

# OSSO: Obtaining Skeletal Shape from Outside

Marilyn Keller<sup>1</sup> Silvia Zuffi<sup>2</sup> Michael J. Black<sup>1</sup> Sergi Pujades<sup>3</sup>

<sup>1</sup>Max Planck Institute for Intelligent Systems, Tübingen, Germany

<sup>2</sup>IMATI-CNR, Milan, Italy

<sup>3</sup>Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, France

{mkeller, black}@tue.mpg.de

silvia@mi.imati.cnr.it sergi.pujades-rocamora@inria.fr



Figure 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 [28] scans.

## Abstract

We address the problem of inferring the anatomic skeleton of a person, in an arbitrary pose, from the 3D surface of the body; i.e. we predict the inside (bones) from the outside (skin). This has many applications in medicine and biomechanics. Existing state-of-the-art biomechanical skeletons are detailed but do not easily generalize to new subjects. Additionally, computer vision and graphics methods that predict skeletons are typically heuristic, not learned from data, do not leverage the full 3D body surface, and are not validated against ground truth. To our knowledge, our system, called OSSO (Obtaining Skeletal Shape from Outside), is the first to learn the mapping from the 3D body surface to the internal skeleton from real data. We do so using 1000 male and 1000 female dual-energy X-ray absorptiometry (DXA) scans. To these, we fit a parametric 3D body shape model (STAR) to capture the body surface and a novel part-based 3D skeleton model to capture the bones. This provides inside/outside training pairs. We model the statistical variation of full skeletons using PCA in a pose-normalized space and train a regressor from body shape parameters to skeleton shape parameters. Given an arbitrary 3D body shape and pose, OSSO predicts a realistic skeleton

inside. In contrast to previous work, we evaluate the accuracy of the skeleton shape quantitatively on held out DXA scans, outperforming the state-of-the-art. We also show 3D skeleton prediction from varied and challenging 3D bodies. The code to infer a skeleton from a body shape is available at <https://osso.is.tue.mpg.de>, and the dataset of paired outer surface (skin) and skeleton (bone) meshes is available as a Biobank Returned Dataset. This research has been conducted using the UK Biobank Resource.

## 1. Introduction

The estimation of 3D human pose and shape (HPS) from images, video, and other data sources is widely studied and has many applications. Current methods for HPS exploit detailed models of the visible surface of the human body learned from 3D scans. While the surface shape is accurate, these models are all based on a “skeleton” that only approximately models the kinematic structure of the body using a small number of linear segments with ball joints. While these simplified skeletons are useful for the animation of virtual characters and action recognition, they are not appropriate for applications in medicine and biomechanics.

Moreover, each 3D body model defines its own kinematic skeleton; transferring 3D pose information between them often requires optimization and introduces errors. To be more widely relevant, HPS methods need to output a skeleton that corresponds to the true, 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 work, 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. 1 left). Given a posed body, our method predicts the skeleton from the body shape, and poses it inside subject to anatomic constraints (Fig. 1 right).

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 [8,9,21,33] represent different body parts: skin, muscles, organs, 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 [1,4,15,43] 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 [3,20,23,41]. These have two elements in common: they model the human external shape, i.e. the skin surface, and they are learned from data. Using thousands of 3D scans, these models capture the statistical variability of the human body shape. Here we use STAR [23], because it has a richer shape representation than SMPL [20]. Such models, however, employ a simplified kinematic skeleton and joints. While they can be readily inferred from data, the idealized skeletal structure means that 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,

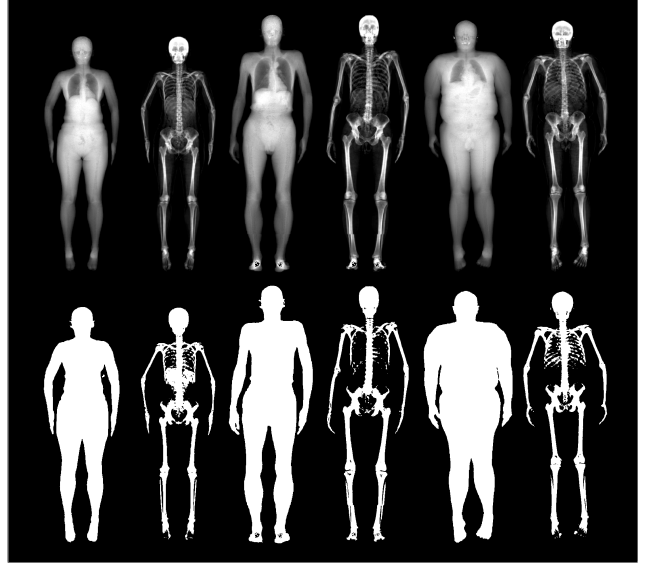


Figure 2. Top: Pairs of soft tissue ( $I_S$ ) and bone ( $I_B$ ) DXA images. Bottom: Computed skin ( $M_S$ ) and bone ( $M_B$ ) masks.

which is harmful to humans; e.g. Computed Tomography (CT) and X-rays. This means that 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 [36]. 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. 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 UK Biobank [36], 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 [1]. We also demonstrate the use of OSSO by estimating skeletons for a variety of body shapes and poses.

OSSO makes the following contributions: (1) We fit a 3D body model to DXA images to obtain 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.

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, our work is the first to learn and validate such a skeleton based on data of the inside and the outside of the body.

## 2. Related Work

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

**Data-driven skin models.** Learned statistical body models [2, 3, 20, 23, 41] leverage large datasets of 3D scans. In our work, we use STAR [23] to represent the body surface with two parameter vectors ( $\beta_S, \theta_S$ ) 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 [29], 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 [5, 7, 11, 12, 16, 17, 26, 27, 35, 37, 39] and was recently surveyed [42]. However, less work fits such models to X-ray images. Pansiot and Boyer [24] leverage video-based surface motion capture to recover a volumetric representation of the hand from planar X-ray images. In our work, we leverage a silhouette term [46] 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. [10] propose a morphing algorithm to register a template skeleton to a target skeleton mesh or 3D im-

age. 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.

Saito et al. [31] simulate the growth of fat, muscle, and bones to generate new body shapes. Kadlecěk et al. [15] propose a physics-based anatomic model that can be adapted to input 3D scans, where, similar to Zhu et al. [43], the bones are deformed using linear blend skinning with bounded bi-harmonic weights. While the obtained bones are visually plausible, these models are neither learned from data nor validated against it.

In Phace [13], 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. [38] 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. [44] 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. [4] infer the skeleton of a subject from RGBD images of the skin. Their skeleton inference method is based on Anatomy Transfer [1] 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 [32] couples a musculoskeletal biomechanical model to the SCAPE model [3]. 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) [1]. 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 our work, we use a similar approach by leveraging the Stitched Puppet graphical model [45]. 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. [40], 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. Data

A key contribution of our work 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 images. The dataset is made available to the community as a Biobank Returned Dataset.

Creating the dataset has several steps: (1) we segment DXA images to get the silhouettes of the body and bones (Sec. 3.1), (2) we create synthetic skeleton silhouettes and use them to learn to predict landmarks (Sec. 3.2), (3) we register STAR [23] to the skin silhouette images (Sec. 3.3), (4) we create a custom skeleton model (Sec. 3.4) and register it to real skeleton binary masks (Sec. 3.5). Fig. 3 shows an overview of the dataset creation procedure.

#### 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. 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 [34], and accurate segmentations are performed manually [6]. Moreover, these methods only focus on a small set of bones. Jamalud et al. [14] use a U-Net to segment body parts from DXA scans that are relevant for scoliosis classification. Unfortunately the code is not available

In our work, 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 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. 2); this is what we need to fit a 3D skeleton to them.

#### 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. Consequently, we train a landmark detector for skeleton binary masks  $M_B$ . To do so, we generate synthetic training data using an initial skeleton mesh similar to the one from Anatomy Transfer (AT) [1]. We rig the skeleton so that we can control it with the STAR shape and pose parameters (see Sup. Mat. for details), and define 29 landmarks on the skeleton mesh,  $\mathcal{L}_I$ , that are in correspondence with the 3D joints of the STAR model. 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. From the synthetic silhouettes of the skeleton, we train the landmark detector using a stacked hourglass network [22]. In Sup. Mat., we provide the details and evaluation of this network.

