

AnthroNet: Conditional Generation of Humans via Anthropometrics

Francesco Picetti* Shrinath Deshpande* Jonathan Leban* Soroosh Shahtalebi† Jay Patel†
 Peifeng Jing Chunpu Wang Charles Metze III Cameron Sun
 Cera Laidlaw James Warren Kathy Huynh River Page
 Jonathan Hogins Adam Crespi Sujoy Ganguly§ Salehe Erfanian Ebadi§

Unity Technologies

Abstract

We present a novel human body model formulated by an extensive set of anthropocentric measurements, which is capable of generating a wide range of human body shapes and poses. The proposed model enables direct modeling of specific human identities through a deep generative architecture, which can produce humans in any arbitrary pose. It is the first of its kind to have been trained end-to-end using only synthetically generated data, which not only provides highly accurate human mesh representations but also allows for precise anthropometry of the body. Moreover, using a highly diverse animation library, we articulated our synthetic humans' body and hands to maximize the diversity of the learnable priors for model training. Our model was trained on a dataset of 100k procedurally-generated posed human meshes and their corresponding anthropometric measurements. Our synthetic data generator can be used to generate millions of unique human identities and poses for non-commercial academic research purposes¹.

1. Introduction

With the rise of digital worlds, and the unprecedented need to replicate and represent digital copies of real-world objects, there has been a significant surge of interest in developing methods for representing objects in 3D. These 3D representations are particularly important for human-centric computer vision and graphics problems like person tracking, human action recognition, and virtual try-on applications. Advanced human body models representing the most extensive range of possible human physiques, expressions, and motions are critical for these applications. Seminal works such as the SMPL(-X) body model [33, 43] and

the subsequent improvements and extensions [31, 38, 48] played a critical role in creating a realistic animated human body model that can represent many different body shapes. Later on, GHUM(L) [59] created an accurate human body model based on a set of deep generative models that each encapsulates variations on the human body, hands, head, and pose. One major limitation of the preliminary works on human shape modeling is their reliance on dense body scans that often require expensive motion capture and scan setups, which prohibit data collection and annotation at large scales. Thus, there has been a growing interest in developing models where the body shape is regressed from monocular images [25–27, 62].

Recently, SHAPY [11] presented a new image-based paradigm, where a combination of linguistic shape attributes and some basic anthropometric measurements – i.e., height, weight, chest circumference, waist circumference, and hips circumference – to condition the regression of the SMPL-X model parameters from monocular RGB images. Although SHAPY outperforms its counterparts in regressing a 3D shape from monocular images, deploying linguistic shape attributes reported subjectively with no ground truth for validation can introduce noise to the model. Moreover, their limited set of anthropometric measures only defines the body's overall shape and limits the degrees of freedom needed to represent a wide variety of shapes.

Inspired by SHAPY, we present a novel non-linear human body model conditioned on an extensive set of objective anthropometric measures that allows us to represent a wide range of human body shapes. Our proposed generative model AnthroNet is trained end-to-end on posed human body meshes (Fig. 1) and their associated anthropometric measures. To train AnthroNet, we procedurally generated 100,000 multi-subject multi-posed human meshes. This dataset includes the mesh and pose data and the anthropometric measures of the meshes. Finally, we provide an anthropomorphic measurement regressor, which after registration with AnthroNet, allows us to easily extract anthro-

*.†Equal contribution.

§Joint corresponding authors.

¹<https://unity-technologies.github.io/AnthroNet/>

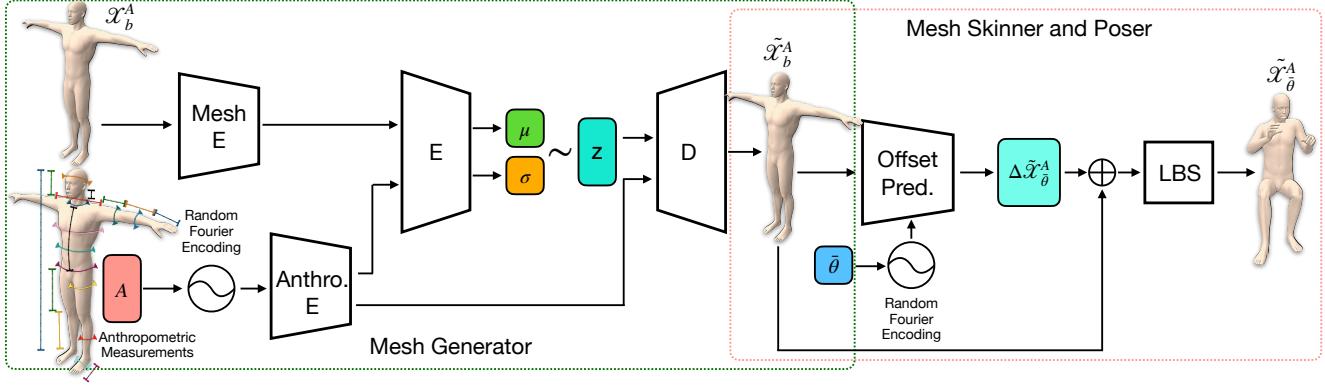


Figure 1. AnthroNet’s end-to-end trainable pipeline. We discard the Mesh Encoder in the Mesh Generator block at inference time. Then we decode the estimated mesh in the bind pose \tilde{x}_b^A using a latent vector Z sampled from a normal distribution combined with the encoded anthropometric measurements A . The mesh Skinner and Poser block predicts the pose and shape corrective offsets $\Delta\tilde{x}_{\bar{\theta}}^A$ required to animate the mesh into the desired pose $\bar{\theta}$ and produce a fully skinned, rigged, and posed mesh $\tilde{x}_{\bar{\theta}}^A$ with the provided measurements A .



Figure 2. Examples from our multi-subject and multi-pose synthetic dataset.

pometric measurements from any real-world scan or other body models without the need to define mesh-specific methods to extract the anthropometric measures carefully. In summary, our contributions are:

- AnthroNet provides a novel expressive high resolutions model of human body shape conditioned on anthropomorphic measurements.
- AnthroNet offers two efficient registration pipelines that enable a bi-directional conversion of SMPL and SMPL-X meshes into AnthroNet meshes.
- Our synthetic data generator which can produce high resolution meshes with their associated metadata is publicly available for non-commercial academic research use.

2. Related Work

The field of human body modeling has seen a rich body of work on 3D articulated mesh reconstruction for the whole body, including hands and face [4, 5, 17, 21, 23, 32, 33, 38, 42, 43, 59] and individual body parts alone for a finer level of expressiveness and control [31, 37, 46, 48, 53]. More recently neural parametric models [39] and implicit parametric models [3, 9, 40, 58] for 3D deformable shapes have been proposed which leverage the flexibility of the learned implicit models that model the body as continuous functions, where the surface is represented as the level set of a neural network that is defined on the 3D space [41].

