DOI: 10.1111/cgf.15046
EUROGRAPHICS 2024 / A. Bermano and E. Kalogerakis
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 2

TAILORME: Self-Supervised Learning of an Anatomically Constrained Volumetric Human Shape Model

S. Wenninger[1†] F. Kemper[1†] U. Schwanecke[2] M. Botsch[1]

[1]TU Dortmund University, Computer Graphics Group, Germany
[2]RheinMain University of Applied Sciences, Computer Vision and Mixed Reality Group, Wiesbaden, Germany

Figure 1: *Our anatomically constrained human shape model allows to infer the skeleton from a surface scan. Due to injecting anthropometric measurements into the latent code, our model can then locally manipulate both the skeleton shape and the soft tissue distribution of a person.*

Abstract

Human shape spaces have been extensively studied, as they are a core element of human shape and pose inference tasks. Classic methods for creating a human shape model register a surface template mesh to a database of 3D scans and use dimensionality reduction techniques, such as Principal Component Analysis, to learn a compact representation. While these shape models enable global shape modifications by correlating anthropometric measurements with the learned subspace, they only provide limited localized *shape control. We instead register a volumetric anatomical template, consisting of skeleton bones and soft tissue, to the surface scans of the CAESAR database. We further enlarge our training data to the full Cartesian product of all skeletons and all soft tissues using physically plausible volumetric deformation transfer. This data is then used to learn an anatomically constrained volumetric human shape model in a self-supervised fashion. The resulting* TAILORME *model enables shape sampling, localized shape manipulation, and fast inference from given surface scans.*

CCS Concepts

• ***Computing methodologies*** → ***Mesh models; Volumetric models; Shape analysis; Learning latent representations;***

1. Introduction

Human shape modeling has been extensively studied due to its application in various fields, such as shape and pose estimation from multi-view stereo or monocular RGB(-D) input. Starting from simple linear PCA models [ASK*05; LMR*15] to recent advances in machine learning models [RBSB18; BBP*19], these models are used as the foundation for many downstream tasks such as body composition estimation [WNT*21], the creation of virtual humans [AWLB17; AMB*19; WAB*20], or generating synthetic training data for image recognition tasks [WBH*21]. Most of the approaches train on commercially available 3D scan databases such as CAESAR [RBD*02] or 3D Scanstore [3DS23]. These 3D scans naturally provide only what is easily observable from the outside: the silhouette of the scanned subject. However, by setting the focus on modelling the *skin layer* of humans, models that want to learn how to accurately *modify* a given virtual human, suffer from missing anatomical information.

† The first two authors contributed equally

Modifying realistic virtual humans has gained attention due to its promising applicability in VR therapy [PSR*14; MTM*18; WDM*22; WMF*22], which can serve as a complementary intervention technique to classic forms of therapy. Such VR systems can immersively expose patients with anorexia or obesity to generic or personalized virtual humans at different levels of weight or Body Mass Index (BMI), allowing patients to reflect on and researchers to gain insight into possibly occurring body image disturbances. However, current models for body weight/BMI modification are typically learnt on surface-only models and employ global models such as Principal Component Analysis [ACPH06; HSS*09; PSR*14; DWM*22], leading to shape modification models providing only limited localized control. Participants have stated the request for changing the composition of specific body parts in addition to a global BMI/weight modification [DWM*22].

In this paper, we present our TAILORME model, which is able to achieve such localized shape manipulation. We leverage recent advances in inferring anatomical structures from surface scans [DLG*13; KIL*16; KWB21; KZBP22] to register a volumetric anatomical template model to the CAESAR database, resulting in pairs of skeleton and skin meshes. The main contribution of our work is to provide a novel statistical model that clearly separates the distribution of skeleton and soft-tissue in its latent space. In order for our model to successfully learn separate parameter sets, we calculate the full Cartesian product of all skeleton shapes and all soft-tissue distributions using volumetric deformation transfer [BSPG06], allowing us to transfer the soft tissue distribution of subject i onto the skeleton of subject j.

The resulting data set is then used to train a neural network that learns the common underlying parameters from a person's bone structure and soft tissue distribution. Examples that share either a common skeleton or a common soft tissue distribution can be sampled from the Cartesian data set. These commonalities are learned and encoded with an autoencoder using the SpiralNet++ [GCBZ19] approach. To separate the skeleton and soft tissue distribution a self-supervised learning approach inspired by Barlow Twins [ZJM*21] is used, which reduces redundancy in the underlying distributions. To allow local modification of body regions, measurements are taken on the example Cartesian data set and are additionally injected into the latent code to reduce the correlation of the remaining parameters with these known measurements.

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

2. Related Work

2.1. Human Shape Models

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

More sophisticated dimensionality reduction techniques have also been applied to human shape models: Ranjan et al. [RBSB18] propose a convolutional mesh autoencoder and introduce a pooling and unpooling operation directly on the mesh surface structure. The Neural 3D Morphable Models (Neural3DMM) network [BBP*19] adjusts the pooling operations and uses a spiral convolutional operator, which has been further refined by Gong et al. [GCBZ19]. Our model uses a similar autoencoder design paired with the self-supervised learning technique Barlow Twins [ZJM*21].

2.2. Modifying Virtual Humans

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

A common approach is to model shape modification by learning linear correlations between a set of anthropometric measurements (e.g., as present in the CAESAR database [RBD*02]) and the low-dimensional shape space [ACPH06; HSS*09; PSR*14; DWM*22]. The modified shape can then be computed by mapping the desired measurement changes into the subspace through learned regressors and then projecting the change in subspace coordinates back into vertex space. Commonly used anthropometric measurements, such as arm length and inseam, are highly correlated. The cited methods cannot completely disentangle this correlation in the anthropometric measurements, leading to limited control over local shape manipulation. Our non-linear model learns to separate the correlations between these measurements, thereby enabling more localized shape manipulations.

For surface models, there is some work on creating more local shape space representations. Tena et al. [TDM11] propose a method for automatically segmenting registered head meshes into several components. A shape space is then learned for each component separately and the resulting submeshes are stitched together. *Sparse PCA* combined with spatially-varying regularization

weights [NVW*13] has also been shown to result in more localized shape models. For an overview about parametric (head) surface models, including global and local models, we refer the reader to the survey by Egger et al. [EST*20]. These methods could achieve localized shape control, but are only trained on surface meshes.

2.3. Anatomical Models

Achenbach et al. [ABG*18] trained a multi-linear model (MLM) to find a lower-dimensional model of skull and corresponding head shape, parameterized by skull shape and soft tissue distribution. The MLM does not completely decouple the two parameter sets, so changing the skin parameters can still affect the skeleton. Our non-linear model better decouples skeleton from skin shape, i.e., when changing skin parameters, the skeleton stays fixed.

Anatomy Transfer [DLG*13] is a method for warping an anatomical template model into a target skin surface via a harmonic space warp while constraining bones to only deform via affine transformations. This can however lead to unnaturally scaled or sheared bones, as discussed in other work [KIL*16; KWB21].

Saito et al. [SZK15] developed a physics-based simulation of muscle and fat growth on a tetrahedral template mesh including an enveloping muscle layer that separates the tetrahedral mesh representing the subcutaneous fat layer from the rest of the template. [KIL*16] present a method for fitting such a physics-based simulation to a set of 3D scans in different poses, to get a personalized anatomical model. Their approach yields visually plausible results but requires a complex numerical optimization strategy taking several minutes and can therefore not be used for interactive VR interventions. Komaritzan et al. [KWB21] follow the approach of Saito et al. [SZK15] by using a multi-layered model to separate skeleton, muscle, and skin surface derived from an anatomical template model [Zyg23]. Their model is then fitted to a given skin layer in a multi-stage optimization scheme. Embedding the high-resolution skeleton and muscle meshes from the anatomical template into the resulting layers is done by a triharmonic RBF warp. However, they do not train a statistical model on the resulting shapes. Additionally, their fitting approach is an order of magnitude slower compared to our method.