#### 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 [19, 30]. 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 [23] 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  $E_\theta$  as in [5]. 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 [46] and effectively solve for the STAR



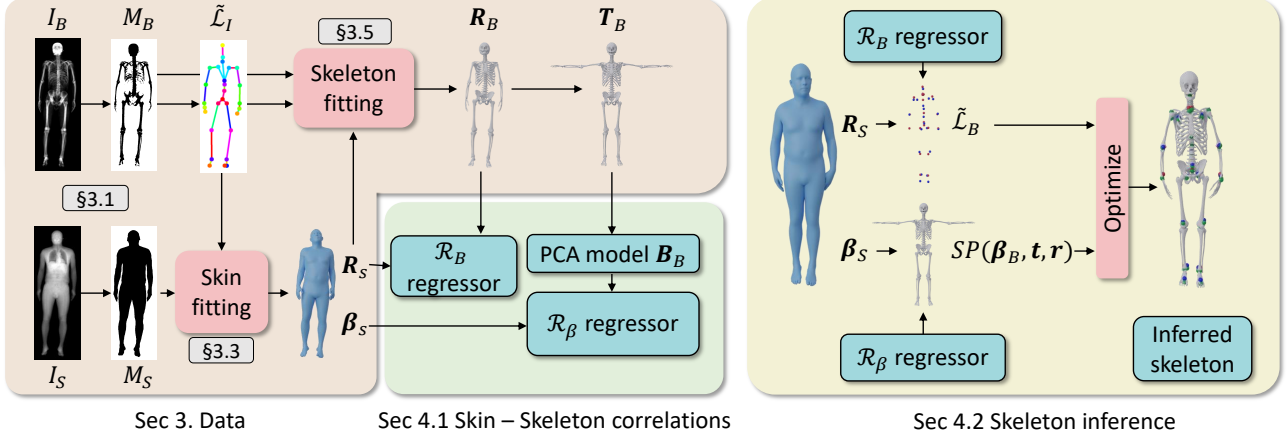


Figure 3. Overview of OSSO: learning and inference. From the input DXA images ( $I_B, I_S$ ) we obtain the skeleton and skin masks ( $M_B, M_S$ ). From the skeleton mask  $M_B$ , we predict 2D landmarks  $\tilde{\mathcal{L}}_I$  and use them to register STAR to  $M_S$  and obtain  $\mathbf{R}_S$  and  $\beta_S$ . We then register our skeleton graphical model to  $M_B, \tilde{\mathcal{L}}_I$  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  $(\beta_S, \beta_B)$  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  $\mathcal{R}_B(\mathbf{R}_S) = \tilde{\mathcal{L}}_B$ .

shape and pose parameters ( $\hat{\beta}_S, \hat{\theta}_S$ ) that minimize:

$$\begin{aligned}
 E_S(\beta_S, \theta_S; M_S, \tilde{\mathcal{L}}_I) = & E_{sil}(ST(\beta_S, \theta_S), M_S) \\
 & + \lambda_I \|P(\mathbf{J}(\beta_S, \theta_S)) - \tilde{\mathcal{L}}_I\| \\
 & + \lambda_\beta \|\beta_S\| + \lambda_\theta E_\theta(\theta_S) + \lambda_h E_h(ST(\beta_S, \theta_S)),
 \end{aligned} \quad (1)$$

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 [46])<sup>1</sup>,  $\tilde{\mathcal{L}}_I$  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  $\beta_S$  in Fig. 3).

This approach works well, but can fail for 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.

### 3.4. Skeleton graphical model

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. 3.2) is not well suited for this task. Thus, we create a new skeleton model, capitalizing on the *stitched puppet* [45] and the synthetic shape deformation space from *GLOSS* [47]. The *stitched puppet* provides an ideal graphical model formulation, allowing each group of bones to move independently, while the stitching potentials enforce the coherence of the full skeleton.

<sup>1</sup>We use the open available implementation at [https://github.com/silviazufi/smalr\\_online](https://github.com/silviazufi/smalr_online)

Starting with the same AT skeleton template as before, we manually define groups of bones that belong to the same anatomic part, and define the interfaces between these parts. Unlike the original formulation [45], the stitching potentials are not defined as springs between corresponding surface points. Instead, we define corresponding points by selecting vertices of connected parts that have a distance below a certain threshold (see Sup. Mat.). Also, unlike [45], 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  $\tilde{\mathcal{L}}_I$  are properly defined.

### 3.5. Skeleton from DXA

Next, 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  $E_B(\beta_B, \mathbf{t}, \mathbf{r}) =$

$$E_{data}(\beta_B, \mathbf{t}, \mathbf{r}; M_B, \tilde{\mathcal{L}}_I, \mathbf{R}_S) + E_{prior}(\beta_B, \mathbf{t}, \mathbf{r}), \quad (2)$$

where  $E_{data}(\beta_B, \mathbf{t}, \mathbf{r}; M_B, \tilde{\mathcal{L}}_I, \mathbf{R}_S) =$

$$\begin{aligned}
 & E_{sil}(P(SP(\beta_B, \mathbf{t}, \mathbf{r})), M_B) \\
 & + \lambda_I \|P(\mathcal{L}_I(SP(\beta_B, \mathbf{t}, \mathbf{r}))) - \tilde{\mathcal{L}}_I\| \\
 & + \lambda_i E_i(SP(\beta_B, \mathbf{t}, \mathbf{r}, \mathbf{R}_S))
 \end{aligned} \quad (3)$$

and  $E_{prior}(\beta_B, \mathbf{t}, \mathbf{r}) =$

$$\begin{aligned}
 & \lambda_{shape} \|\beta_B\| + \lambda_{pose} \|\mathbf{r} - \mathbf{r}_T\| + \\
 & + \lambda_{sti} E_{sti}(SP(\beta_B, \mathbf{t}, \mathbf{r})) + \lambda_{sy} E_{sy}(SP(\beta_B, \mathbf{t}, \mathbf{r}))
 \end{aligned} \quad (4)$$

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 [45], 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). The implementation details of this cost are found in Sup. Mat. 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  $\mathbf{R}_B = SP(\hat{\beta}_B, \hat{\mathbf{t}}, \hat{\mathbf{r}})$  (see  $\mathbf{R}_B$  in Fig. 3).

**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  $\mathbf{d}_p^0 = (\mathbf{R}_S[sn_p] - \mathbf{R}_B[sk_p])$  as their initial 3D offset.

This allows us to define a signed distance cost between the unposed meshes  $E_d(\mathbf{T}_S, \mathbf{T}_B) = \sum_p w_p \cdot (\mathbf{T}_S[sn_p] - \mathbf{T}_B[sk_p]) - \mathbf{d}_p^0$ , with  $w_p = \text{sign}(\mathbf{T}_S[sn_p] - \mathbf{T}_B[sk_p]) \cdot N(\mathbf{T}_B[sk_p])$  where  $N(\mathbf{T}_B[sk_p])$  is the normal on the skeleton mesh at vertex  $sk_p$ . We fix  $\beta_B = \hat{\beta}_B$  and find  $(\hat{\mathbf{t}}_u, \hat{\mathbf{r}}_u)$  that minimize  $E_u(\mathbf{t}, \mathbf{r}) =$

$$\lambda_{sti} E_{sti}(SP(\hat{\beta}_B, \mathbf{t}, \mathbf{r})) + \lambda_d E_d(\mathbf{T}_S, SP(\hat{\beta}_B, \mathbf{t}, \mathbf{r})) \quad (5)$$

to obtain  $\mathbf{T}_B = SP(\hat{\beta}_B, \hat{\mathbf{t}}_u, \hat{\mathbf{r}}_u)$  (see  $\mathbf{T}_B$  in Fig. 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  $\hat{\mathcal{L}}_B$  from the skin vertices  $\mathbf{R}_S$  (Sec. 4.1). Then, at test time, given an arbitrary STAR body shape in an arbitrary pose, we predict a skeleton mesh (Sec. 4.2) and then repose it to match the input skin pose (Sec. 4.3). See Fig. 3.

### 4.1. Skeleton statistics and correlation to skin shape

We next learn the correlation between the skin and the bones. With the unposed skeletons  $\mathbf{T}_B$ , we first compute a low-dimensional linear subspace  $\mathcal{B}_B$ , 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  $\beta_B \in \mathcal{B}_B$  from the STAR shape space coefficients  $\beta_S$  computed in Sec. 3.3.

To properly constrain the 3D location of the 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.  $\mathcal{L}_B(\mathbf{R}_B)$ . We formulate the problem as a non-negative least squares problem and solve it with an active set method [18].

### 4.2. Inferring the skeleton from the skin

We now have all 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. 4.1) we predict the subject’s skeleton shape  $\beta_B = \mathcal{R}_\beta(\beta_S)$  from the body surface shape  $\beta_S$ . Then, to properly position the skeleton inside the body, we pose the body surface in a normalized lying pose  $\theta_S^L$ , 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  $\beta_B$  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:

$$E(\mathbf{t}, \mathbf{r}) = \lambda_L ||\mathcal{L}_B(SP(\beta_B, \mathbf{t}, \mathbf{r})) - \tilde{\mathcal{L}}_B|| + \lambda_{ct} E_{ct}(SP(\beta_B, \mathbf{t}, \mathbf{r}), \mathbf{R}_S(\theta_S^L)), \quad (6)$$

where  $E_{ct}$  forces the contact between the skeleton and the skin, see Sup. Mat. for the detailed definition. The obtained mesh is our skeleton prediction.

### 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. We need a more precise anatomical model of the joints. This could be very detailed (e.g. OpenSim [33]) but would complicate optimization. Hence, we strike a balance between simplicity and realism and 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 stay at a similar distance.

To replicate the ligaments of the human knee we define pairs of vertices at attachment points and an energy  $E_l$  to constrain their distance. In Sup. Mat. we provide implementation details of  $E_j$  and  $E_l$ .

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  $SP_0 = SP(\beta_B, \mathbf{t}_0, \mathbf{r}_0)$  inside the corresponding lying down skin mesh  $ST_0 = ST(\theta_S^L, \beta_S)$ , and a skin mesh in a specific pose  $ST_\theta = ST(\theta_S^P, \beta_S)$ , we pose the skeleton inside

$ST_\theta$ . 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 Eq. (5) to position the skeleton inside the posed body  $ST_\theta$ . Then, to enforce more realistic anatomic joints we add

$$E(\mathbf{t}, \mathbf{r}) = \lambda_t E_j(\mathbf{t}, \mathbf{r}; SP_0) + \lambda_j E_l(\mathbf{t}, \mathbf{r}; SP_0) \quad (7)$$

to Eq. (5) and optimize again.