Most of these techniques relied on dense point clouds obtained from 3D scans or used an already trained parametric model of the human body to reconstruct the 3D shape from images, videos, motion capture data, or other modalities.

ties. To date, SMPL and SMPL-X remain the most widely used body models for their simplicity and ease of integration [1, 2, 7, 21, 25–28, 35, 36, 45, 47, 49, 61, 63]. More recently, GHUM and GHUML [59] proposed an end-to-end trainable deep architecture that allows for higher resolution complete body modeling with 10,168 and 3194 mesh vertices respectively. However, there is no mechanism by which a specific body shape from the embedding of their trained model can be sampled and reconstructed. In contrast, our method allows for such sampling, as the body shape information is used as a conditioning augmentation vector in our generative pipeline. Since we trained our model on synthetic data, we opted for the highest resolution mesh representation with 36,078 mesh vertices, more than 3 \times that of GHUM. This higher-resolution representation allows for more expressive facial, hand, and body shapes. Using synthetic data for training means we do not need to rely on mesh registration pipelines from 3D point clouds [20, 30]. Instead, we obtain perfect mesh ground truth, which helps with consistent and error-free modeling of articulations on the mesh.

A large body of work has been dedicated to 3D pose and shape estimation from images and videos [1, 2, 7, 21, 22, 25–28, 43, 62]. Most techniques rely on SMPL and its variants due to the simplicity of the PCA priors for modeling the human body. More advanced dimensionality reduction techniques have also been proposed based on Gaussian Mixture Models [7] and Variational Autoencoders (VAEs) [43] that allow for higher accuracy and generalization to more poses. GHUM [59] combined VAEs and normalizing flow representations for the body and facial deformation and the skeleton kinematics, respectively. In our technique, we use a single conditional VAE (CVAE) that allows for anatomically consistent and robust pose and shape modeling while enabling the users to define strict priors for the body representation. This opens a new avenue for technical artists and other less artistically inclined users to take advantage of a powerful creation tool that can statistically represent high-resolution 3D articulated human body shape and pose in a single learned embedding with fast inference.

In SHAPY [11], sparse anthropometric measurements and linguistic attributes, *e.g.*, “short/tall”, “long legs”, “pear-shaped”, were used as constraints for 3D shape prediction. The sparsity of those measurements and the natural ambiguity that arises from using linguistic attributes to describe a human body shape only provide coarse and global body proportions and embedded correlation in proportionality and shape present in the SMPL training data. For example, a “muscly” person may produce a person whose entire body is proportionally large. In contrast, people with a larger upper body than a lower body can exist. Our model fills this knowledge gap as it was trained with dense anthropometric measurements covering the whole human body.

The following sections introduce our synthetic data generation pipeline and describe our human body model.

3. Synthetic Data Pipeline

Conventionally, models have been trained on collections of registered meshes with pre-defined topologies onto real scans of actors in motion capture studios. This process is limited by the number and diversity of people scanned and the set of actions they can perform. Additionally, this data suffers from small-scale registration errors. Some efforts have been made to provide human-understandable textual descriptions for human bodies [11]. However, they are still far from providing accurate, consistent, and true-to-life measurements of the subjects. Such scans also lack the diversity of human bodies they represent, as clothing models are not representative of the entire population.

Synthetic data can generate millions of unique humans with perfect measurements and annotations, providing more diversity and accuracy compared with their real data counterparts. Prior work has demonstrated how diverse and perfectly labeled synthetic data can improve model performance on large benchmark tasks [13, 14, 56]. Access to the mesh in a game engine enables us to take accurate and consistent anthropometric measurements of our characters, just like an experienced tailor would in reality. These accurate anthropometric measurements facilitate training a single model representing all humans, children, and adults, with all body variations and types. The same anthropometric measurements used as conditioning augmentation can help us obtain more disentanglement for body representations in the learned manifold of our model, directly resulting in training a much more expressive body model.

A team of technical artists hand-sculpted extreme ends of the body morphology and a few intermediary states to create a procedural pipeline capable of generating such diverse labeled data. Heavy and lean characters were hand sculpted in ZBrush for males and females separately while respecting the realistic physical constraints of the human body. Then using blend shape logic in Houdini, intermediary states can be obtained for characters of varying weight. The same was done for age, height, ethnicity, and sex. They were then imported into the Unity game engine as Vertex Animated Textures (VAT), which stores the blendshapes in a storage-friendly texture format for use in the Unity Synthetic Humans package. This enables us to sample from the entire set of possible humans randomly. As the blend shape logic can vary the mesh vertex positions by infinitesimally small real-valued deltas, our characters’ search space is practically infinite. The resulting meshes are high-poly representations with 36,078 mesh vertices and 68,414 mesh faces (triangles). In addition, technical artists hand-drew the appropriate skinning weights for male and female characters, valid across the entire set of produced meshes. It

can deform well to a large realistic motion capture dataset. We later use these same skinning weights along with the joint-vertex correspondences in our mesh skinning and posing pipeline to re-create a fully rigged, skinned, and posed human character given an unseen input set of anthropometric measurements and desired pose.

4. Data Preparation

Using our synthetic data generator, by random sampling from a uniform distribution along each axis of variation, e.g., sex, age, height, ethnicity, weight, and facial identity, we generated a total of 100,000 distinct human body shapes that are used in our training and validation pipeline. Each human is generated in bind pose (T-pose) and a random pose, where the latter is randomly sampled from a large dataset of 1778 diverse and complex animation clips that are created from real motion capture sequences performed by multiple subjects. Since our generated characters’ morphology may widely diverge from those of the actors in the motion capture sequences, some self-intersection is inevitable due to the simple re-targeting pipeline we used. However, character animation and re-targeting are outside the scope of this work, and the posed characters produce sufficient results for training a human body model. The Unity Perception package [8] provides highly accurate 3D keypoint locations and rotations for our characters. In our mesh generation pipeline, we rely on 3D vertex locations on the mesh, represented in the local character space with the root joint (hip) at the origin. Furthermore, we use the mesh faces (triangles) in the optimization to enforce surface smoothness. The 3D keypoint locations (joint locations) in the posed mesh are used in our mesh poser pipeline.

5. Methodology

Our end-to-end trainable pipeline comprises a conditional variation autoencoder (CVAE) for human body modeling and mesh generation and a deep mesh rigging, skinning, and posing component. Our mesh generation component produces a human mesh in bind pose, given a set of anthropometric measurements as conditioning vectors, each of which we represent as Fourier encodings. The second component takes a mesh in bind pose and produces the vertices offsets relative to the desired pose. We obtain the final posed mesh by combining such offsets and the bind pose vertices through a skinning function.

5.1. Anthropometric Measurements

Anthropometric measurements consistently and accurately represent information for many human body shapes and sizes. Extensive studies have been devoted to qualitative and quantitative measurement of human populations [6, 15, 16, 19, 54, 57, 60]. We defined a dense set of 36

anthropometric measurements over the human body, representing the height and weight of the human, characterized by measurements over the head, neck, shoulders, torso, chest, waist, hips, arms, hands, thighs, legs, and feet. For each measurement, a technical artist has defined a set of vertices over which the measurement is made, such that it consistently represents a tape measure in real life as close as possible. The body part lengths are the euclidean distances between specific vertices on the mesh. By registering any 3D human scanned data or human mesh to its closest approximation with our human body model, we can provide the anthropometric measurements of any 3D human scan similarly to [54]. Since our synthetic human dataset was created with true-to-life scales and proportions, our anthropometric measurements provide close to actual real-life metric body size measurements. Leveraging such information for all existing human scan data is a valuable contribution that we shall leave to future work, due to time constraints. Furthermore, whilst we opted for the 36 anthropometric measurements over our meshes – which provide the maximal meaningful information about all our meshes – obtaining more measurements on every edge loop on the synthetic meshes is a straightforward task.