The recent work *OSSO* [KZBP22] combines the *STAR* model [OBB20] for human body shapes and a model of skeleton shapes based on the Stitched Puppet Model [ZB15]. By fitting these two models to a set of DXA images, the authors learn to regress skeletal shape from skin shape in PCA space. In the follow-up method *SKEL* [KWS*23] the authors present a parametric biomechanical skeleton and skin model with shared shape and pose parameters and anatomically constrained degrees of freedom. The skeleton model is registered to a subset of the *AMASS* dataset [MGT*19] by optimizing bone scalings and poses via a biomechanical optimization framework [WRS*22]. The skeletons inferred with both the OSSO and SKEL approach may however show self-intersections with the given skin mesh. In contrast, our model learns non-linear correlations between skeleton and skin, provides a localized shape modification model, and produces intersection-free pairs of skeleton and skin meshes.

3. Training Data

We start by deriving all the parts of our template model and registering it to surface scans of the CAESAR database, yielding pairs of skeleton and skin meshes (Section 3.1). We enlarge this data set by computing the full Cartesian product of skeleton shape and soft tissue distribution via volumetric deformation transfer (Section 3.2). The resulting data set then constitutes the training data for our TAILORME model. See Figure 2 for an overview of our method.

3.1. Skin and Skeleton Registration

Existing anatomical models, as provided for example by Zygote [Zyg23] or 3D Scanstore [3DS23], are only available with prohibitive licensing. In order to make our model publicly available, we commissioned a 3D artist to build an anatomical template model. It provides a male and female template, both including meshes for skin, eyes, mouth, teeth, muscle, and skeleton (Figure 3). All meshes in the male template are consistently topologized with their counterparts in the female template. With 23752 vertices, our skin mesh has approximately 3.5 times more vertices than the popular SMPL [LMR*15] or STAR [OBB20] models, allowing us to more accurately model skin geometry. We follow the *layered model* approach and generate a skeleton wrap that envelopes the high-detail skeleton mesh and has the same triangulation as the skin layer[KWB21]. This provides a trivial correspondence between skin and skeletal layers.

Our skin surface input data is derived from the European subset of the *CAESAR* database [RBD*02], consisting of about 1700 3D scans annotated with 3D landmarks and anthropometric measurements. To bring all scans into uniform topology and pose, we employ the template fitting approach proposed by Achenbach et al. [AWLB17], adapted to use the skin surface of our template model. This leaves us with 776 male and 919 female skin meshes denoted by $\mathcal{S}_i$. In the following, all computations are done on the male and female data set separately, due to the anatomical differences especially in the hip and shoulder region.

From the fitted skin meshes, we use an author-provided implementation of InsideHumans [KWB21] to estimate skeleton layers $\mathcal{B}_i$, resulting in non-intersecting pairs of skeleton and skin meshes $(\mathcal{B}_i, \mathcal{S}_i)$. Since the InsideHumans approach excludes the head, hands, and feet region from the skeleton layer, we inherit this limitation. We denote the set of vertices belonging to these regions by $\mathcal{E}$. Equipped with this data, we can now enlarge our training data set by computing the Cartesian product of skeleton shape and soft tissue distribution in a physically plausible way.

3.2. Volumetric Deformation Transfer

We train our model on the Cartesian product of two shape dimensions: skeleton shape and soft tissue distribution. To this end, we transfer the soft tissue of subject i onto the skeleton of subject j, which we achieve through deformation transfer [SP04; BSPG06].

In the standard formulation of deformation transfer, the deformation gradients are computed from a triangle on $\mathcal{B}_i$ to the corresponding triangle on $\mathcal{S}_i$. These deformations are then applied to $\mathcal{B}_j$ in order to generate $\mathcal{S}_j$. However, as seen in Figure 4 (center right),

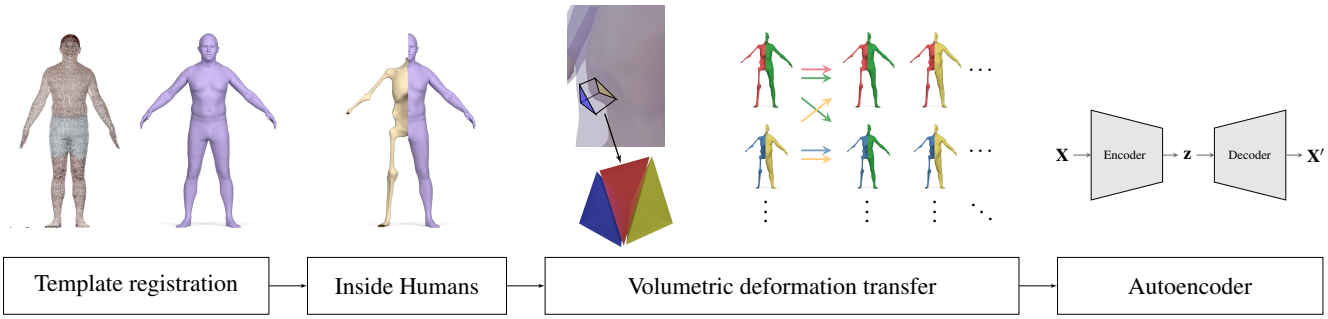

Figure 2: *Overview of our data processing pipeline. We first fit our template model to the CAESAR database, resulting in registered skin surfaces. We use the InsideHumans method [KWB21] to infer anatomical structures from these surface fits. To learn a separated parameter space for skeleton and soft-tissue distribution, we generate the Cartesian product of all soft-tissue distributions and all skeleton shapes, using volumetric deformation transfer. We train our autoencoder by sampling pairs, which share either a common skeleton or soft-tissue distribution from our Cartesian data set.*

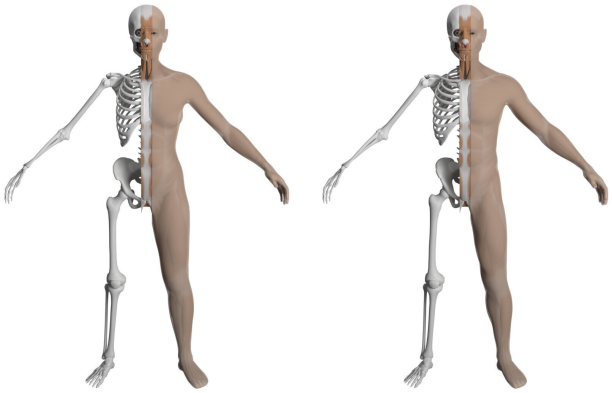

Figure 3: *Male and female template model. In this work, we derive an additional skeleton layer that wraps the high resolution skeleton mesh and shares the triangulation with the skin layer.*

this formulation can lead to interpenetrations of skin and skeleton. These artifacts can happen because the triangle-based deformation gradients between $\mathcal{B}_i$ and $\mathcal{S}_i$ do not encode any volumetric information of the soft tissue enclosed in between these two surfaces. To alleviate this problem we formulate the soft tissue transfer as a *volumetric* deformation transfer problem.

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

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

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

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

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

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

$$\left(\mathbf{G}^\mathsf{T}\mathbf{D}\mathbf{G}\right)\mathbf{X}_j = \left(\mathbf{G}^\mathsf{T}\mathbf{D}\right)\mathbf{F}, \tag{2}$$

with Dirichlet boundary constraints for every vertex belonging to $\mathcal{B}_j \cup \mathcal{E}$. The matrices $\mathbf{G}^\mathsf{T}\mathbf{D}$ and $\mathbf{G}$ represent the discrete divergence and gradient operators for tetrahedral meshes [BSPG06], and $\mathbf{F}$ vertically stacks the desired deformation gradients $\mathbf{F}^t$. Solving this linear system yields new skin vertices $\mathbf{X}_j^{\mathcal{S}}$.