## 5. Experiments

To evaluate our approach, we first quantify how accurately the skin registrations  $\mathbf{R}_S$  match the skin masks  $M_S$  (Sec. 5.1). We then evaluate how accurately our learned regressors predict the 3D bone landmarks from the skin (Sec. 5.2). Finally, we quantitatively and qualitatively evaluate how the projections of the computed bones overlap with the DXA bones masks (Sec. 5.3).

In our experimental setting, for each gender, we use a training set of 1000 subjects and a test set of 200 subjects from the UK Biobank dataset [36]. We made sure both sets have the same Body Mass Index distribution. We compare OSSO with Anatomy Transfer (AT) [1] on the 200 male and 200 female test subjects held out from any learning.

### 5.1. 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) 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 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. Overall, the skin registrations faithfully capture the skin masks (see examples in Sup. Mat.).

### 5.2. Skin to 3D landmark regressors

We next evaluate the accuracy of the regressors  $\mathcal{L}_B$  (Sec. 4.1) by predicting skeleton landmark locations from the body surface on the test set. We train the regressors on the train set and we evaluate the 3D distance between the landmarks on the aligned skeleton  $\mathcal{L}_B(\mathbf{R}_B)$  and the predicted landmarks  $\mathcal{R}_B(\mathbf{R}_S) = \tilde{\mathcal{L}}_B$ . In Sup. Mat., we present the table with the distances for all landmarks (male and female). Our predictions have a mean distance (MD) below 1 cm:  $8.0 \pm 6.1$  mm for males and  $8.4 \pm 6.7$  mm for females and all individual landmarks results are consistent among male and female ( $\pm 1$ mm). The more accurate landmarks correspond to the upper skull ( $MD < 2$  mm) and feet ( $MD < 4$  mm), whereas the least accurate belong 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. 2).

	Male	Female	Male	Female
Method	$\cap_R(\%) \uparrow$		HD (px) $\downarrow$	
$\mathbf{R}_B$	92	94	$8.2 \pm 2.6$	$5.6 \pm 1.7$
OSSO	<b>88</b>	<b>89</b>	<b><math>10.6 \pm 3.2</math></b>	<b><math>9.1 \pm 2.3</math></b>
AT~ [1]	84	88	$14.4 \pm 2.9$	$11.5 \pm 3.1$

Table 1. Quantitative comparison of OSSO and AT [1]. The  $\cap_R$  score standard deviations are all below 2%.

### 5.3. 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 1 presents the results on the test set. In the first row we evaluate the skeletons  $\mathbf{R}_B$  from Sec. 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 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 ( $\cap_R = 1$ ,  $HD = 0$ ). In Fig. 4 and Sup. Mat. we show that visually, OSSO’s predictions are coherent and match the DXA bone images better than Anatomy Transfer. In Sup. Mat. we provide examples of subjects with high Body Mass Index, for which Anatomy Transfer predicts a stretched skeleton, while ours are closer to the DXA skeleton mask.

### 5.4. 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. 4.3). We show several examples in Fig. 1, Fig. 5 and Sup. Mat. The clothed scans are from RenderPeople [28] and are part of the AGORA dataset [25], 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.



Figure 4. From left to right: input  $R_S$ , AT prediction, overlap with  $M_B$ , OSSO prediction, overlap with  $M_B$ ,  $I_B$  DXA. Overlap image color code: orange is  $M_B$  only (false negative), white is the intersection of both (true positive) and green predicted only.



Figure 5. Qualitative evaluation of the skeleton inference in arbitrary poses. OSSO yields visually plausible results.

Although some minor skin interpenetrations remain, the obtained results are visually plausible.

## 6. Conclusion

OSSO addresses the problem of predicting the skeleton of a person from their external body shape. We use STAR [23] 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 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. To our knowledge, this is the first method to

quantitatively validate the accuracy of a skeleton predicted from the body surface.

**Limitations, Future Work, and Risks.** OSSO predicts a person’s skeleton from their body shape. If the parametric body model does not accurately represent someone’s shape, then the skeleton prediction is likely to be poor. We use STAR instead of SMPL because it captures a broader range of body shapes, but it still has limitations. It does not well represent bodies that are extremely thin, extremely obese, with scoliosis, the elderly or children, amputees, and transgender individuals. In Sup. Mat., we show failure cases for out of distribution bodies.

Our current skeleton and its joints, while more realistic than those of STAR, are still an approximation to the true human skeleton. For example, we use simplified models of the spine and lower arm. We plan to add more anatomical detail to the model in future work. One possible path would be to integrate our method with existing anatomical models like OpenSim [33]. Since models like STAR are not controlled by an anatomic skeleton, posed models may deviate from valid human shapes. Future work should retrain models like STAR with our more realistic skeleton. This could be done by using OSSO to infer the skeleton for 3D training scans in a variety of poses. The blend weights and pose-corrective shapes of a STAR-like model could then be learned and controlled by the true skeleton.

A limitation of the present study is that our evaluation is only performed for lying down poses. Ideally, we would like to train and test our method on complex upright poses. Currently, this is not possible because there are no datasets that capture the inside and the outside of the body in various poses. Possible 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.

Our work is motivated by applications in medicine, biomechanics, sports science, etc. Possible negative uses of the technology would involve capturing the skeletal data of a person (e.g. from video) without their permission. If future work shows that the skeleton is accurate enough to diagnose diseases like arthritis, the technology could be used, without consent, to learn about someone’s risk of disease.

**Acknowledgments.** This research has been conducted using the UK Biobank Resource under the Approved Project ID 51951. Authors thank the International Max Planck Research School for Intelligent Systems for supporting MK. MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH. MJB’s research was performed solely at MPI. SP’s work was funded by the ANR SEMBA project. We thank Anatoscope for the initial skeleton mesh.



# Appendices

In this supplementary material, we provide further details of our method and elaborate on the results presented in the main paper. Specifically:

In Sec. A we detail how we train a 2D landmark predictor from DXA silhouettes and quantitatively evaluate the accuracy of the 2D predicted landmarks on the synthetic data. This section extends Sec. 3.2 of the main document.

In Sec. B, we provide further details about the skin and skeleton registrations to the DXA images. This section provides further details of Sec.s 3 and 4 of the main paper.

In Sec. C we present an evaluation of the skeleton shape space obtained in Sec. 4.1 of the main paper.

In Sec. D we provide quantitative and qualitative results to complement the Sec. 5 from the main document.

In Tab. 2 we summarize the notation used in the paper for an easy reference.

## A. Predicting 2D landmarks on DXA scans

In order to register the skin and skeleton models to the DXA scans, we need 2D landmarks on the scans. In this section we explain how we generate the synthetic dataset (Sec. A.1, Sec. A.2) to train a 2D landmark predictor from DXA skeleton silhouettes (Sec. A.3) and evaluate the prediction (Sec. A.4). The 2D landmark prediction from DXA silhouette is illustrated in Fig. 6.

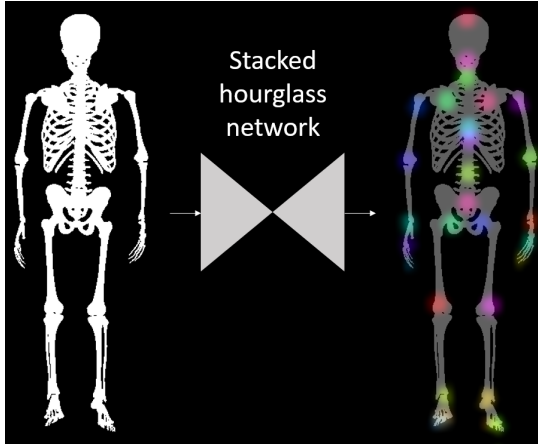


Figure 6. From a skeleton mask, a stacked hourglass network predicts the 2D locations of the landmarks  $\tilde{\mathcal{L}}_I$ .