5.2. Random Fourier Encoding

Fourier encoding [52] is an approach that maps continuous input coordinates into higher dimensional space to help recover high-frequency details from the input data. We deal with 3D mesh representations encapsulating position-dependent information; further, the exact anthropometric measurements over body regions could come from two slightly different human body shapes; also, each measure has different scales; therefore, the Fourier encoding of such input data seems to be a natural choice. For us, this directly translated to much higher fidelity in the preservation of geometry and further disentanglement of multiple highly correlated anthropometric measurements from one another.

We apply our Fourier encoding on our anthropometric measurements and the pose input θ to the network in the form of global locations of body joints. For each input point y and given a random Gaussian matrix B , whose entries are drawn independently from a normal distribution $\mathcal{N}(0, \sigma^2)$ we encode our input as:

$$\mathcal{F}(y) = [\cos(2\pi By) \quad \sin(2\pi By)]^T, \quad (1)$$

which is a deterministic mapping and can be computed feasibly before training.

5.3. Conditional Human Mesh Generation

We formulate the human mesh generation objective as learning the conditional distribution of observed human meshes with respect to their corresponding anthropometric measurements. We employ a conditional latent vari-

able generative model. The high-level architecture of this model was based on Conditional Variational Autoencoder (CVAE) [24] as shown in Fig. 1. Forward pass of the CVAE architecture starts with an encoder that takes in the input as a tuple of vectors, namely human mesh in bind pose \mathcal{X}_b^A , and a vector of conditionals A which are the Fourier encoded anthropometric measurements. Each of the vectors are passed through separate sub-encoding modules which are fused later to produce a probability distribution $p(z)$ for embedding, similar to VAE. A sample is taken from this probability distribution, along with encoded conditional vector at the decoder stage to produce an output mesh $\tilde{\mathcal{X}}_b^A$. We train this generative model on the total loss given by

$$\mathcal{L}_G = \alpha \mathcal{L}_{\text{rec}} \left(\mathcal{X}_b^A, \tilde{\mathcal{X}}_b^A \right) + \beta \mathcal{L}_{\text{KL}}(z) + \gamma \mathcal{L}_{\text{Lapl}} \left(\tilde{\mathcal{X}}_b^A \right), \quad (2)$$

where \mathcal{L}_{rec} and \mathcal{L}_{KL} are the reconstruction and Kullback-Leibler divergence losses similarly to the VAE loss. Specifically, \mathcal{L}_{rec} is the Huber distance between the generated mesh $\tilde{\mathcal{X}}_b^A$ and the target mesh \mathcal{X}_b^A :

$$\mathcal{L}_{\text{rec}} = \begin{cases} \frac{1}{2} \left(\mathcal{X}_b^A - \tilde{\mathcal{X}}_b^A \right)^2 & \text{if } \left| \mathcal{X}_b^A - \tilde{\mathcal{X}}_b^A \right| \leq 1 \\ \left| \mathcal{X}_b^A - \tilde{\mathcal{X}}_b^A \right| - \frac{1}{2} & \text{otherwise.} \end{cases} \quad (3)$$

\mathcal{L}_{KL} quantifies the divergence between the latent space distribution $p(z)$ and a reference distribution $\mathcal{N}(0, 1)$:

$$\mathcal{L}_{\text{KL}} = \sum_{z \in \mathcal{Z}} p(z) \log \frac{p(z)}{\mathcal{N}(0, 1)}. \quad (4)$$

We include a Laplacian loss that penalizes high frequency variations on the reconstructed mesh surfaces:

$$\mathcal{L}_{\text{Lapl}} \left(\tilde{\mathcal{X}}_b^A \right) = \sum_{v=1}^V \sum_{n \in \mathcal{R}(v)} \|\ell_{v,n} (\tilde{x}_v^A - \tilde{x}_n^A)\|_2^2, \quad (5)$$

where V is the number of vertices in the mesh, the ring $\mathcal{R}(v)$ is the set of neighboring vertices to the v -th vertex and $\ell_{v,n}$ are cotangent-based Laplacian weights.

We performed ablation experiments where random Fourier encoding on the conditionals before passing them to the encoder yielded the best results on our test set. During inference, the mesh encoder is discarded, and we only use decoder side of the model, which takes in the encoded anthropometric information and a latent vector and produces a sample from the conditional distribution given by

$$\tilde{\mathcal{X}}_b^A = D(A, z). \quad (6)$$

5.4. Deep Mesh Rigging, Skinning, and Posing

Many popular human body models [33, 59] and 3D Digital Content Creation (DCC) applications use some form

of Linear Blend Skinning (LBS) or Dual Quaternion Skinning in order to bind the mesh to the skeleton and animate the mesh along with the skeleton. We opted to create a body model that is compatible with other graphics software and can be used as a drop-in placement in applications that use the same skinning pipeline. In our end-to-end trainable pipeline, we use a variant of LBS, where the skinning weights $\omega_{v,j}$ for each vertex v and joint j , are provided to us from the DCC application and are consistent across all body sizes and poses. In particular, the skinning weights are differently computed for male and female meshes.

For each character, we use the global joint locations θ from the bind pose and the global joint locations $\bar{\theta}$ from the current arbitrary pose which are calculated from the global mesh vertex positions in the bind pose and current pose, respectively. Then given the pose and shape-dependent corrective offsets $\Delta \tilde{\mathcal{X}}_\theta^A$ that are learned in our deep pose corrective module, we are able to skin the mesh using the modified LBS formulation. As such, each vertex on the posed mesh can be formulated as follows in homogeneous coordinate system:

$$x_v^A = \omega_{v,j} T_j(\theta) T_j(\bar{\theta})^{-1} \begin{bmatrix} \tilde{x}_v^A + \Delta \tilde{x}_{v,\theta}^A \\ 1 \end{bmatrix} \quad (7)$$

$$T_j(\theta) = \begin{bmatrix} 1 & \theta \\ 0 & 1 \end{bmatrix} \in SE(3), \quad (8)$$

where $T_j(\theta)$ is the global transformation matrix for joint j . Then we compute the transformation from bind pose to posed mesh by multiplying the inverse of the global transformation matrix at bind pose θ and the global transformation matrix at current pose $\bar{\theta}$.

Prior works [33, 59] have relied on learning pose-corrective and shape-corrective blend shapes separately. In contrast, we learn both the pose and shape-dependent corrective offsets in a single model. This allows us to learn multi-subject and multi-pose dependent corrective offsets in the hidden space of our mesh skinner regression model. The mesh skinner model takes as input the predicted bind pose mesh $\tilde{\mathcal{X}}_b^A$ from our mesh generator, given the anthropometric measurements A . It then fuses them with the current pose $\mathcal{F}(\bar{\theta})$ which are encoded with our random Fourier encoding \mathcal{F} . The mesh skinner is then asked to predict the pose and shape-dependent corrective offsets $\Delta \tilde{\mathcal{X}}_\theta^A$ that can produce the skinned and posed mesh $\tilde{\mathcal{X}}_\theta^A$ in the LBS formulation above. To this end, the loss function which the skinner is trained on is

$$\mathcal{L}_S = \rho \mathcal{L}_{\text{rec}} \left(\Delta \mathcal{X}_\theta^A, \Delta \tilde{\mathcal{X}}_\theta^A \right) + \eta \mathcal{L}_{\text{rec}} \left(\mathcal{X}_\theta^A, \tilde{\mathcal{X}}_\theta^A \right) + \lambda \mathcal{L}_{\text{Lapl}} \left(\tilde{\mathcal{X}}_\theta^A \right), \quad (9)$$

where the first term is the Huber distance between the desired offsets and the network output; the second term is the

Huber distance between the desired and reconstructed posed meshes; and the third term constrains the optimization to obtain smooth meshes.