In order to smoothly blend into the boundary region $\mathcal{E}$, we define per-tetrahedron interpolation weights $w^t \in [0,1]$, which decrease based on the distance to $\mathcal{E}$. We use w^t to linearly interpolate between the desired deformation gradients $\mathbf{F}^t$ and the deformation gradients computed from $\mathbb{S}_\mu$ to the target subject $\mathbb{S}_j$, thereby ensuring a smooth transition into $\mathcal{E}$.

In the presented volumetric formulation, the deformation gradients $\mathbf{F}^t$ include information about the volumetric stretching and

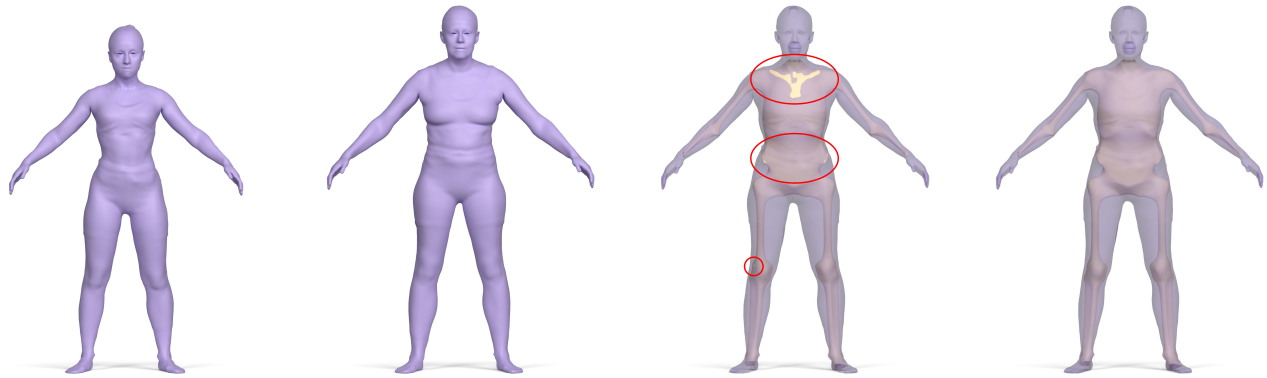

Figure 4: *Transferring the soft tissue of the left model onto the skeleton of the center-left model using* surface-based *deformation transfer, the skeleton wrap protrudes the skin (center-right). Our volumetric deformation transfer successfully avoids these artifacts (right).*

compression of the tetrahedron t of the soft tissue layer. Since $\mathbb{S}_\mu$ and $\mathbb{S}_i$ do not exhibit inverted elements, the deformation gradients $\mathbf{F}^t$ do not contain any inversions. As such, solving Equation (2) avoids self-intersections between skin and skeleton (shown in Figure 4, right), since those would require tetrahedra to invert, which in turn would lead to a high deviation from the target deformation gradient. Figure 5 shows several examples of transferring the soft tissue distribution of a set of subjects with different height and weight characteristics onto the same skeleton.

4. Model Learning

Our objective is to learn a compact representation of human body shapes. We do so using a specific autoencoder architecture. To enable guided and localized shape manipulation, we inject anthropometric measurements into the autoencoder's latent representation. We measure the length of the torso, arms, and legs on the skeleton, and the circumference of chest, waist, abdomen, and hips on the skin meshes. By injecting normalized values of those measurements into our latent representation, we form an expressive latent code. Our neural network architecture is a convolutional autoencoder with local mesh convolutions based on SpiralNet++ [GCBZ19]. Decoupling the two shape dimensions is accomplished by splitting the latent code into two parameter sets: one for skeleton shape and one for soft tissue distribution (Section 4.1). We define a loss function based on the Barlow Twins method [ZJM*21], which allows us to reduce the redundancy in the latent code (Section 4.2).

4.1. Network architecture

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

As a first design for the shape compression task, we experimented with utilizing two separate PCA models for the skeleton and soft tissue distribution. The PCA weights then formed the input to our autoencoder, which learned to decouple the two shape dimensions. This is in line with the OSSO approach [KZBP22], where the correlation between skin and skeleton shape is learned by a linear regressor between two PCA subspaces. We found that the resulting model separates the skeleton and soft tissue distribution, but only provides global shape control when modifying semantic parameters in the latent space, due to the global influence of the PCA weights. Figure 6 shows an example of the global influence, where modifying arm length also changes the body height.

To mitigate the global effects, we opt for an autoencoder using the SpiralNet++ approach [GCBZ19], which utilizes a mesh convolution and pooling operator. This design enables local shape control when modifying the entries in the latent space that correspond to the anthropometric measurements.

The structure of our autoencoder is shown in Figure 7. The samples $\mathbf{X} \in \mathbb{R}^{37143\times 3}$ drawn from our Cartesian product data set (Section 3.2) consist of the vertex positions $\mathbf{X}^{\mathcal{S}}$ and $\mathbf{X}^{\mathcal{B}}$ of the skin mesh and the skeleton wrap (the latter excluding vertices in $\mathcal{E}$ belonging to head, hands, and feet). We normalize the vertex positions before using them as input for our autoencoder. Our latent code utilizes 48 parameters for the skeleton and soft tissue distribution each. We take three measurements on the skeleton (torso, arm and leg length), four measurements on the skin (chest, waist, abdomen and hip circumference), and append them to the resulting embeddings via skip connections. This results in a total of 103 parameters in the latent space.

4.2. Cross-correlation Loss

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

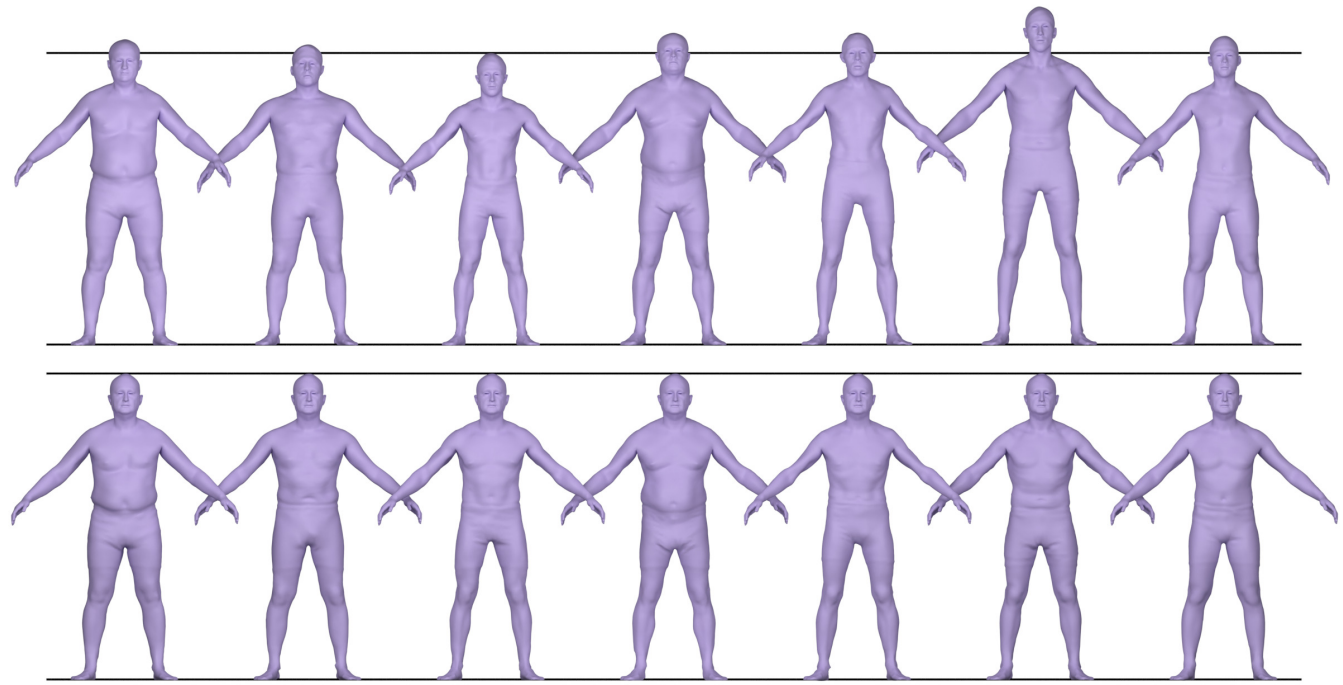

Figure 5: *Exemplary results of transferring the soft tissue of various people (top row) onto a single target skeleton via volumetric deformation transfer (bottom row). Note that soft tissue characteristics of the top row and skeletal dimensions of the bottom row are faithfully preserved.*