### A.1. 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 [23] parameters.

Table of notation in OSSO

$I_S$	$\triangleq$	Dxa soft tissue image (skin)
$I_B$	$\triangleq$	Dxa bone image (skeleton)
$M_S$	$\triangleq$	Skin mask segmented from $I_S$
$M_B$	$\triangleq$	Skeleton mask segmented from $I_B$
$ST(\beta_S, \theta_S)$	$\triangleq$	STAR body model [23]
$\mathcal{B}_S$	$\triangleq$	STAR shape space
$K(\beta_S, \theta_S)$	$\triangleq$	The initial skeleton model rigged to the STAR shape and pose parameters
$SP(t, r, \beta_B)$	$\triangleq$	Our <i>skeleton stitched puppet</i> model
$\hat{M}_B$	$\triangleq$	Synthetic skeleton mask generated with $K$
$\mathcal{L}_I$	$\triangleq$	29 3D landmarks whose 24 firsts correspond to STAR joints location and the closest skeleton vertices in $K$
$\tilde{\mathcal{L}}_I$	$\triangleq$	2D landmarks predicted from $M_B$ .
$\mathbf{R}_S$	$\triangleq$	STAR body model registered to $M_S$
$\mathbf{R}_B$	$\triangleq$	Our <i>skeleton stitched puppet</i> model registered to $M_S$
$\mathbf{T}_B$	$\triangleq$	$\mathbf{R}_B$ unposed in T pose
$\mathcal{L}_B$	$\triangleq$	63 3D landmarks defined as vertices on the skeleton mesh template
$\mathcal{R}_B$	$\triangleq$	Regressor to predict skeleton landmarks $\mathcal{L}_B$ from a STAR body model registration $\mathbf{R}_S$
$\tilde{\mathcal{L}}_B$	$\triangleq$	$\mathcal{L}_B$ landmarks location inferred with the regressor $\mathcal{R}_B$
$\mathcal{B}_B$	$\triangleq$	PCA model of the skeleton learned from $\mathbf{T}_B$
$\mathcal{B}_S$	$\triangleq$	STAR PCA shape space
$\mathcal{R}_\beta$	$\triangleq$	Regressor to predict skeleton shape components $\beta_B \in \mathcal{B}_B$ from STAR shape components $\beta_S \in \mathcal{B}_S$
$SI_{AT}$	$\triangleq$	Skeleton mesh inferred with AT
$SI_{OSSO}$	$\triangleq$	Skeleton mesh inferred with OSSO

Table 2. Table of Notation

We first generate 21 STAR bodies by sampling the STAR shape space  $\mathcal{B}_S$ . 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) [1], we register a template skeleton mesh to each of these body shapes. Effectively we enforce the skin of the AT mesh to match 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  $\mathcal{B}_i = \mathbf{T}_{\beta_i=2} - \mathbf{T}_{\beta_i=-2}$  for  $i$  in  $[0, n_\beta]$ , else  $\mathcal{B}_i = \mathbf{0}$ .

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.

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. 7). 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.

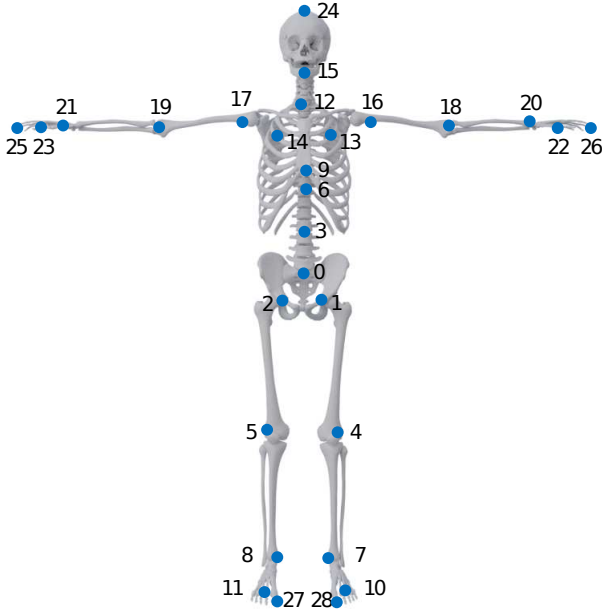


Figure 7. Position of the 3D landmarks  $\mathcal{L}_I$  on the Stitched Puppet skeleton model  $P_B$ . These markers correspond to the location of the STAR 3D joints plus 5 additional landmarks.

## A.2. 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  $\beta$  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*  $\theta_L$  and sample new angles from a uniform distribution centered at  $\theta_L$  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. 8.

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

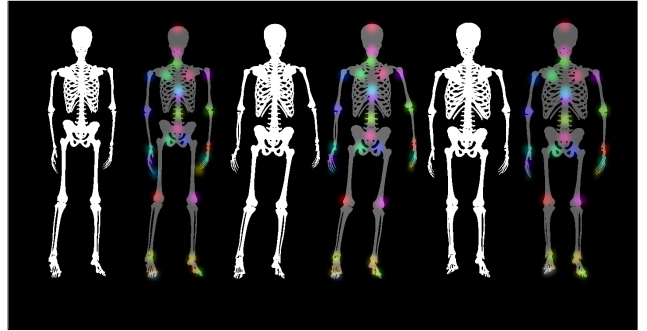


Figure 8. Pairs of synthetic skeleton masks (in white) and 2D landmarks  $\tilde{\mathcal{L}}_I$  (color-coded) overlaid on the mask (in gray).

## A.3. 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 [22] 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  $\tilde{\mathcal{L}}_I$ .

In Fig. 9, we show qualitative results of the predicted landmarks on binary masks from real DXA images. We visually inspected the predicted 2D landmarks and observe that the silhouette simplification strategy combined with our data augmentation technique allows to obtain very good qualitative results on real DXA images.

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

Table 3. 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. 7.

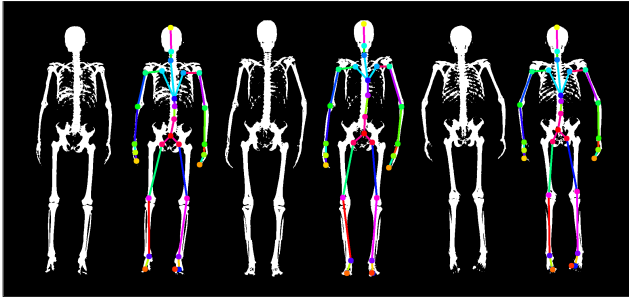


Figure 9. Pairs of input and predicted 2D landmarks  $\tilde{\mathcal{L}}_I$  on real DXAs. The network learned on synthetic data generalizes well to real data.

#### A.4. 2D landmarks prediction evaluation

As the original DXA images do not have annotations, we only evaluate quantitatively on the synthetic dataset. 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.

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.

## B. Skin and skeleton registrations to DXA

This section provides further details to complement the sections 3.3, 3.4 and 3.5 of the main paper.

### B.1. Skeleton model based on Stitched Puppet

We create a parametric skeleton model to align to the DXA skeleton silhouettes based on the *stitched puppet* [45].

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 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 [45] 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.

In our skeleton model, we manually define 21 groups of bones that belong to the same anatomic part, and define the interfaces between these parts. In Fig. 10 we show the different parts with color codes, their interfaces, as well as the 3D landmarks  $\mathcal{L}_B$  defined on the bones.

### B.2. Registration costs

In this section, we detail the costs used for the skeleton registration (Sec 3.5 of the main paper) and the final reposing (Sec 4.3 of the main paper). In this section, 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.

**Skeleton to DXA registration.** In Sec. 3.5 of the main paper, we introduce the cost  $E_i$  to constrain the skeleton inside the body. We decompose  $E_i$  as  $E_i = E_{in} + E_p + E_{ct}$  and illustrate the intuition of each cost in Fig. 11.

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

$$E_{in} = \max(0, D_z(SP(\beta_B, \mathbf{t}, \mathbf{r}), \mathbf{R}_S)) \quad (8)$$

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

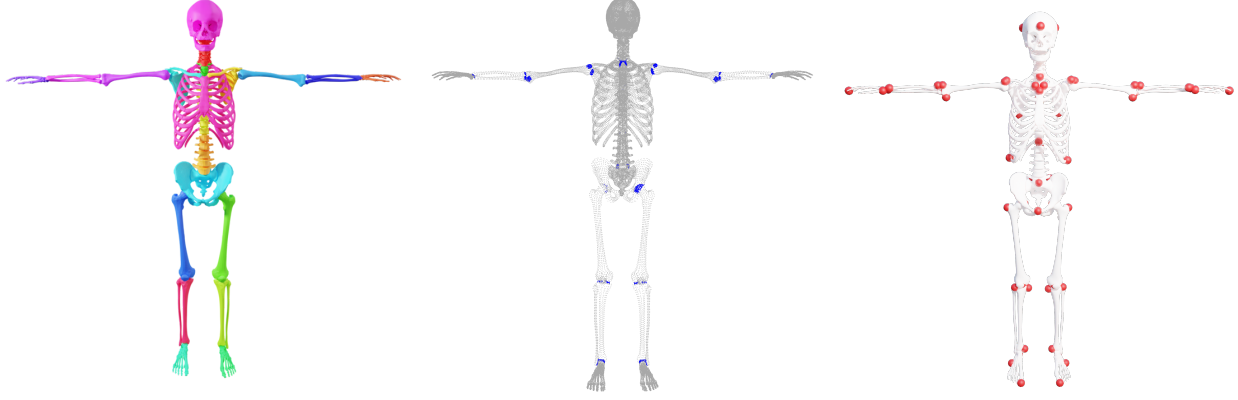


Figure 10. Our *stitched puppet* skeleton model, with the different bone groups (left), the interface point between the groups (center) and the 3D landmarks  $\mathcal{L}_B$  (right).

area  $A$ , we define

$$E_p = v_{sp}^z - \sum_{v_{st} \in A} (v_{st}^z). \quad (9)$$

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

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:

$$E_{ct} = v_{sp} - (v_{st} - e \cdot v_{st}^n) \quad (10)$$

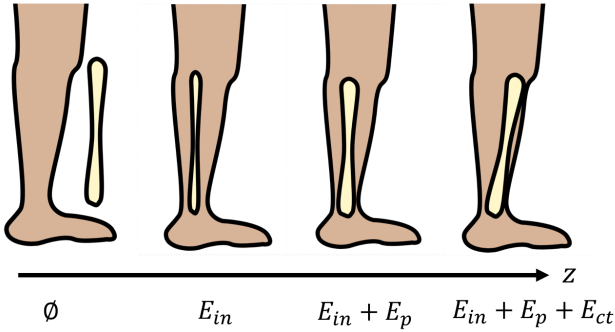


Figure 11. We illustrate the intuition behind the costs on a profile view of the tibia in the leg. From the frontal projected silhouette, there is no constraint for the bone to be inside the body along the  $z$  axis. We use  $E_{in}$  to force it to be inside. Forcing it inside is not enough as it could squeeze and collapse; thus, we enforce the bone to be close to the skin surface with  $E_p$ . 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.