5.5. Mesh Registration Pipeline

Mesh registration is an optimization routine that we use at different stages of our pipeline in order to convert body meshes from different models to each other. Due to the popularity of SMPL-X, we developed the following two pipelines. The reader is referred to the appendix for further details on the qualitative results.

AnthroNet-to-SMPLX Here the goal is to build an AnthroNet mesh that fits a given SMPL-X mesh, so as to compare the expressiveness of AnthroNet to SMPL-X. We define c and θ as the learnable variables which will be updated through a gradient-based optimization. Here $c \in \mathbb{R}^{37}$ represents concatenation of sex with the list of anthropometric measures, and θ denotes the vector of joint locations. The mesh generator and the pose module, both with frozen parameters, take c and θ as inputs. We minimize the Chamfer loss between the generated posed mesh and the target SMPL-X mesh, and the gradients are calculated accordingly to update the learnable parameters.

SMPLX-to-AnthroNet This pipeline follows a similar logic as in AnthroNet-to-SMPLX but with the opposite goal of converting the SMPL-X meshes to AnthroNet. The major difference here is that prior to calculating the Chamfer loss over the vertices, we first minimize the loss on the joint locations until it reaches a certain range; afterwards the Chamfer loss comes into effect for the optimization process.

6. Experiments and Results

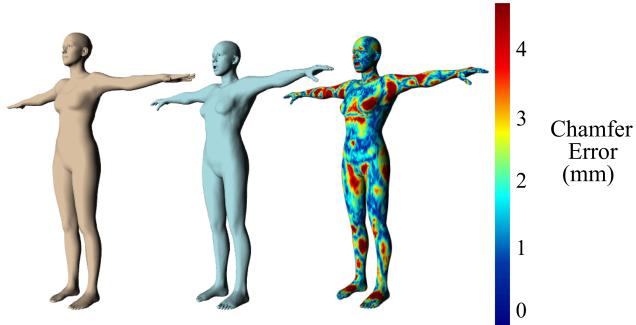


Figure 3. **Illustration of Registration Error when AnthroNet is fitted to a SMPL-X mesh.** A large portion of error attributes are due to the fact that SMPL-X meshes are registered on humans wearing tight clothing, whereas our synthetic humans are without clothes.

We evaluate the performance of AnthroNet as a whole and its individual modules, namely Mesh Generator, Poser,

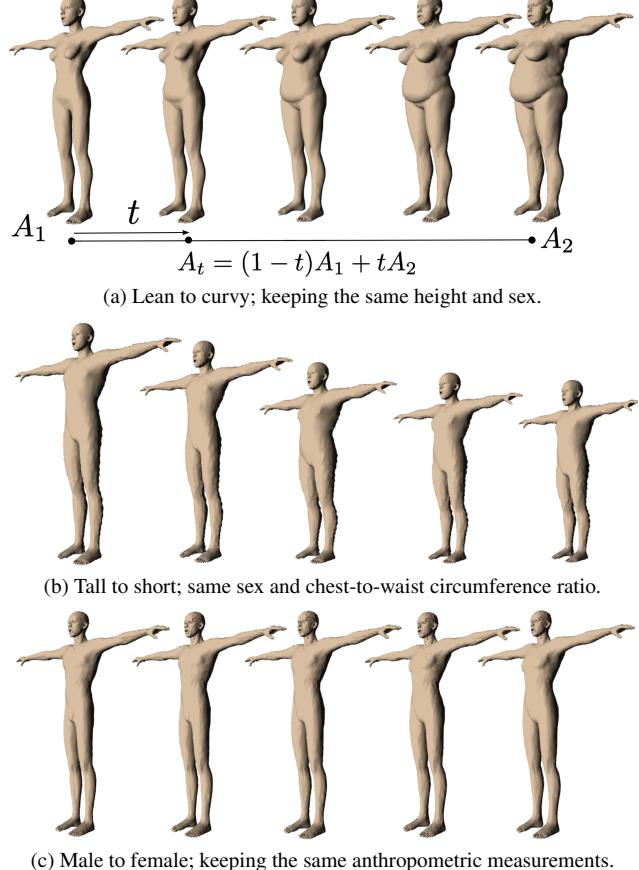


Figure 4. **Human meshes generated by AnthroNet conditioned on measurements and sex.** As depicted in (a) a mesh is generated using conditional A_t , which is obtained from the linear interpolation between two measurement vectors (A_1, A_2).

and the Registration module. We have also performed an extensive set of supplementary experiments and ablations which are provided in the appendix. Throughout this section, the Mean Point-to-Point Error (P2P) [11] is reported as the error between two meshes.

Generative Performance The Mesh Generator is tasked with producing a T-posited human mesh that corresponds with a set of anthropometric measurements. We first evaluate this module’s performance on our held out test set produced by our synthetic data generator. We use this criteria to determine the best conditional mesh generator on the synthetic data. This was done by conducting ablation studies to understand the impact of different components in the compute graph (for details see appendix). The ablation studies indicated that random Fourier encoding lead to better conditional generation as summarized in Tab. 1.

Mesh Registration Performance The SMPL-X body model is widely used in the literature; thus we perform registration on 1185 SMPL-X meshes obtained from SHAPY

fashion model dataset [11]. Since our human mesh contains significantly more vertices, we re-sample vertices on the SMPL-X mesh such that they are equal to the vertices in our model.

Tab. 2 presents statistics of registration results, which indicate a good overall fit based on Euclidean Chamfer distance between our model’s generated mesh and the re-sampled SMPL-X mesh target. As it can be seen, there is a high amount of error in the chest area particularly for female meshes; this is due to the fact that scans of female subjects used to train the SMPL-X model were wearing tight clothing and sports bras. Fig. 3 depicts visual error analysis on the registration, highlighting zones with topological differences in the modeling.

Interpretability of Conditionals Having trained with precise anthropometric data as conditionals, AnthroNet has interpretable and precise control over generation of human meshes. We demonstrate this ability in Fig. 4, where only certain anthropometric measurements are varied to generate a desired human shape using equation (6). The meshes shown are computed by taking two conditional points in 37 dimensional space (36 measurements plus sex). The sample points are selected such that they describe the desired qualitative aspects of a person; e.g., a lean female of a certain height. Then a measurement vector is computed as the linear interpolation between the two sets. Fig. 5 shows more smooth interpolations in our learned shape manifold where the same pose is applied on all generated meshes. Here, the conditional vector A is obtained as linear interpolation between conditionals (A_1, A_2, A_3, A_4).