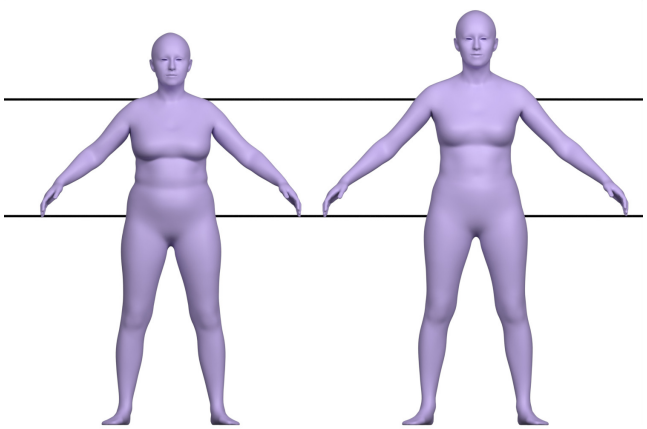

Figure 6: *Representing skeletons and skins in PCA subspaces separates their parameters, but the global nature of PCA prevents localized changes: Increasing the arm length of the left model also causes the body height to increase (right).*

share soft tissue distribution should result in identical soft tissue embeddings $\mathbf{z}^{(\mathcal{S})}$. We achieve this using a self-supervised learning approach based on Barlow Twins [ZJM*21], where the loss formulation penalizes dissimilar embeddings for similar input samples.

We extend the concept of pairs in Barlow Twins by using quadruplets, built by all four combinations of skeletal and soft tissue distribution from two samples each for pairs of skeleton and skin meshes. We randomly select two different indices k,l for skeletons and two different indices for the distribution of soft tissues m,n from our training data set. Let $\mathbf{X}_{km}$ denote the vertex positions resulting from transferring the soft tissue distribution of subject m onto the skeleton of subject k via volumetric deformation transfer (Section 3.2). From the chosen indices (k,l,m,n) we create a quadruplet containing the entries $(\mathbf{X}_{km},\mathbf{X}_{kn},\mathbf{X}_{lm},\mathbf{X}_{ln})$, such that each entry shares either its skeleton or its soft tissue distribution with two of the other entries. These quadruplets are processed in batches by our autoencoder. The forward process of the encoder for one quadruplet in a batch is visualized in Figure 8.

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

$$\mathcal{L}_{\mathrm{BT}} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2, \tag{3}$$

with

$$C_{ij} = \frac{1}{\mathrm{BS}} \sum_b z^A_{b,i} \cdot z^B_{b,j}. \tag{4}$$

The batch size is marked as BS, the batch dimension is denoted by b, and the index dimensions of the network output are represented by i and j. λ is a trainable hyperparameter to weight the

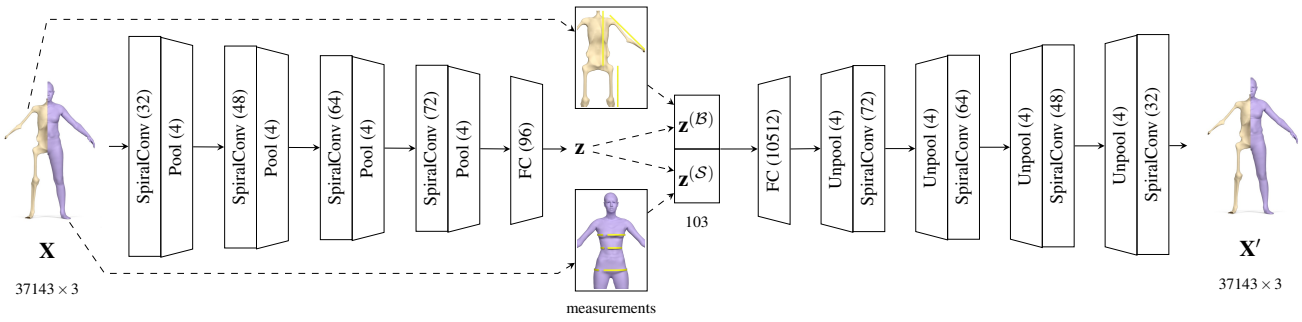

Figure 7: *Our network architecture is based on four SpiralNet++[GCBZ19] convolution and pooling blocks in the encoder. A final dense layer is connecting the last layer of the encoder to achieve our embeddings* $\mathbf{z}$*. We divide the embeddings into two parts, one for the skeleton and the other one for the soft tissue distribution. We append the normalized values of the measurements (the lengths of torso, arms and legs and the circumferences of chest, waist, abdomen, and hips) taken on the input mesh via a skip connection to the latent code. The decoder is the reversed order of the encoder using four unpooling and convolution blocks.*

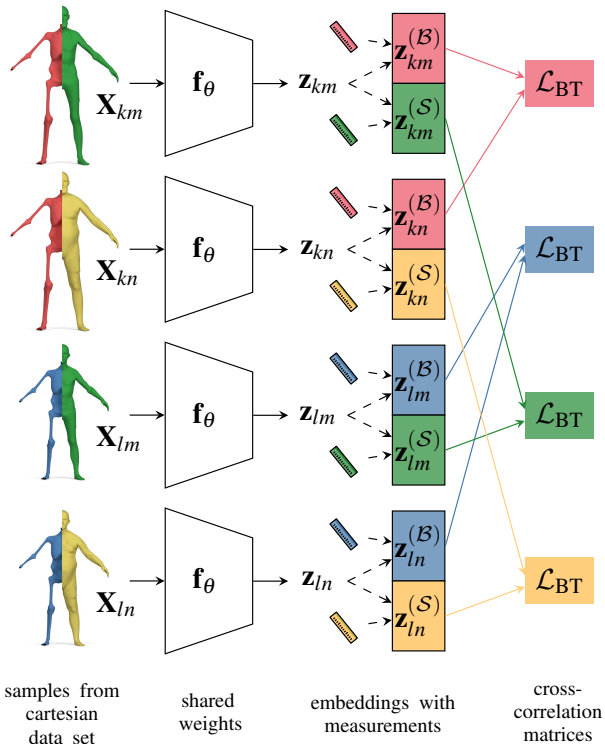

Figure 8: *Processing a quadruplet of samples in the encoder and dividing the embeddings into two parts: one for the skeleton and one for the soft tissue distribution. After separation the normalized measurements are appended to the respective embeddings. If the divided embeddings are noted the same way, they are merged for the entire batch. In order to calculate the cross-correlation loss between pairs, the cross-correlation matrices are calculated for the positions colored in the same way.*

importance of off-diagonal entries being close to 0 in the empirical cross-correlation matrices. $\mathbf{z}^{(A)}$ and $\mathbf{z}^{(B)}$ are batches of embeddings, which are selected as described in the following. Note that Equation (4) differs from the original definition in that we do not use batch normalization on the embeddings before calculating the entries of the cross-correlation matrices C_{ij}. Our model achieves greater accuracy without batch normalization and enables more efficient manual modifications to the reconstructed meshes.