**Skeleton unposing.** In Sec. 3.5 of the main paper, we introduce  $E_d$ , a cost that enforces the conservation of the skeleton to skin distance when changing the pose. In Fig. 12 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 be constant independent of the 3D pose.

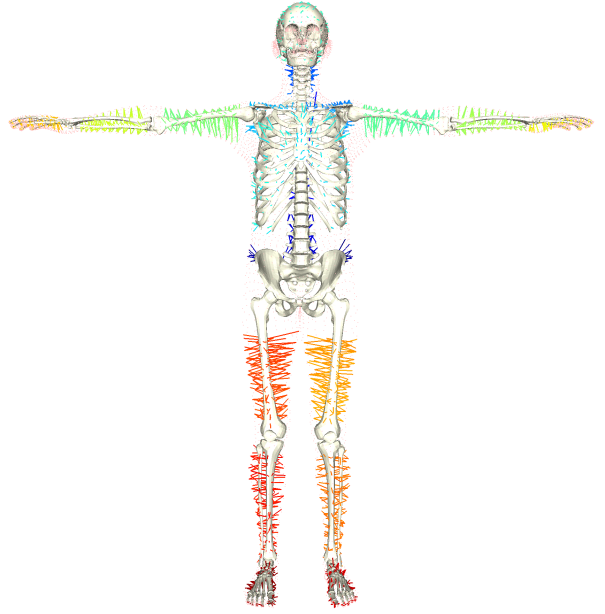


Figure 12. Skin to skeleton pairs used in the cost  $E_d$ . We color the links in each part with a different color for visualization purposes.

**Skeleton reposing.** In Sec. 4.3 of the main paper, we use the costs  $E_j$  and  $E_l$  in the skeleton inference optimization.

The term  $E_j$  models ball joints in the shoulders, elbows and hips. It forces the bone heads to stay in their sockets.



For each articulation, we define vertices  $s_i, s_j$  on the skeleton template that define a joint socket of a bone head. At each optimization step, we fit spheres with centers  $S_i, S_j$  to each groups of vertex and force each of spheres to stay at a similar distance during the optimization:

$$E_j(\mathbf{t}, \mathbf{r}; SP_0) = ||S_i(\mathbf{t}, \mathbf{r}) - S_j(\mathbf{t}, \mathbf{r})|| - d_{s0} \quad (11)$$

This cost is not sufficient to model the knee movement, so we define stitching costs 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  $E_l = ||l_i - l_j|| - d_{l0}$ .

The distances  $d_{l0}$  and  $d_{s0}$  are defined such that  $E_j(\mathbf{t}_0, \mathbf{r}_0; SP_0) = 0$  and  $E_l(\mathbf{t}_0, \mathbf{r}_0; SP_0) = 0$ .

## C. Skeleton shape space evaluation

In section 3.6 of the main paper, we detail how we learn a skeleton shape space from the unposed skeleton meshes. In this section, we present an evaluation of the compactness of the shape space as well as its generalization ability.

### C.1. Variance

To evaluate the compactness of our skeleton shape space, we compute the variance explained by each component of the PCA space. The cumulative variance plot is shown Fig. 13. With 3, 5 and 10 components, the male PCA model respectively captures 91.1%, 94.8% and 97.8% of the skeleton’s variance. The female model respectively 92.7%, 95.6% and 98.1%.

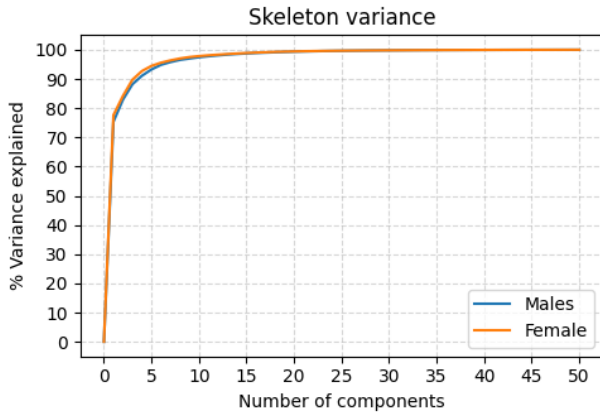


Figure 13. Cumulative variance of the skeleton shape space.

### C.2. Shape space generalisation

We next evaluate how the skeleton shape space generalises to unseen subjects. We compute the skeleton shape space from the training dataset and we evaluate how accurately it can reconstruct 200 left out unposed skeletons. We

Nb components	error (mm) (mean $\pm$ std)	
	Male	Female
3	7.59 $\pm$ 4.79	7.79 $\pm$ 4.86
5	5.55 $\pm$ 3.49	5.14 $\pm$ 3.27
10	3.14 $\pm$ 2.19	3.02 $\pm$ 2.14

Table 4. Skeleton reconstruction error given the number of principal components used. The errors are in millimeters.

project each of the test set meshes onto the first  $N$  basis vectors of the shape space and we reconstruct the bones using only these coefficients.

We then measure how much information is lost in this projection by computing the per-vertex distance between the original mesh and the projected and reconstructed mesh. We aggregate this per-vertex error for each mesh and obtain the errors reported in Table 4.

As we can see, with a small number of components, such as 5, mean errors are below 6 mm. When using 10 components, the reconstruction mean errors are below 4 mm. The created bones shape space can capture the shape of left out subjects with errors below 4 millimeters.

## D. Extended results

This section complements the presented results in Sec. 5 of the main document.

### D.1. Skin alignment qualitative evaluation

In this section we illustrate the alignment results of the STAR model on the DXA images. Those alignments were obtained with the optimization presented in Sec. 3.3 of the main paper. These results complement the quantitative evaluation reported in Sec. 5.1 of the main manuscript, where the intersection over union coefficient between the DXA mask  $M_S$  and the computed skin silhouette is 94% for females and 95% for males. In Figure 14, we show the qualitative results. 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, 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. 15, we show some failure cases with low intersection 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.

### D.2. Skeleton 3D landmarks regression evaluation

In Sec. 4.1 of the main paper, we explain how we train a regressor that, taking as input the vertices of the skin, pre-

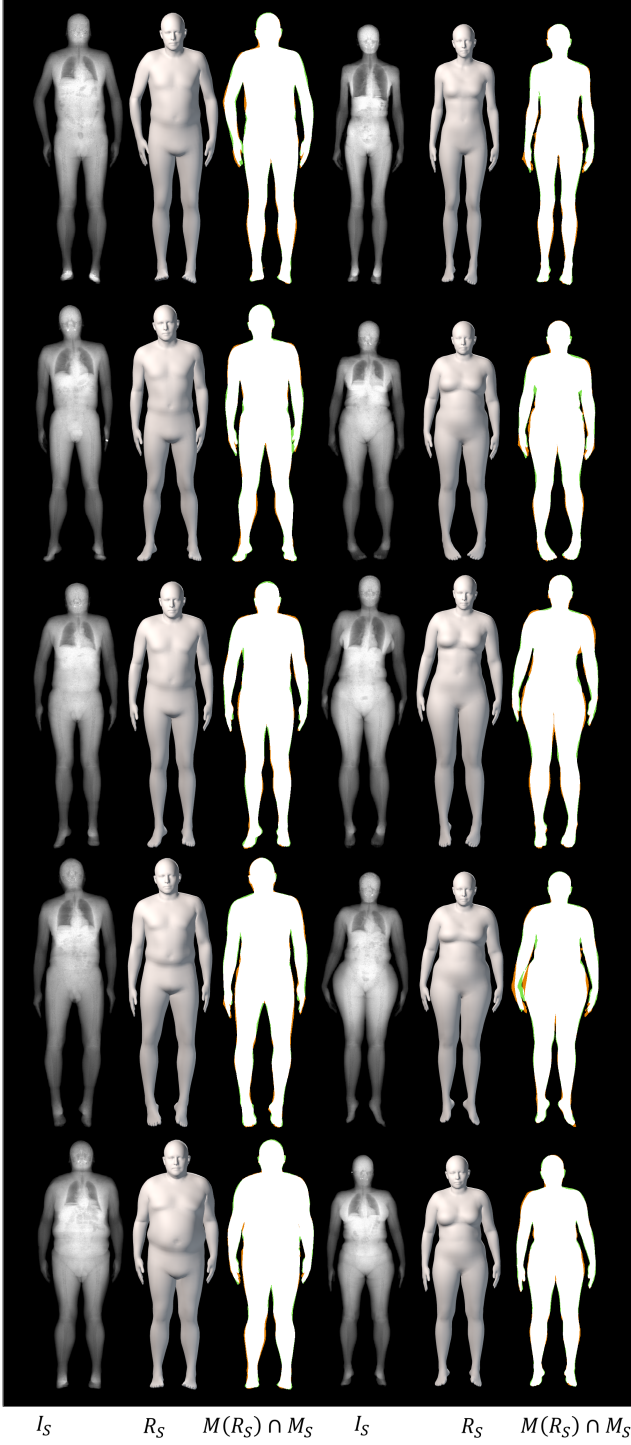


Figure 14. Comparison of the aligned STAR models  $\mathbf{R}_S$  with the target DXA masks  $M_S$  for subjects sampled from the curated dataset. On the left we show males and on the right females. The masks intersection is color-coded as follow: green:  $\mathbf{R}_S$  only, orange:  $M_S$  only, white: both.