Regressing Body Shape from Monocular Images To compare the performance of the proposed AnthroNet with its counterparts on regressing the 3D shape from monocular images, we employ the Human Bodies in the Wild (HBW) dataset [11]. HBW contains 3D body scans of 35 subjects (15 male and 20 female) along with their registered SMPL-X body models and a total of 2543 photos of the subjects. We deploy the image-regression module in SHAPY as a pretrained model with frozen parameters. The goal is to register AnthroNet meshes to SMPL-X meshes with high fidelity and then evaluate the accuracy of the measured anthropometrics on the SMPL-X mesh.

The image-regression module takes RGB images as input and outputs an estimate of the SMPL-X shape parameters. To lay a baseline for our comparisons, we apply the measurement module of the SHAPY framework to estimate a sparse set of anthropometric measurements based on predicted SMPL-X parameters. Afterwards, these parameters are passed through our SMPLX-to-AnthroNet pipeline and obtain the corresponding AnthroNet meshes, where our dense set of anthropometric measurements can be calculated. Finally, we find the correspondence between the

Hyper Parameters	P2P	Chest	Hip	Waist
Full Anthropometrics, Sex, $\gamma = 0$	3.83	46.9	9.79	12.3
Full Anthropometrics, Sex, $\gamma = 0.01$	5.55	90.7	38.7	17.6
Height, Weight, Sex, $\gamma = 0.01$	6.55	79.4	27.9	21.3
Height, Sex, Select Circumferences, $\gamma = 0.005$	6.75	41.8	14.2	13.4

Table 1. **High-level overview of key ablations.** Fourier Encoding of the conditionals outperforms models without them when tested on held out samples. Models with all anthropometric measurements outperform a subset of select anthropometric measurements. γ is Laplacian loss factor as shown in equation 2.

Body Section	P2P Chamfer distance [mm]
Full Body	4.63 ± 0.07
Chest	6.67 ± 0.42
Waist	5.06 ± 0.27
Hip	6.03 ± 0.34
Thighs	5.15 ± 0.16
Calves	4.67 ± 0.08
Arms	7.75 ± 0.07

Table 2. **Registration error of fitting AnthroNet on 1185 SMPL-X Meshes obtained from Fashion Model Dataset [11].**

Method	Model	Height	Chest	Waist	Hips	P2P
SMPLR [34]	SMPL	182	267	309	305	69
STRAPS [50]	SMPL	135	167	145	102	47
SPIN [29]	SMPL	59	92	78	101	29
TUCH [44]	SMPL	58	89	75	57	26
Sengupta et al. [51]	SMPL	82	133	107	63	32
ExPose [12]	SMPL-X	85	99	92	94	35
SHAPY [11]	SMPL-X	51	65	69	57	21
AnthroNet (ours)	SMPL-X	51 ± 47	65 ± 16	69 ± 22	57 ± 16	21 ± 4.5

Table 3. **Performance comparison of AnthroNet when coupled with a pretrained image regressor.** Values are the error in millimeters and evaluated over the HBW dataset [11].

dense and sparse measurements and calculate the root mean squared error between the anthropometric measurements and mean vertex error in millimeters. The results of this experiment are reported in Tab. 3. It is worth highlighting that the deployment of a pretrained image regressor would most likely hinder the whole AnthroNet pipeline to achieve its peak performance due to the potential mismatch between the distributions of data across the two training domains. However, thanks to the registration pipeline, AnthroNet manages to convert SMPL-X models to AnthroNet native mesh system with negligible loss, such that comparable performance on predicting the set of sparse anthropometric measures over HBW dataset is obtained.

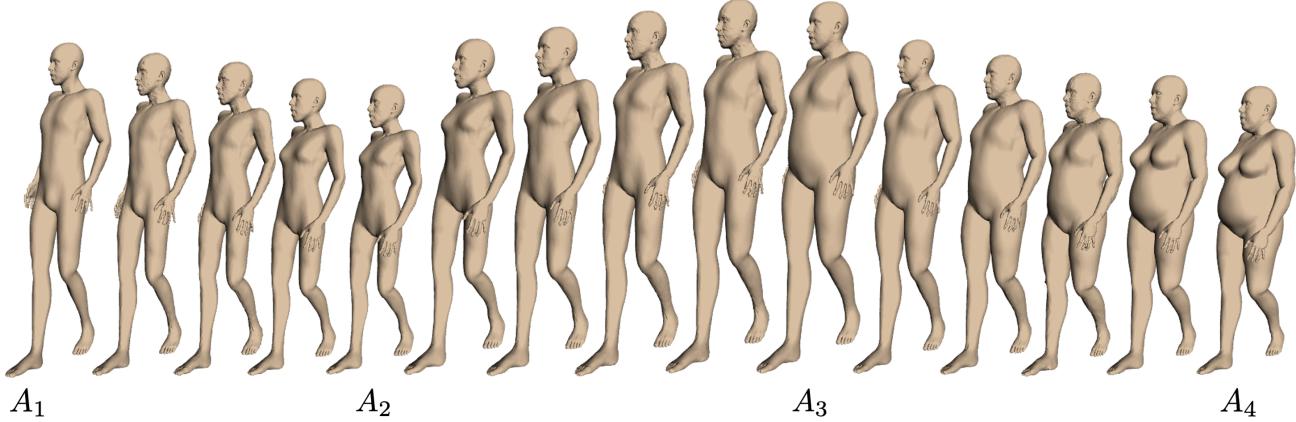


Figure 5. **AnthroNet with Poses.** The figure shows posed meshes obtained by applying LBS to generated meshes on conditionals computed by piece-wise linear interpolation between four sets of full anthropometric measurements (A_1, A_2, A_3, A_4); for details see the appendix.

7. Discussion and Remarks

Limitations While our synthetic data is created with rigor and supported by years of experience of technical character artists, it is not free from unintentional errors and biases introduced during the character sculpting, character animation, and retargeting in our animation pipeline, and motion capture animation data that may not represent all possible human motions. We know that the set of identities and ethnicities our synthetic data generator produces does not represent the entire human population. However, we aimed to do due diligence in ensuring that we are not introducing significant biases by over-representing or under-representing specific body shapes, sexes, ethnicities, or ages; hence, we opted for random uniform sampling across the entire set of available parameters for representing a human body.

We measured our anthropometric descriptors over regions of the body that professional tailors use for making garments. Numerous studies on anthropometric measurements [6, 15, 16, 19, 54, 57, 60] suggest that these set of anthropometric measurements best describe variations seen on the human body shape if additional anthropometric measurements are needed, our synthetic data pipeline provisions for such modifications. Other human body models are generally trained on 3D scans of individuals who self-identified as male or female or exhibited physical characteristics of one or the other sex. One approach to avoiding limitations arising from over or under-representation of either sex is to train a joint body model on data from both sexes and create a “neutral” body model. While this approach has its benefits, we could not do so, as we made the skinning weight data for males and females separately. We shall leave such improvements as future work.

Our generative model aims to learn low-dimensional representations that encapsulate the statistics of the given input

data. Its ability is limited by the input data provided and, secondly, by the optimization and metrics used in the learning process. Inevitably, better architectural or model parameter choices and loss function designs will exist that may or may not yield more accurate human shape manifolds.