We rearrange the embeddings in a batch to group all components with similarities on the source data set. These embeddings should have a minimal redundancy when originating from the same distribution. This is indicated when sharing one of their indices in the training data set k, l, m, or n. We can form four cross-correlation matrices for the quadruplets in the batch: $\mathbf{z}^{(\mathcal{B})}_{km} \otimes \mathbf{z}^{(\mathcal{B})}_{kn}$, $\mathbf{z}^{(\mathcal{B})}_{lm} \otimes \mathbf{z}^{(\mathcal{B})}_{ln}$, $\mathbf{z}^{(\mathcal{S})}_{km} \otimes \mathbf{z}^{(\mathcal{S})}_{lm}$, and $\mathbf{z}^{(\mathcal{S})}_{kn} \otimes \mathbf{z}^{(\mathcal{S})}_{ln}$, where $\otimes$ denotes the outer product of the batched embeddings. We calculate the cross-correlation loss $\mathcal{L}_{\mathrm{BT}}$ for each matrix using Equation (3) and sum up the total loss for all four matrices of the batch to compute our redundancy loss $\mathcal{L}_Q$. Note that we do not need to minimize the redundancy between skeleton and soft tissue parameters directly to learn a separation.

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

$$\mathcal{L} = \mathcal{L}_R + \beta \mathcal{L}_Q, \tag{5}$$

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

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

rameter for off-diagonal entries to achieve best performance was $\lambda = 2.3 \cdot 10^{-2}$ for females and $\lambda = 4.3 \cdot 10^{-2}$ for males.

5. Post-Processing

After the inference of the decoder, the resulting meshes might show certain artifacts. We observed asymmetries in the face region, which are amplified when modifying the latent code towards the boundary of the learned distribution. This is in part due to the fact that the variance of the face region was not part of the modification in the training data set (Section 3.1). To mitigate potential artifacts, we apply three post-processing steps after decoding: (i) the face region is symmetrized, (ii) the resulting skin surface is smoothed and (iii) intersections between the skeleton and skin layer are resolved. As a final step, we embed the high-resolution anatomical skeleton into the skeleton wrap using a triharmonic space warp.

5.1. Face Symmetrizing & Smoothing

After the inference of the decoder, we approximately symmetrize the face region by adapting the approach of Mitra et al. [MGP07]. A reflective symmetry plane is defined at the center of the head, based on which corresponding vertex pairs (v_i, v_j) on both sides can be determined. The y-coordinate of these vertex pairs are adjusted to match approximately: $y_i' = \frac{3}{4} y_i + \frac{1}{4} y_j$. Afterwards, one explicit smoothing step [DMSB99] is performed on the skin mesh in order to reduce high frequency noise, which may occur when applying drastic changes to the latent parameters.

5.2. Intersection Avoidance

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

When inferring the skeleton for a skin $\mathcal{S}$ given by a 3D scan (as demonstrated in Section 6.5), we want to keep the vertices on $\mathcal{S}$ fixed, as they can be considered ground truth. As such, in order to resolve the detected collisions, we solve for a new skeleton layer $\mathcal{B}$ by minimizing

$$E(\mathcal{B}) = E_{\text{reg}}(\mathcal{B}, \bar{\mathcal{B}}) + E_{\text{close}}(\mathcal{B}, \bar{\mathcal{B}}) + w_{\text{coll}} E_{\text{coll}}(\mathcal{B}, \mathcal{S}), \quad (6)$$

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

$$E_{\text{reg}}(\mathcal{B}, \bar{\mathcal{B}}) = \frac{1}{2} \sum_{\mathbf{x}_i \in \mathcal{B}} A_i \left\| \Delta \mathbf{x}_i - \mathbf{R}_i \Delta \bar{\mathbf{x}}_i \right\|^2. \quad (7)$$

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

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

$$E_{\text{close}}(\mathcal{B}, \bar{\mathcal{B}}) = \frac{1}{|\mathcal{V} \setminus \mathcal{C}|} \sum_{\mathbf{x}_i \notin \mathcal{C}} \left\| \mathbf{x}_i - \bar{\mathbf{x}}_i \right\|^2, \quad (8)$$

and E_{coll} defines the collision avoidance term:

$$E_{\text{coll}}(\mathcal{B}, \mathcal{S}) = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{x}_i \in \mathcal{C}} w_i \left\| \mathbf{x}_i - \pi_{\mathcal{S}}(\mathbf{x}_i) \right\|^2, \quad (9)$$

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

We iteratively minimize Equation (6) via the projective dynamics solver implemented in the ShapeOp library [DDB*15]. We set $w_{\text{coll}} = 50$, $w_i = 1$, and progressively increase the per-vertex collision avoidance weight w_i by 1 each iteration in which the collision could not be resolved. After each iteration, the Laplacian of the initial state $\bar{\mathcal{B}}$ in Equation (7) is updated to the current solution, thereby making the skeleton layer slightly less rigid. Following this optimization scheme, we could reliably resolve all between-layer-collisions in our tests.

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

5.3. Embedding High-resolution Skeleton

Once an intersection-free pair $(\mathcal{B}, \mathcal{S})$ is generated, we embed the high-resolution skeleton mesh $\mathcal{SK}$ by following Komaritzan et al. [KWB21], i.e., using a space warp based on triharmonic radial basis functions [BK05]. The matrix of the involved linear system depends on the template skeleton $\tilde{\mathcal{B}}$ only and hence can be prefactorized. After generating a new skeleton $\mathcal{B}$, the solution can be inferred by back-substitution, and the space warp can efficiently be evaluated to embed the high-resolution anatomical skeleton.

6. Results and Applications

The resulting TAILORME model allows local shape manipulation based on the injected measurements in the latent space. For a demonstration of the final model, we refer the reader to the accompanying video. In the following, we evaluate the performance of the model on our test data set, and compare our approach to the related approaches of OSSO [KZBP22], SKEL [KWS*23], MLM [ABG*18] and standard PCA approaches [ACPH06; PSR*14]. Finally, we demonstrate the modification of 3D-scanned realistic virtual humans.

6.1. Model Evaluation

To quantitatively evaluate the fit of our trained model defined in Section 4, we do not perform the post-processing decribed in Section 5. We use separate data sets for training and model evaluation. The training subset is utilized for model optimization, while a validation subset is used to conduct an automated evaluation and to estimate the models capability for generalization. To minimize the validation set bias, we utilize a third subset of our data set for testing. The splitting is done such that the skeleton and soft tissue distribution of an individual ends up in only one of these subsets.

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

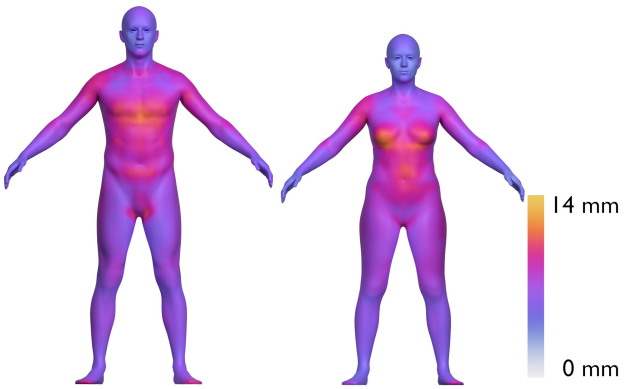

Figure 9: *Mean Euclidean vertex distance when evaluating over all samples in the male (left) and female (right) test data set. Our model achieves a maximum Euclidean distance of 11.5 mm for males and 13.9 mm for females.*

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

Figure 9 displays our model's reproduction error for the skin, measured as per-vertex distance averaged over all meshes in the test data set, when using the decoder back propagation method. The vertex distances of the fitted skins are evenly distributed on the limbs and face, but in the chest and abdomen regions the largest average deviation from the target is observed. Overall, our model attains a maximum per-vertex error of 13.9 mm over all samples in the test data set.

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

$$\mathcal{L}_1 = \frac{1}{3n} \left\| \mathbf{X} - \mathbf{X}' \right\|_1 = \frac{1}{3n} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \mathbf{x}'_i \right\|_1 . \tag{10}$$

When using the encoder and decoder for reproduction of the test data set, we achieve a mean absolute error for the skeleton wrap and the skin of 5.2 ± 1.5 mm for females and 5.4 ± 1.5 mm for males. The cross-correlation matrices in Equation (4) converge to the identity matrix. The individual mesh measurements positively correlate with each other as the circumferential measurements on the original meshes show a strong connection.