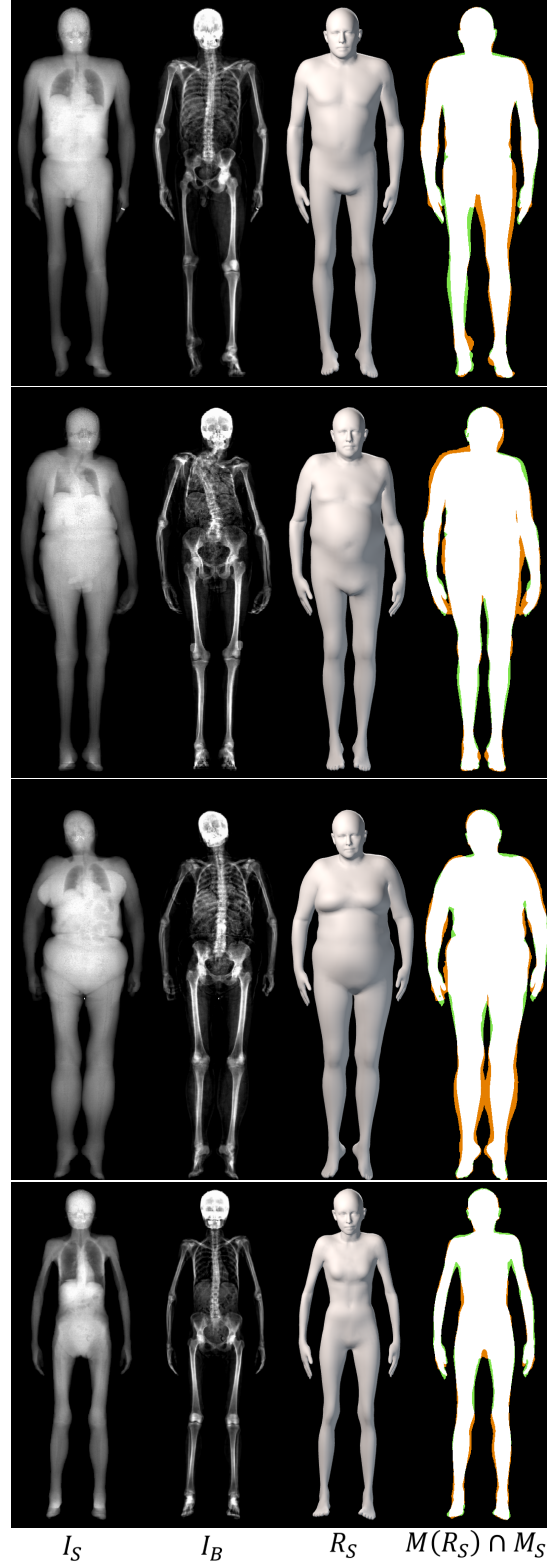


Figure 15. Failure cases. For each subject, we show  $I_S$ ,  $I_B$ , the fitted skin mesh  $\mathbf{R}_S$  and the intersection of both masks. The masks intersection is color-coded as follow: green:  $\mathbf{R}_S$  only, orange:  $M_S$  only, white: both. The STAR model can not faithfully capture the shape of these subjects.

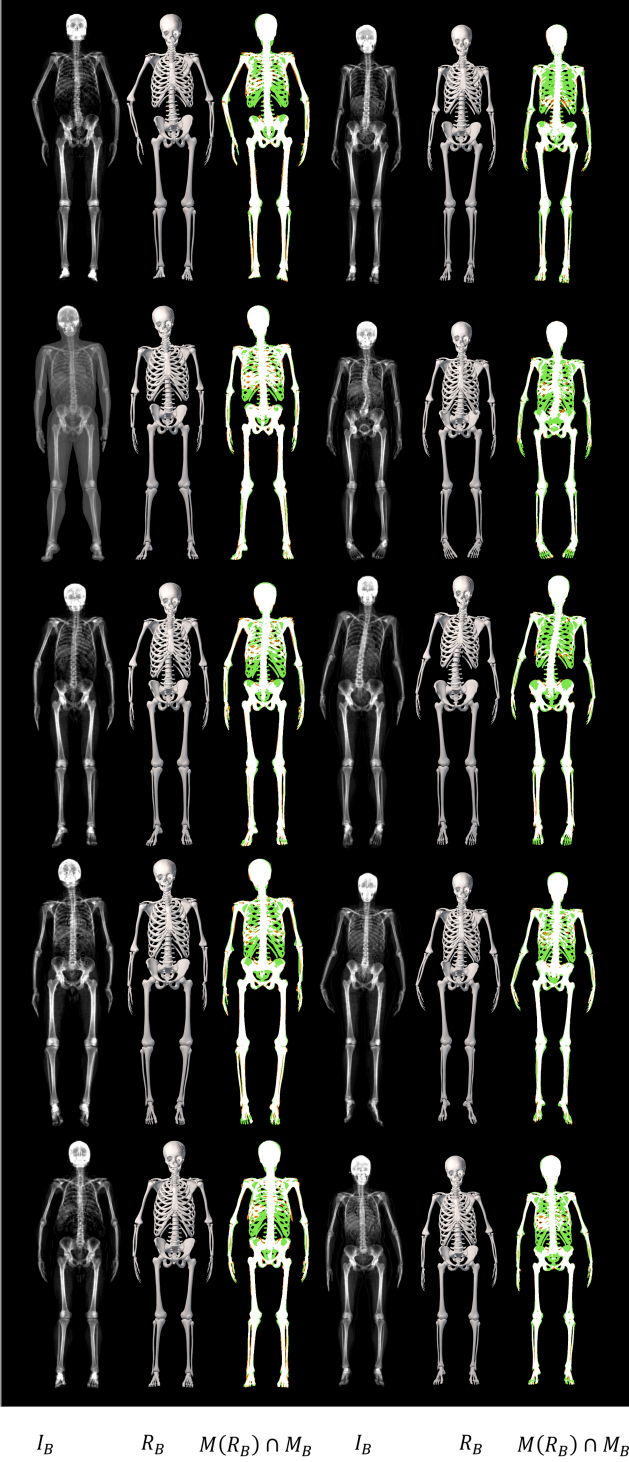


Figure 16. Comparison of the registered skeleton  $R_B$  with the target DXA masks  $M_B$  for subjects sampled from the training dataset. On the left we show males and on the right females. The masks difference is color-coded as follow: green:  $R_B$  only, orange:  $M_B$  only, white: both.

dicts the 3D location of the landmarks  $\mathcal{L}_B$  (presented in Fig. 10 right). This regression is learned in a normalized lying down pose as illustrated in Fig. 17.

To evaluate the  $\mathcal{L}_B$  regressor accuracy, we learn the regressor from the 1000 train subjects and evaluate on the 200 left out subjects. We compute the 3D distance between the regressed landmarks position and its ground truth position. In Sec. 5.2 of the main paper we provide a general evaluation on the accuracy of the regressor as well as a discussion of the results. The detailed per landmark errors are listed in Table 5.

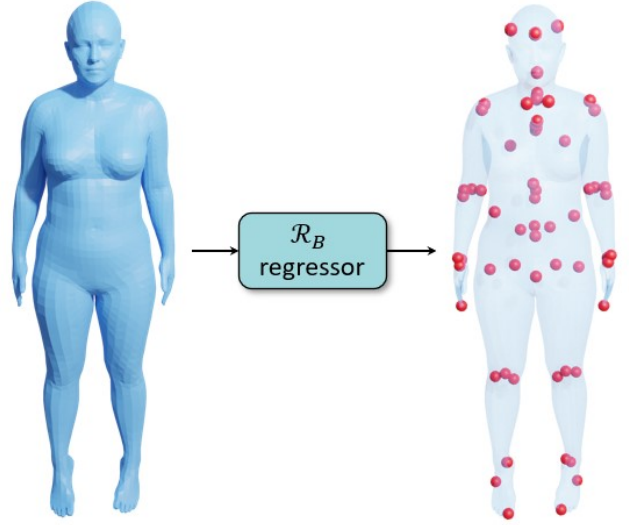


Figure 17. Given a skin mesh, the landmark regressor lets us compute the landmark 3D locations as a linear combination of the skin mesh vertices locations.

### D.3. Skeleton registration qualitative evaluation

Next we show qualitative results of the skeleton registrations  $R_B$  in Fig. 16. The subjects are the same as in Fig. 14. These results complement the Sec. 5.3 of the main document, and precisely, the numeric value reported in the first row of Table 1 in the main document.