Societal Impacts and Ethical Considerations Capturing, processing, storing, and transferring any real human data is subject to strict data and privacy regulations and guidelines across the world [10, 55]. On the other hand, synthetic data allows for preserving privacy and safety and avoiding ethical concerns and biases that may exist with real human data. We sought to provide an alternative to using real human data that addresses some of these concerns. We describe in detail how we have created our synthetic data and body model, and with the provided data, other researchers can reproduce and validate the results.

We make our synthetic data generator publicly available for non-commercial use and academic research. The authors created and vetted the synthetic and motion capture dataset, free from offensive and explicit content and personally identifiable information. We do not anticipate any apparent societal or ethical harm from our work. We believe that relieving the computer vision community of the burden of collecting and annotating vast amounts of human data – that may bear serious privacy, safety, and ethical concerns – delivers a net positive impact. We aimed to provide and pave the path for body models that can better describe attributes relating to the human body shape. We did so by giving dense measurements over the meshes. We hope that this will enable a more inclusive and more expressive model. We also allow the researchers to increase upon the diversity of the synthetic humans by open-sourcing our synthetic data generator.

8. Conclusion

In this work, we presented AnthroNet, a deep generative human body model that takes an extensive set of anthropometric measurements as input and generates body meshes that are more than three times higher resolution than SMPL-X meshes, with high fidelity to anthropometric descriptions of a human body. The whole pipeline is trained end-to-end solely on 100,000 synthetically generated human body meshes, which not only provide a wide variety of shapes and poses but also alleviate the privacy concerns as well as data collection costs with 3D scanning of real humans. Lastly, we provide AnthroNet with two registration modules that allow for the bi-directional transformation of SMPL-X meshes to the AnthroNet, which enables obtaining dense anthropometric measurements on all 3D human body scans that can be registered onto the SMPL-X model.

References

- [1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*. CVPR Spotlight Paper. [3](#)
- [2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2293–2303, 2019. [3](#)
- [3] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5461–5470, 2021. [2](#)
- [4] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. *ACM transactions on graphics (TOG)*, 22(3):587–594, 2003. [2](#)
- [5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. [2](#)
- [6] Murtaza Aslam, Fozia Rajbdad, Shahid Khattak, and Shoaib Azmat. Automatic measurement of anthropometric dimensions using frontal and lateral silhouettes. *IET Computer Vision*, 11(6):434–447, 2017. [4, 8](#)
- [7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016, Lecture Notes in Computer Science*. Springer International Publishing, Oct. 2016. [3](#)
- [8] Steve Borkman, Adam Crespi, Saurav Dhakad, Sujoy Ganguly, Jonathan Hogins, You-Cyuan Jhang, Mohsen Kamalzadeh, Bowen Li, Steven Leal, Pete Parisi, et al. Unity Perception: Generate synthetic data for computer vision. *arXiv preprint arXiv:2107.04259*, 2021. [4](#)
- [9] Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1450–1459, 2021. [2](#)
- [10] Preston Bukaty. *The California Consumer Privacy Act (CCPA): An implementation guide*. IT Governance Publishing, 2019. [8](#)
- [11] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3d body shape regression using metric and semantic attributes. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. [1, 3, 6, 7, 12, 16](#)
- [12] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40. Springer, 2020. [7, 16](#)
- [13] Salehe Erfanian Ebadi, Saurav Dhakad, Sanjay Vishwakarma, Chunpu Wang, You-Cyuan Jhang, Maciek Chociej, Adam Crespi, Alex Thaman, and Sujoy Ganguly. Psp-hdri+: A synthetic dataset generator for pre-training of human-centric computer vision models. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. [3](#)
- [14] Salehe Erfanian Ebadi, You-Cyuan Jhang, Alex Zook, Saurav Dhakad, Adam Crespi, Pete Parisi, Steven Borkman, Jonathan Hogins, and Sujoy Ganguly. Peoplesanspeople: A synthetic data generator for human-centric computer vision. 2021. [3](#)
- [15] Centers for Disease Control and Prevention (CDC). Body measurements (anthropometry). *National Health And Nutrition Examination Survey III*, 1988. [4, 8](#)
- [16] Claire C Gordon, Cynthia L Blackwell, Bruce Bradtmiller, Joseph L Parham, Patricia Barrientos, Stephen P Paquette, Brian D Corner, Jeremy M Carson, Joseph C Venezia, Belva M Rockwell, et al. 2012 anthropometric survey of us army personnel: Methods and summary statistics. Technical report, Army Natick Soldier Research Development and Engineering Center MA, 2014. [4, 8, 12, 13](#)
- [17] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer graphics forum*, volume 28, pages 337–346. Wiley Online Library, 2009. [2](#)
- [18] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994. [12](#)
- [19] Luděk Hynčík, Hana Čechová, Tomasz Bońkowski, Gabriela Kavalířová, Petra Špotová, Viola Hampejsová, and Hong Meng. Personalization of a human body model using subject-specific dimensions for designing clothing patterns. *Applied Sciences*, 11(21):10138, 2021. [4, 8](#)
- [20] Haiyong Jiang, Jianfei Cai, and Jianmin Zheng. Skeleton-aware 3d human shape reconstruction from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5431–5441, 2019. [3](#)
- [21] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and

- bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 2, 3
- [22] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 3
- [23] Isinsu Katircioglu, Bugra Tekin, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Learning latent representations of 3d human pose with deep neural networks. *International Journal of Computer Vision*, 126(12):1326–1341, 2018. 2
- [24] Sohn Kihyuk, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 28(6):3483–3491, 2015. 5
- [25] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 1, 3
- [26] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 1, 3
- [27] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11035–11045, 2021. 1, 3
- [28] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 3
- [29] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 7, 16
- [30] Chun-Liang Li, Tomas Simon, Jason Saragih, Barnabás Póczos, and Yaser Sheikh. Lbs autoencoder: Self-supervised fitting of articulated meshes to point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11967–11976, 2019. 3
- [31] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 1, 2
- [32] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (ToG)*, 33(6):1–13, 2014. 2
- [33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 2, 5
- [34] Meysam Madadi, Hugo Bertiche, and Sergio Escalera. Smplr: Deep learning based smpl reverse for 3d human pose and shape recovery. *Pattern Recognition*, 106:107472, 2020. 7, 16
- [35] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 3
- [36] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9990–9999, 2021. 3
- [37] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Training a feedback loop for hand pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3316–3324, 2015. 2
- [38] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pages 598–613, 2020. 1, 2
- [39] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12695–12705, 2021. 2
- [40] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. Spams: Structured implicit parametric models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12851–12860, 2022. 2
- [41] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2
- [42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2
- [43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 2, 3
- [44] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 7, 16
- [45] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 3
- [46] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d