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

6.2. Comparison to OSSO and SKEL

We qualitatively compare our work to the OSSO approach [KZBP22]. This method computes a linear regressor between skin and skeleton PCA shape spaces, after fitting both shape models to a set of DXA Scans. As DXA scans are taken in a lying pose, OSSO first reposes a given skin mesh to this pose, infers skeleton shape there, and finally reposes the given result to the input pose. As seen in Figure 10, when compared to our skeleton prediction, the skeleton inferred by OSSO exhibits a skewed and unnaturally shifted rib cage as well as large gaps between bone structures, such as the elbow region or in between ribs and spine. We also compare our work to the SKEL approach [KWS*23], where a combined parametric model for biomechanical skeleton and skin shape is presented. The final model is able to infer an animatable biomechanical skeleton from given SMPL parameters as shown in Figure 10. We observe that SKEL's template skeleton misses bones for the clavicles and the lower spine unnaturally detaches from the pelvis, when fitting the model to a target skin.

Moreover, both OSSO's and SKEL's skeleton protrudes through the given skin, while our method resolves these intersections (Section 5.2). We evaluated the number of skeleton and skin intersections by fitting our model, OSSO, and SKEL to 1697 samples from the European subset of the CAESAR data set [RBD*02]. For OSSO and SKEL, there is no single sample which is free of intersections. Our model produces self-intersections in only 1.30 % of cases, and then only due to the RBF warp in the head region, where hairs are not correctly handled in our method – a limitation we inherit from the InsideHumans approach [KWB21].

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

6.3. Comparison to MLM

We compare our model to the multilinear model (MLM) presented in [ABG*18]. The MLM approach requires the computation of a 3D tensor to separate the skeleton and soft tissue dimensions. Given our model with 51 skeleton and 52 soft tissue parameters, the MLM requires a total of 292M parameters. Our method requires two orders of magnitude fewer parameters (2.1M) to process the same number of input variables. When applying the MLM approach to our training data, we found that the decoupling process of the two parameter sets is incomplete. This effect can also be observed in the original work. Achenbach et al. [ABG*18] provide a demo application at `https://ls7-gv.cs.tu-dortmund.de/publications/2018/vcbm18/vcbm18.html`, where changing the first skull parameter causes subsequent changes to the first *soft tissue* parameter to also change the resulting skull shape.

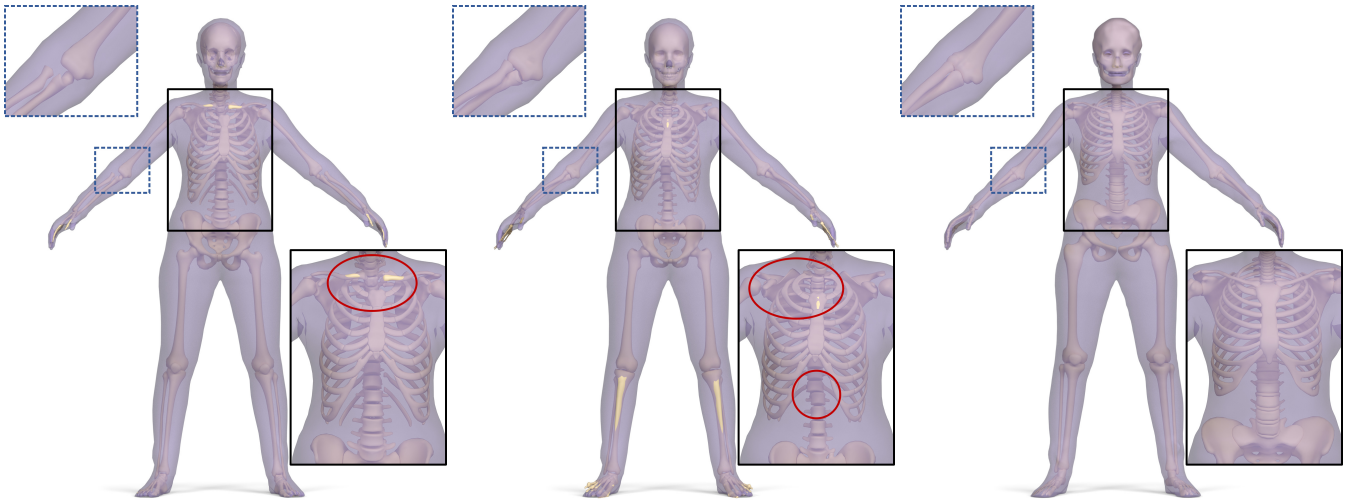

Figure 10: *Comparison of OSSO [KZBP22] (left) and SKEL [KWS*23] (center) with our approach (right). The OSSO skeleton protrudes through the skin (yellow protrusions circled in red), we resolve these kinds of collisions (right). The rib cage inferred by OSSO is skewed (black inset) and the stitched puppet model results in gaps between bone structures (dashed blue inset). The SKEL skeleton (center) is missing the clavicles (circled in red). The individual spine segments (lumbar, thoracic and cervical) are not properly aligned (black inset).*

6.4. Comparison to Surface PCA

To show the benefits of our local and non-linear autoencoder and mesh convolution design (Section 4), we compare our method to the common approach of modelling anthropometric shape manipulation by correlating measurements with a global and linear PCA subspace learned from surface scans [ACPH06; PSR*14]. These methods allow global shape manipulation, but provide only limited *local* control. For an example of this effect, we compare the results of shortening arm length with our model and the model proposed by Piryankova et al. [PSR*14]. Figure 11 shows that our model provides local control of arm length, whereas the surface based approach results in notable changes in the leg region. To interactively explore the shortcomings of the surface based approach, we invite the reader to experiment with the demo application at `https://bodyvisualizer.com`, and try to change the inseam parameter, while keeping the other measurements fixed. This changes the model's arm length in addition to the desired effect of changing the leg length.

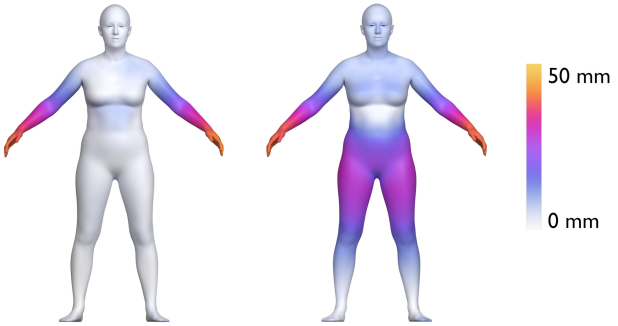

Figure 11: *Comparison of our localized shape modification (left) with a global PCA approach [PSR*14] (right). Both models were used to shorten arm length by* 38 mm*. The vertex distance from the original to the modified mesh is color coded. Our model enables more localized shape changes, while the global PCA approach considerably changes the leg region when modifying arm length.*

6.5. Modifying Virtual Humans

As demonstrated in Figure 1, we can also fit our model to surface scans of clothed humans, allowing us to modify virtual humans with our TAILORME model. To this end, given a registered surface scan conforming to our skin layer topology, we let our model infer skin and skeleton shape by optimizing the mean absolute error with additional weight decay in order to prevent overfitting.

To determine a fit for the skeleton and soft tissue distribution of a person, gradient descent is performed on the latent parameters $\mathbf{z}$ using the Adam optimizer [KB15] implemented in PyTorch and running on the GPU. We apply a weight decay of $7.5 \cdot 10^{-5}$ to prevent fitting values that are too far outside of the learned embeddings. Fitting the skeleton and soft tissue parameters takes $<$ 100 ms on a desktop PC equipped with an NVidia RTX 3090 GPU and an Intel Core i9 10850K CPU. The post-processing steps (Section 5) add another 700 ms to the total inference time. This is an order of magnitude faster than the Inside Humans approach [KWB21], whose authors report a total time of approximately 20 s on similar hardware. We measure an inference time of approximately 2 min for the publicly available implementation of OSSO [KZBP22].