### D.4. OSSO VS Anatomy Transfer comparison

In Figure 19, we present a qualitative comparison between our OSSO predictions and the ones from Anatomy Transfer. This results complement Sec. 5.3 of the main document.

From the DXA test set, we select 5 subjects spanning the dataset BMI distribution. From the skin alignment  $R_S$ , 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  $SI$ .

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

Table 5. 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. 18.

As can be seen from the images, our predictions do better capture the global shape of the skeletons. Particularly, Anatomy Transfer often estimates the location of the hips to be too low with respect to the actual hips location. Our method predicts a skeleton which is visually closer to the one observed in the DXA images.

### D.5. Skeleton inference qualitative evaluation

**Lateral view** Fig. 20 shows side views of the inference result in T-pose. While there is no ground truth to evaluate this pose with, the results are plausible.

**Inference on subjects from AGORA [25]** Fig. 21 shows the inferred skeletons for subjects with different shapes and poses.

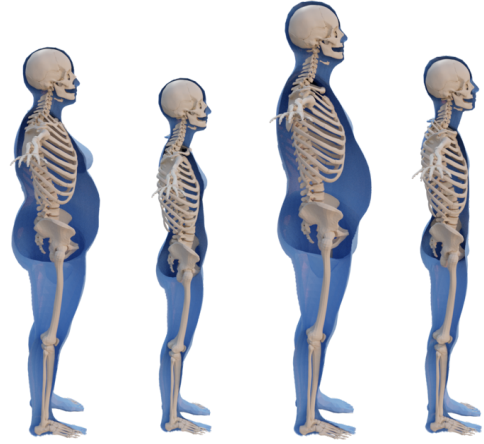


Figure 20. Lateral views of skeletons inferred with OSSO.



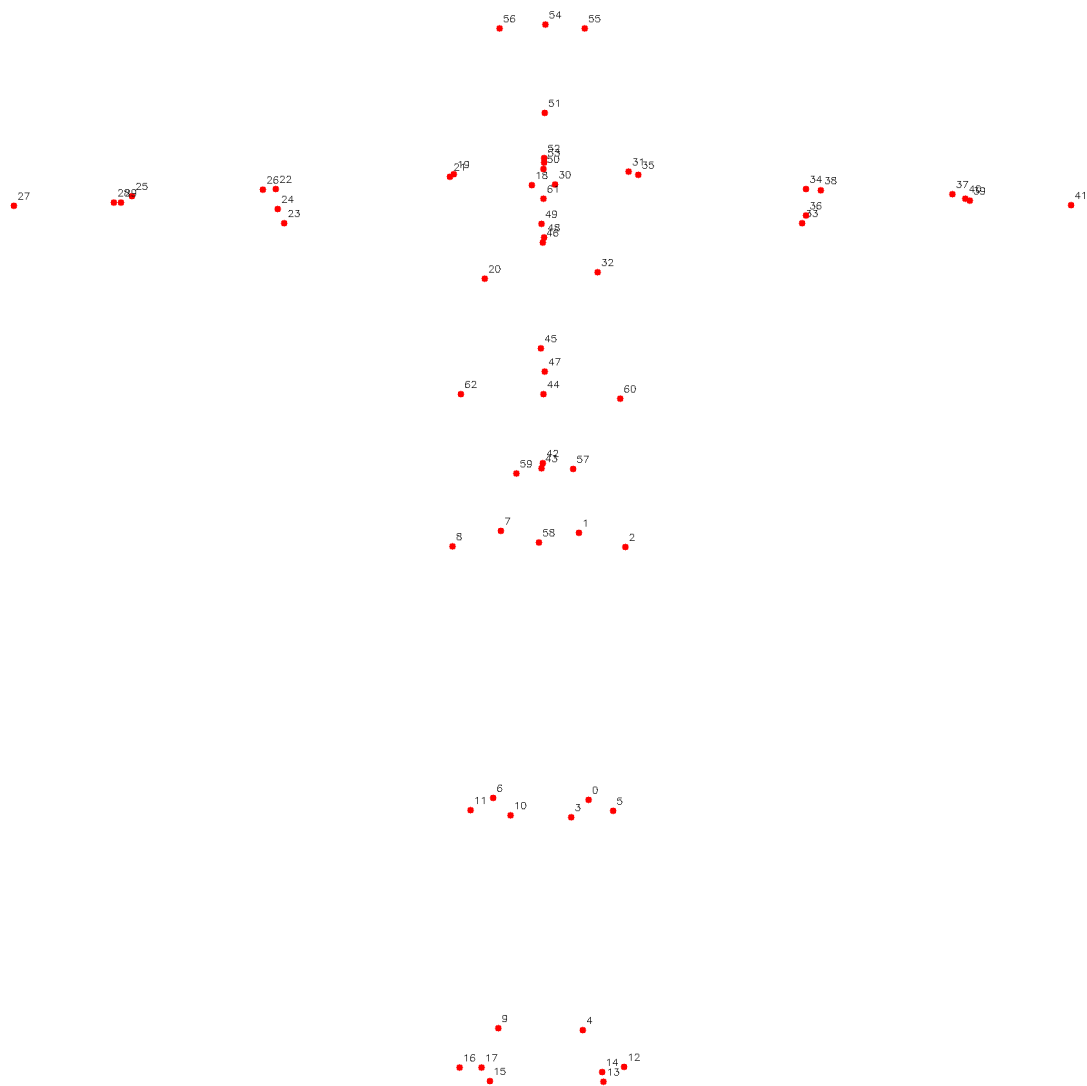
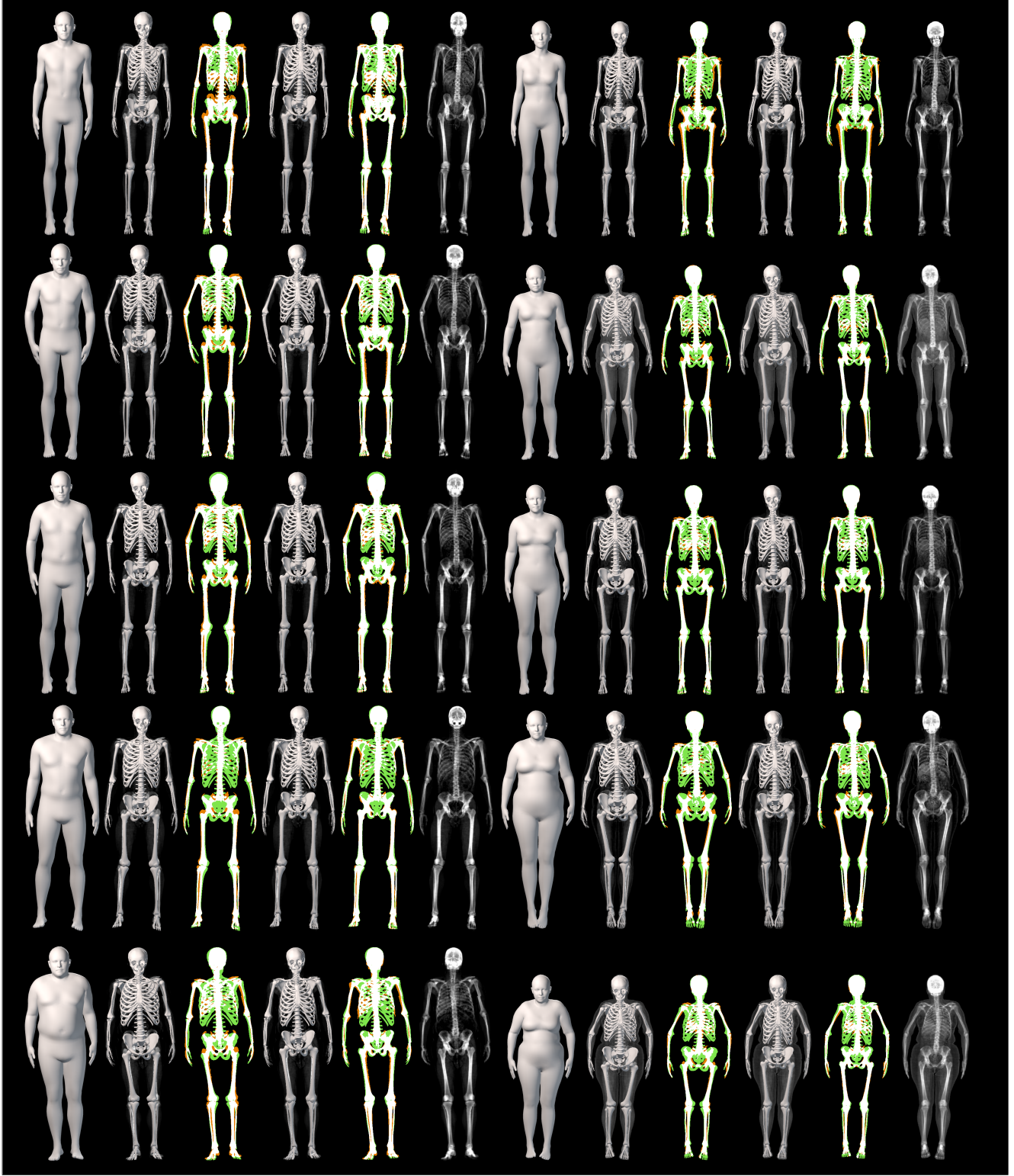


Figure 18. Landmarks  $\mathcal{L}_B$  on the skeleton mesh with landmark number.