- morphable models: A large scale face-and-head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10934–10943, 2019. 2
- [47] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1810–1819, 2021. 3
- [48] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 1, 2
- [49] Igor Santesteban, Elena Garces, Miguel A Otaduy, and Dan Casas. Softsmpl: Data-driven modeling of nonlinear soft-tissue dynamics for parametric humans. In *Computer Graphics Forum*, volume 39, pages 65–75. Wiley Online Library, 2020. 3
- [50] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. *arXiv preprint arXiv:2009.10013*, 2020. 7, 16
- [51] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11219–11229, 2021. 7, 16
- [52] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020. 4
- [53] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (ToG)*, 35(4):1–12, 2016. 2
- [54] Aggeliki Tsoli, Matthew Loper, and Michael J Black. Model-based anthropometry: Predicting measurements from 3d human scans in multiple poses. In *IEEE winter conference on applications of computer vision*, pages 83–90. IEEE, 2014. 4, 8
- [55] Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 1st edition, 2017. 8
- [56] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 3
- [57] Stefanie Wuhrer and Chang Shu. Estimating 3d human shapes from measurements. *Machine vision and applications*, 24(6):1133–1147, 2013. 4, 8
- [58] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2
- [59] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 1, 2, 3, 5
- [60] Song Yan, Johan Wirta, and Joni-Kristian Kämäräinen. Anthropometric clothing measurements from 3d body scans. *Machine Vision and Applications*, 31(1):1–11, 2020. 4, 8
- [61] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 3
- [62] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11446–11456, 2021. 1, 3
- [63] Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G Narasimhan, and Minh Vo. Texmesh: Reconstructing detailed human texture and geometry from rgb-d video. In *European Conference on Computer Vision*, pages 492–509. Springer, 2020. 3

Appendix

A. Design Parameters

This section provides a description of all the deep neural networks and the details of the anthropometric measurements introduced in this work.

A.1. Mesh Generator architecture

In this section, we describe the design details of the deep neural network responsible for the mesh generation. In this conditional Variational Auto-Encoder (cVAE), we can identify four main components:

- i. a mesh encoder takes the vertices target mesh in bind pose \mathcal{X}_b^A , and produces a feature vector h_v ;
- ii. a condition encoder takes the anthropometric measurements A (after Fourier encoding) and produces a feature vector h_a :
- iii. a fusion encoder combines h_v and h_a to get the latent representations μ and σ ;
- iv. a decoder fuses the reparameterized latent space z and the condition feature vector h_a and reconstructs the mesh vertices $\tilde{\mathcal{X}}_b^A$.

All the encoders share the same core block, composed of a fully connected (FC) layer followed by a batch normalization (BN) layer and a parametric rectified linear unit (PReLU) activation function:

- The mesh encoder squeezes the input $3V$ coordinates to 128 features.
- The condition encoder takes 37 (36 anthropometric measurements and 1 sex identifier) $\times 2f$ (f being the number of Fourier components) inputs and produces 128 features.
- The fusion module takes $[h_v, h_a]^T$ (256 features) and passes them in the concatenation of two core blocks. The output of this encoding scheme is a 64 features vector that is finally processed by two separate fully connected layers to produce μ and σ whose output dimension is 10.
- The decoder is the concatenation of three core blocks; the first one takes $[z, h_a]^T$ and produces 64 features; the second and the third scales the features up to 128 and 256, respectively. Finally, we include an additional FC layer to produce $3V$ coordinates.

A.2. Anthropometric Measurements

As mentioned in the core paper, we use 36 measures for the anthropometry of human body. Those measurements are listed in Tab. 4 along with the mean, the standard deviation, the min and max of each measure from our $100k$ dataset. To lay a baseline, we have reported the statistics of such anthropometric measurements obtained from 7435 male and 3922 female subjects, as reported in [16]. Empty cells denote the absence of a true correspondence between the anthropometric measurement in our dataset and the real measurements. It is observed that the standard deviation of 34 out of 36 anthropometric measurements of our $100k$ dataset is higher than the one of real measurements which underlines on the diversity of our dataset.

A.3. Details of the Anthropometric regressor

In order to facilitate the evaluation of our system on the Humans in the Wild dataset [11], we trained a regression model that extracts the anthropometric measurements from registered meshes.

For each of the 36 anthropometric measurements defined in sec. 5.1 in core paper, we trained an expert Multi Layer Perceptron [18] comprised of three layers. To ensure the disentanglement between the anthropometric measures, we designed a mask for each measure to isolate the vertices on the mesh related to the measure as shown in Fig. 6. This design prevents interference from other vertices in the prediction of a measure on a specific body part. During training, for each measurement the input mesh is masked with a specific mask, as depicted in Fig. 6. At inference time, we aggregate the outputs of all the 36 models.

We conducted different experiments corresponding to specialized models for each anthropometric measurement using no mask or a binary mask or a soft exponential mask (described in the Fig. 6) but also one general model which predicts all the anthropometric measurements at once. The results are reported in the Tab. 5. The specialized models using soft exponential masks have proven the best performance, thus they were chosen as reference.

A.4. Deep Skinner architecture

In this section, we describe the design details of the deep neural network responsible for the rigging, skinning and posing of the meshes. The goal of this network is to estimate the pose correctives $\Delta\tilde{\mathcal{X}}_{\bar{\theta}}^A$ relative to the generated mesh in bind pose $\tilde{\mathcal{X}}_b^A$ and the desired pose $\bar{\theta}$. The forward pass of this network is twofold:

- i. an encoder concatenates the bind pose vertices \mathcal{X}_b^A and the desired pose $\bar{\theta}$ (after Fourier encoding), and produces a feature vector h ;
- ii. a decoder reconstructs the offsets $\Delta\tilde{\mathcal{X}}_{\bar{\theta}}^A$ from h .

Name	100k Synthetic				Real [16]							
					Male				Female			
	mean	std	min	max	mean	std	min	max	mean	std	min	max
Waist circumference	0.828	0.174	0.55	1.250	0.941	0.112	0.648	1.379	0.861	0.100	0.611	1.334
Chest circumference	1.061	0.159	0.76	1.522	1.059	0.087	0.774	1.469	0.947	0.083	0.695	1.266
Hip circumference	0.988	0.154	0.72	1.368	0.102	0.077	0.737	1.305	1.021	0.076	0.798	1.341
Height	1.760	0.144	1.49	1.991	1.756	0.0686	1.491	1.993	1.629	0.0642	1.410	1.829
Shoulder width	0.295	0.012	0.248	0.335	0.416	0.0192	0.337	0.489	0.365	0.0183	0.283	0.422
Torso height from back	0.376	0.030	0.310	0.425	0.382	0.0259	0.293	0.483	0.350	0.0223	0.284	0.435
Torso height from front	0.324	0.029	0.256	0.394	0.478	0.027	0.383	0.598	0.425	0.026	0.345	0.532
Head circumference	0.546	0.009	0.521	0.554	0.574	0.016	0.516	0.633	0.561	0.019	0.500	0.635
Neck circumference	0.431	0.044	0.320	0.541	0.398	0.026	0.311	0.514	0.330	0.019	0.275	0.424
Head height	0.245	0.004	0.232	0.251	-	-	-	-	-	-	-	-
Mid-line neck length	0.092	0.013	0.056	0.115	-	-	-	-	-	-	-	-
Lateral neck length	0.120	0.012	0.078	0.141	0.108	0.012	0.062	0.151	0.106	0.012	0.071	0.150
Hand size	0.177	0.011	0.149	0.197	0.193	0.010	0.164	0.239	0.181	0.01	0.145	0.220
Arm circumference	0.283	0.052	0.165	0.409	0.358	0.035	0.246	0.490	0.306	0.031	0.216	0.435
Arm length	0.329	0.029	0.270	0.395	0.364	0.018	0.298	0.423	0.334	0.017	0.271	0.398
Forearm circumference	0.270	0.038	0.184	0.367	0.310	0.022	0.233	0.402	0.264	0.019	0.200	0.342
Forearm length	0.254	0.027	0.197	0.300	0.268	0.015	0.216	0.328	0.241	0.015	0.169	0.297
Thigh circumference	0.551	0.073	0.395	0.733	0.625	0.059	0.412	0.843	0.616	0.056	0.448	0.870
Thigh length	0.512	0.046	0.415	0.590	0.409	0.028	0.298	0.520	0.379	0.024	0.304	0.474
Calf circumference	0.354	0.042	0.248	0.458	0.392	0.030	0.266	0.523	0.373	0.029	0.282	0.482
Calf length	0.424	0.040	0.348	0.504	0.419	0.025	0.344	0.519	0.403	0.026	0.294	0.506
Foot width	0.098	0.007	0.083	0.115	0.102	0.005	0.079	0.126	0.093	0.005	0.077	0.108
Heel to ball length	0.198	0.014	0.169	0.225	0.201	0.011	0.156	0.245	0.182	0.009	0.151	0.216
Heel to toe length	0.269	0.019	0.229	0.306	0.271	0.013	0.216	0.323	0.246	0.012	0.198	0.286