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

Figure 12: *The results of fitting our model to different scan poses (top row). We first unpose the scan by fitting our human skin surface template. The skeleton inference is then performed in the trained A-pose. Finally, we employ linear blending skinning to repose the results back to the observed scan pose (bottom row). Note that the fingers of the second pose are not faithfully reproduced by the pose estimation of the employed template fitting technique.*

latent parameters $\mathbf{z}$ we apply the difference of decoding $\tilde{\mathbf{z}}$ and $\mathbf{z}$ to the scan, resulting in the modified person $\mathbf{X}'$:

$$\mathbf{X}' = \mathbf{X} + (\mathbf{g}_\theta(\mathbf{z}) - \mathbf{g}_\theta(\tilde{\mathbf{z}})). \tag{11}$$

To prevent unnatural deformation of the head region, we stitch the original head of the scanned person back onto the resulting mesh using differential coordinates similar to Döllinger et al. [DWM*22].

6.6. Experiments with different poses

In Figure 12, we show examples of fitting our model to poses which differ from our trained A-pose. We follow the OSSO approach [KZBP22] and first unpose a given scan by employing the surface-based template fitting approach of Achenbach et al. [AWLB17]. Our model is then fit to the resulting A-pose surface mesh, resulting in an unposed high-resolution skeleton embedding. Since our template skeleton shares its animation rig with the skin surface, we can then use linear blend skinning to repose the resulting skeleton.

Note that after applying the very simplistic linear blend skinning, the resulting skeleton and skin meshes may show self-intersections, as we only resolve these intersections in the A-pose of our template. As linear blend skinning is obviously not an anatomically correct animation technique, more sophisticated animation methods such as proposed in SKEL [KWS*23] or Fast Projective Skinning [KB18] should be incorporated in future work to allow anatomically sound animations based on our model.

7. Limitations

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

Our training data lacks information about the bone structure underlying the head, hands, and feet of our subjects. Therefore, our model cannot properly reproduce these areas. Due to this fact, we observe asymmetric structures, especially occurring in the facial region, which are amplified when modifying the latent code.

Although our resulting meshes are free of self-intersections, this property only holds in the A-Pose. When animating the resulting skeleton and skin, due to our simplistic use of linear blend skinning we cannot guarantee that the skeleton does not protrude the skin.

8. Conclusion

We presented TAILORME, a novel approach for learning a volumetric anatomically constrained human shape model. We computed the Cartesian product of skeleton shapes and soft tissue distributions from the CAESAR database using volumetric deformation transfer. To decouple the two shape dimensions, we utilized a Barlow Twins inspired learning approach to train our autoencoder from pairs of skeleton and soft tissue distribution. The resulting model can be used for shape sampling, e.g., generating various soft tissue distributions on the same skeleton. It provides localized shape manipulation due to the injected measurements in the latent space of our autoencoder. Compared to other methods, our model better decouples the skeleton and soft tissue shape dimensions, allows more localized shape manipulation, and provides significantly faster inference time.

In future work our model can be extended to include other anatomical details such as the muscles used to generate and modify humans. The skeleton, muscle, and soft tissue layers then can be separated by our model applying a triple Barlow twins loss, where pairs of eight are processed in a batch. Incorporating a skeleton into the hands and feet will be beneficial rather than keeping them static. Fitting a skull inside the head, similar to the approach of Achenbach et al. [ABG*18], enables a more accurate modification of the person's facial features.

Being able to infer anatomical structures can be used to improve the animation of virtual humans. Further investigations into developing an animation method for volumetric virtual humans that avoids self-intersections is an interesting direction for future work.

9. Acknowledgements

We would like to thank Alexander Boerner (https://www.alexanderborner.de/) for creating the template models and allowing us to make our model publicly available. We thank Martin Komaritzan for his support when porting the InsideHumans implementation to our template model. Stephan Wenninger has been funded by "Stiftung Innovation in der Hochschullehre" through the project "Hybrid Learning Center" (FBM2020-EA-690-01130). Open Access funding enabled and organized by Projekt DEAL.

References

[3DS23] 3DSCANSTORE. https://www.3dscanstore.com/. 2023 1, 3.

[ABG*18] ACHENBACH, JASCHA, BRYLKA, ROBERT, GIETZEN, THOMAS, et al. "A Multilinear Model for Bidirectional Craniofacial Reconstruction". *Proc. of Eurographics Workshop on Visual Computing for Biology and Medicine*. 2018, 67–76 3, 8, 9, 11.

[ACPH06] ALLEN, BRETT, CURLESS, BRIAN, POPOVIĆ, ZORAN, and HERTZMANN, AARON. "Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis". *Proc. of the ACM SIGGRAPH/Eurographics symposium on Computer animation*. 2006, 147–156 2, 8, 10.

[AMB*19] ALLDIECK, THIEMO, MAGNOR, MARCUS, BHATNAGAR, BHARAT LAL, et al. "Learning to Reconstruct People in Clothing From a Single RGB Camera". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 1.

[ASK*05] ANGUELOV, DRAGOMIR, SRINIVASAN, PRAVEEN, KOLLER, DAPHNE, et al. "SCAPE: Shape Completion and Animation of People". *ACM Transactions on Graphics* 24.3 (2005), 408–416 1, 2.

[AWLB17] ACHENBACH, JASCHA, WALTEMATE, THOMAS, LATOSCHIK, MARC ERICH, and BOTSCH, MARIO. "Fast Generation of Realistic Virtual Humans". *Proc. of ACM Symposium on Virtual Reality Software and Technology*. 2017, 12:1–12:10 1, 3, 11.

[BBP*19] BOURITSAS, GIORGOS, BOKHNYAK, SERGIY, PLOUMPIS, STYLIANOS, et al. "Neural 3D Morphable Models: Spiral Convolutional Networks for 3D Shape Representation Learning and Generation". *The IEEE International Conference on Computer Vision (ICCV)*. 2019 1, 2.

[BK05] BOTSCH, MARIO and KOBBELT, LEIF. "Real-Time Shape Editing using Radial Basis Functions". *Computer Graphics Forum* 24.3 (2005), 611–621 8.

[BKP*10] BOTSCH, MARIO, KOBBELT, LEIF, PAULY, MARK, et al. *Polygon Mesh Processing*. AK Peters, CRC press, 2010 8.

[BSPG06] BOTSCH, MARIO, SUMNER, ROBERT, PAULY, MARK, and GROSS, MARKUS. "Deformation Transfer for Detail-Preserving Surface Editing". *Vision, Modeling, and Visualization* (2006), 357–364 2–4.

[DDB*15] DEUSS, MARIO, DELEURAN, ANDERS HOLDEN, BOUAZIZ, SOFIEN, et al. "ShapeOp – A Robust and Extensible Geometric Modelling Paradigm". *Proc. of Design Modelling Symposium*. 2015, 505–515 8.

[DLG*13] DICKO, ALI-HAMADI, LIU, TIANTIAN, GILLES, BENJAMIN, et al. "Anatomy transfer". *ACM Transactions on Graphics* 32.6 (2013), 188:1–188:8 2, 3.

[DMSB99] DESBRUN, MATHIEU, MEYER, MARK, SCHRÖDER, PETER, and BARR, ALAN H. "Implicit Fairing of Irregular Meshes Using Diffusion and Curvature Flow". *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '99. 1999, 317–324 8.

[DWM*22] DÖLLINGER, NINA, WOLF, ERIK, MAL, DAVID, et al. "Resize Me! Exploring the User Experience of Embodied Realistic Modulatable Avatars for Body Image Intervention in Virtual Reality". *Frontiers in Virtual Reality* 3 (2022) 2, 11.