$R_S$      $SI_{AT}$  on  $I_B$      $M(SI_{AT}) \cap M_B$      $SI_{OSSO}$  on  $I_B$      $M(SI_{OSSO}) \cap M_B$      $I_B$      $R_S$      $SI_{AT}$  on  $I_B$      $M(SI_{AT}) \cap M_B$      $SI_{OSSO}$  on  $I_B$      $M(SI_{OSSO}) \cap M_B$      $I_B$

Figure 19. For each subject, we show in the order (1)  $R_S$ , (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  $I_B$

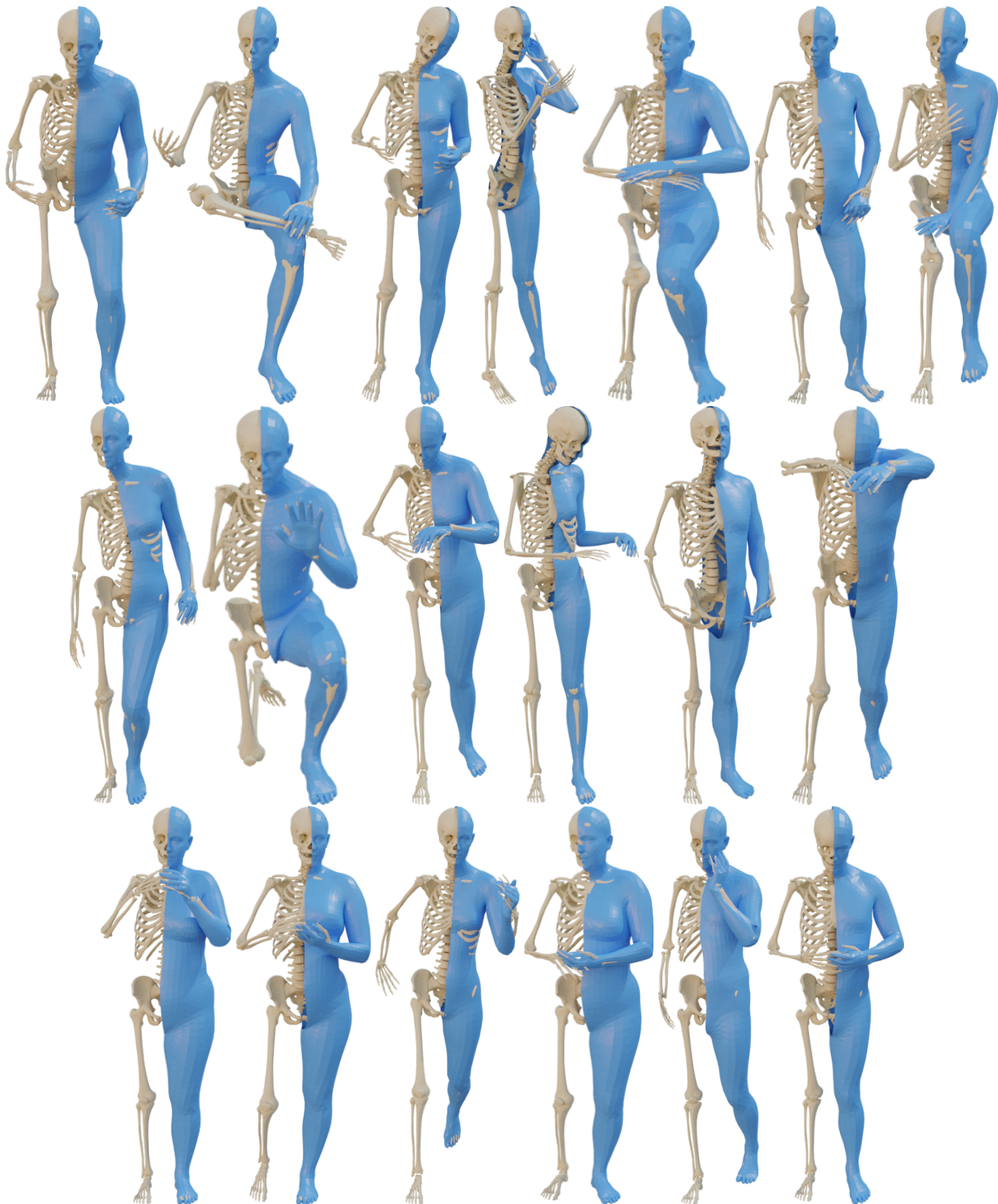


Figure 21. Given SMPL bodies aligned to RenderPeople subjects [25, 28], we use OSSO to infer the underlying skeleton.

## References

- [1] Dicko Ali-Hamadi, Tiantian Liu, Benjamin Gilles, Ladislav Kavan, François Faure, Olivier Palombi, and Marie-Paule Cani. Anatomy transfer. *ACM Transactions on Graphics*, 32(6):1–8, Nov. 2013. [2](#), [3](#), [4](#), [7](#), [9](#)
- [2] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Transactions on Graphics*, 22(3):587–594, July 2003. [3](#)
- [3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, July 2005. [2](#), [3](#)
- [4] Armelle Bauer. *Modélisation anatomique utilisateur-spécifique et animation temps-réel. Application à l'apprentissage de l'anatomie*. Theses, Université Grenoble Alpes, Nov. 2016. [2](#), [3](#)
- [5] 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)*, Lecture Notes in Computer Science, pages 561–578. Springer International Publishing, Oct. 2016. [3](#), [4](#)
- [6] 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. [4](#)
- [7] 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. [3](#)
- [8] 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. [2](#)
- [9] 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. [2](#)
- [10] Benjamin Gilles, Lionel Reveret, and Dinesh Pai. Creating and animating subject-specific anatomical models. *Computer Graphics Forum*, 29(8):2340–2351, Dec. 2010. [3](#)
- [11] Peng Guan, Alexander Weiss, Alexandru Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *International Conf. on Computer Vision (ICCV)*, pages 1381–1388, 2009. [3](#)
- [12] 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. [3](#)
- [13] Alexandru-Eugen Ichim, Petr Kadleček, Ladislav Kavan, and Mark Pauly. Phace: Physics-based face modeling and animation. *ACM Transactions on Graphics*, 36(4):1–14, 2017. [3](#)
- [14] 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. [4](#)
- [15] Petr Kadleček, Alexandru-Eugen Ichim, Tiantian Liu, Jaroslav Krivánek, and Ladislav Kavan. Reconstructing personalized anatomical models for physics-based body animation. *ACM Transactions on Graphics*, 35(6):1–13, Nov. 2016. [2](#), [3](#)
- [16] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. [3](#)
- [17] 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. [3](#)
- [18] Charles L Lawson and Richard J Hanson. *Solving least squares problems*. Society for Industrial and Applied Mathematics, 1995. [6](#)
- [19] Mei Kay Lee, Ngoc Sang Le, Anthony C Fang, and Michael TH 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. [4](#)
- [20] 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. [2](#), [3](#)
- [21] Shawn P McGuan. Human modeling—from bubblemen to skeletons. In *SAE Digital Human Modeling for Design and Engineering Conference*, pages 26–28, 2001. [2](#)
- [22] 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. [4](#), [10](#)
- [23] 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, Aug. 2020. [2](#), [3](#), [4](#), [8](#), [9](#)
- [24] 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. [3](#)
- [25] 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. [7](#), [16](#), [19](#)
- [26] 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 Com-*



- puter Vision and Pattern Recognition (CVPR), pages 10975–10985, 2019. 3
- [27] 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. 3
- [28] RenderPeople. <https://renderpeople.com>, 2020. 1, 7, 19
- [29] Cornelius J. F. Reyneke, Marcel Lüthi, Valérie Burdin, Tania S. Douglas, Thomas Vetter, and Tinashe E. M. Mutsavanga. 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. 3
- [30] Marcel M. Rossi, A. 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. 4
- [31] Shunsuke Saito, Zi-Ye Zhou, and Ladislav Kavan. Computational bodybuilding: Anatomically-based modeling of human bodies. *ACM Transactions on Graphics*, 34(4):1–12, 2015. 3
- [32] 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 *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, pages 25–36, 2021. 3
- [33] 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. 2, 6, 8
- [34] 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. 4
- [35] 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. 3
- [36] 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 Medicine*, 12(3):e1001779, 2015. 2, 7
- [37] 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. 3
- [38] Bohan Wang, George Matcuk, and Jernej Barbič. Hand modeling and simulation using stabilized magnetic resonance imaging. *ACM Transactions on Graphics*, 38(4):1–14, July 2019. 3
- [39] 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. 3
- [40] 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. 4
- [41] 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. 2, 3
- [42] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *arXiv:2012.13392*, 2020. 3
- [43] Lifeng Zhu, Xiaoyan Hu, and Ladislav Kavan. Adaptable anatomical models for realistic bone motion reconstruction. *Computer Graphics Forum*, 34(2):459–471, 2015. 2, 3
- [44] Gaspard Zoss, Thabo Beeler, Markus Gross, and Derek Bradley. Accurate markerless jaw tracking for facial performance capture. *ACM Transactions on Graphics*, 38(4):1–8, July 2019. 3
- [45] 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. 4, 5, 6, 11
- [46] 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. 3, 4, 5
- [47] 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. 5