[EST*20] EGGER, BERNHARD, SMITH, WILLIAM A. P., TEWARI, AYUSH, et al. "3D Morphable Face Models—Past, Present, and Future". *ACM Transactions on Graphics* 39.5 (2020) 3.

[FB12] FREIFELD, OREN and BLACK, MICHAEL J. "Lie Bodies: A Manifold Representation of 3D Human Shape". *Computer Vision – ECCV*. 2012, 1–14 2.

[GCBZ19] GONG, SHUNWANG, CHEN, LEI, BRONSTEIN, MICHAEL, and ZAFEIRIOU, STEFANOS. "SpiralNet++: A Fast and Highly Efficient Mesh Convolution Operator". *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2019, 4141–4148 2, 5, 7.

[HSS*09] HASLER, N., STOLL, C., SUNKEL, M., et al. "A Statistical Model of Human Pose and Body Shape". *Computer Graphics Forum* 28.2 (2009), 337–346 2.

[KB15] KINGMA, DIEDERIK P. and BA, JIMMY. "Adam: A Method for Stochastic Optimization". *International Conference on Learning Representations (ICLR)*. 2015 7, 9, 10.

[KB18] KOMARITZAN, MARTIN and BOTSCH, MARIO. "Projective Skinning". *Proc. of the ACM on Computer Graphics and Interactive Techniques* 1.1 (2018) 11.

[KIL*16] KADLEČEK, PETR, ICHIM, ALEXANDRU-EUGEN, LIU, TIANTIAN, et al. "Reconstructing Personalized Anatomical Models for Physics-based Body Animation". *ACM Transactions on Graphics* 35.6 (2016), 213:1–213:13 2, 3.

[KWB21] KOMARITZAN, MARTIN, WENNINGER, STEPHAN, and BOTSCH, MARIO. "Inside Humans: Creating a Simple Layered Anatomical Model from Human Surface Scans". *Frontiers in Virtual Reality* 2 (2021). ISSN: 2673-4192 2–4, 8–10.

[KWS*23] KELLER, MARILYN, WERLING, KEENON, SHIN, SOYONG, et al. "From Skin to Skeleton: Towards Biomechanically Accurate 3D Digital Humans". *ACM Transactions on Graphics* 42.6 (2023) 3, 8–11.

[KZBP22] KELLER, MARILYN, ZUFFI, SILVIA, BLACK, MICHAEL J., and PUJADES, SERGI. "OSSO: Obtaining Skeletal Shape from Outside". *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, 20492–20501 2, 3, 5, 8–11.

[LMR*15] LOPER, MATTHEW, MAHMOOD, NAUREEN, ROMERO, JAVIER, et al. "SMPL: A Skinned Multi-Person Linear Model". *ACM Transactions on Graphics* 34.6 (2015) 1–3.

[MGP07] MITRA, NILOY J., GUIBAS, LEONIDAS, and PAULY, MARK. "Symmetrization". *ACM Transactions on Graphics (SIGGRAPH)* 3 (2007), #63, 1–8 8.

[MGT*19] MAHMOOD, NAUREEN, GHORBANI, NIMA, TROJE, NIKOLAUS F., et al. "AMASS: Archive of Motion Capture As Surface Shapes". *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019 3.

[MTM*18] MÖLBERT, SIMONE CLAIRE, THALER, ANNE, MOHLER, BETTY J, et al. "Assessing body image in anorexia nervosa using biometric self-avatars in virtual reality: Attitudinal components rather than visual body size estimation are distorted". *Psychological Medicine* 48.4 (2018), 642–653 2.

[NVW*13] NEUMANN, THOMAS, VARANASI, KIRAN, WENGER, STEPHAN, et al. "Sparse Localized Deformation Components". *ACM Transactions on Graphics* 32.6 (2013) 3.

[OBB20] OSMAN, AHMED A A, BOLKART, TIMO, and BLACK, MICHAEL J. "STAR: A Sparse Trained Articulated Human Body Regressor". *Computer Vision – ECCV*. 2020, 598–613 2, 3.

[PSR*14] PIRYANKOVA, I., STEFANUCCI, J., ROMERO, J., et al. "Can I recognize my body's weight? The influence of shape and texture on the perception of self". *ACM Transactions on Applied Perception for the Symposium on Applied Perception* 11.3 (2014), 13:1–13:18 2, 8, 10.

[PWH*17] PISHCHULIN, LEONID, WUHRER, STEFANIE, HELTEN, THOMAS, et al. "Building Statistical Shape Spaces for 3D Human Modeling". *Pattern Recognition* (2017), 276–286 2.

[RBD*02] ROBINETTE, KATHLEEN M, BLACKWELL, SHERRI, DAANEN, HEIN, et al. *Civilian american and european surface anthropometry resource (CAESAR), final report. volume 1: Summary*. Tech. rep. Sytronics Inc, 2002 1–3, 9.

[RBSB18] RANJAN, ANURAG, BOLKART, TIMO, SANYAL, SOUBHIK, and BLACK, MICHAEL J. "Generating 3D faces using Convolutional Mesh Autoencoders". *Computer Vision – ECCV*. 2018, 725–741 1, 2.

[SP04] SUMNER, ROBERT W. and POPOVIĆ, JOVAN. "Deformation Transfer for Triangle Meshes". *ACM Transactions on Graphics* 23.3 (2004), 399–405 3.

[SZK15] SAITO, SHUNSUKE, ZHOU, ZI-YE, and KAVAN, LADISLAV. "Computational Bodybuilding: Anatomically-based Modeling of Human Bodies". *ACM Transactions on Graphics* 34.4 (2015), 41:1–41:12 3.

[TDM11] TENA, J. RAFAEL, DE LA TORRE, FERNANDO, and MATTHEWS, IAIN. "Interactive Region-Based Linear 3D Face Models". *ACM Transactions on Graphics* 30.4 (2011), 76:1–76:10 2.

[WAB*20] WENNINGER, STEPHAN, ACHENBACH, JASCHA, BARTL, ANDREA, et al. "Realistic Virtual Humans from Smartphone Videos". *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology*. VRST '20. 2020 1.

[WBH*21] WOOD, ERROLL, BALTRUŠAITIS, TADAS, HEWITT, CHARLIE, et al. "Fake It Till You Make It: Face Analysis in the Wild Using Synthetic Data Alone". *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, 3681–3691 1, 11.

[WDM*22] WOLF, ERIK, DÖLLINGER, NINA, MAL, DAVID, et al. "Does Distance Matter? Embodiment and Perception of Personalized Avatars in Relation to the Self-Observation Distance in Virtual Reality". *Frontiers in Virtual Reality* 3 (2022) 2.

[WMF*22] WOLF, ERIK, MAL, DAVID, FROHNAPFEL, VIKTOR, et al. "Plausibility and Perception of Personalized Virtual Humans between Virtual and Augmented Reality". *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2022, 489–498 2.

[WNT*21] WONG, MICHAEL C., NG, BENNETT K., TIAN, ISAAC, et al. "A pose-independent method for accurate and precise body composition from 3D optical scans". *Obesity* 29.11 (2021), 1835–1847 1.

[WRS*22] WERLING, KEENON, RAITOR, MICHAEL, STINGEL, JON, et al. "Rapid bilevel optimization to concurrently solve musculoskeletal scaling, marker registration, and inverse kinematic problems for human motion reconstruction". *bioRxiv* (2022) 3.

[ZB15] ZUFFI, SILVIA and BLACK, MICHAEL J. "The Stitched Puppet: A Graphical Model of 3D Human Shape and Pose". *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, 3537–3546 3.

[ZJM*21] ZBONTAR, JURE, JING, LI, MISRA, ISHAN, et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". *Proceedings of the 38th International Conference on Machine Learning*. Ed. by MEILA, MARINA and ZHANG, TONG. Vol. 139. Proceedings of Machine Learning Research. 2021, 12310–12320 2, 5, 6.

[Zyg23] ZYGOTE. https://www.zygote.com. 2023 3.