Table 4. Statistics of the anthropometric measures (in mm) on the 100k dataset compared with real measurements collected from [16]. The minimum, maximum values along with the mean μ and the standard deviation σ are reported.

	MSE error [mm]
no mask	0.086
binary mask	0.0011
soft exponential	0.000142
one model	0.0010

Table 5. Mean Euclidean distance on the test dataset for the Anthro regressor model for different masks.

The encoder shares the same core block of the generative network, composed of a fully connected (FC) layer followed by a batch normalization (BN) layer and a parametric rectified linear unit (PReLU) activation function. The decoding stage is made of a FC layer with 128 input features and 128 output features, and a FC layer with 128 input features and $3V$ output features. The latter passes through a PReLU activation function to produce the offsets estimate.

We performed ablation studies to determine the best design, in terms of mesh reconstruction ability. Tab. 6 reports different combinations of architectures and number of Fourier encoding components, along with the mean vertex-to-vertex distance on the test dataset. From our experiments, we can state that a simpler network achieves better

encoding blocks	decoding FC layers	BN	Frequencies	P2P
1	2	yes	32	4
1	2	yes	16	4
3	2	yes	16	15
1	1	no	16	29
1	2	yes	0	56

Table 6. Mean vertex-to-vertex distance (in mm) on held out test dataset for the Skinner model with different architecture designs, and number of Fourier components.

results than deeper networks. Moreover, the batch normalization layers are required to drive the optimization towards the desired solutions.

B. Extended Experiments and Results

B.1. Random Fourier Encoding

Tab. 7 presents models with varying levels of Fourier Frequencies used for encoding of conditionals. By visual inspecting the training and test samples, we concluded that Fourier-encoding helped in preventing overfitting both the generator and skinner networks.

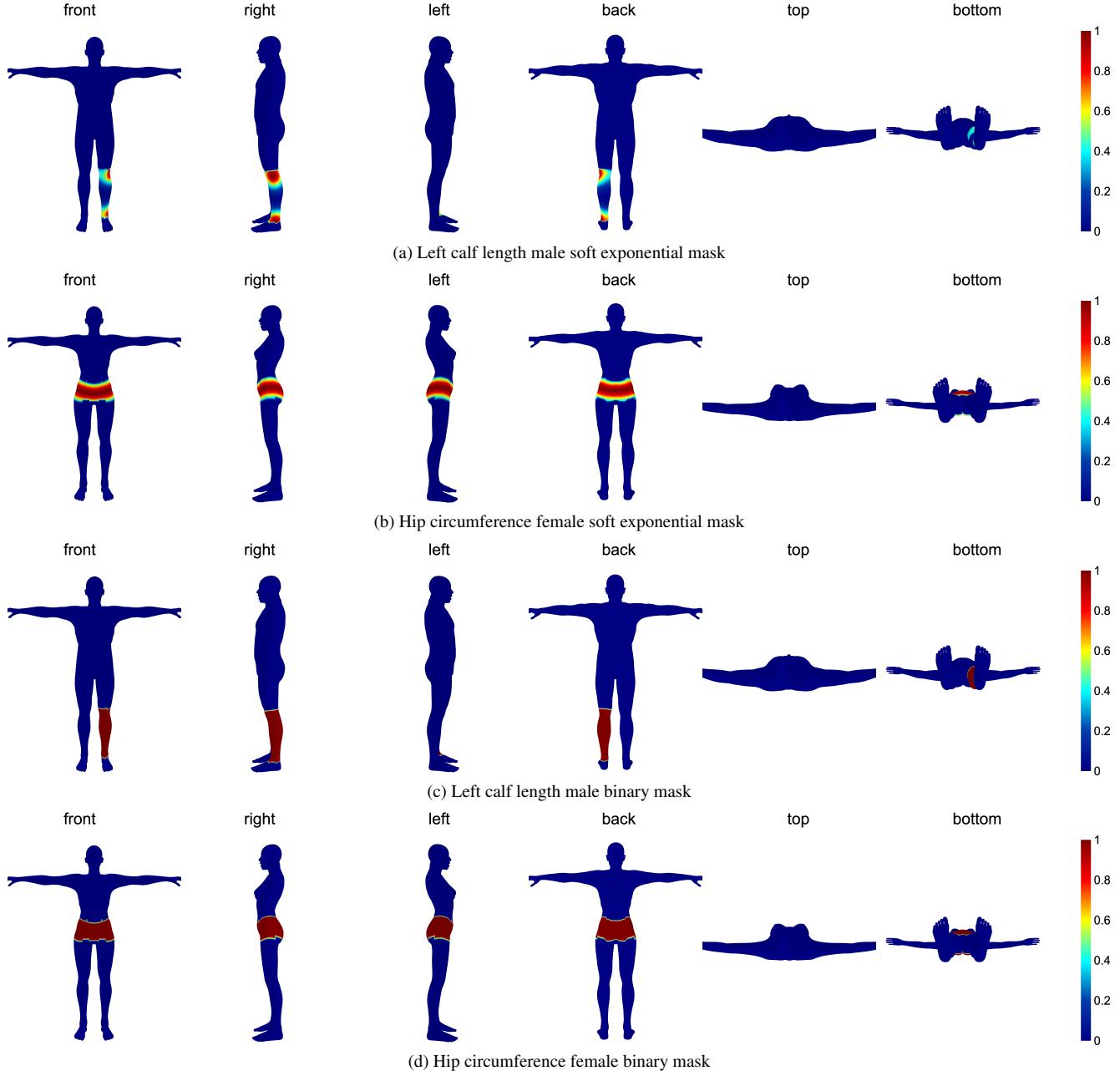


Figure 6. After isolating the mesh related volume, we calculated the minimal Euclidean distance d_{\min} between each vertices of the volume and the ones used for the anthropometric measure. For the soft exponential mask (Figs. 6a and 6b), the final weight in the mask associated to the vertex point is $e^{-d_{\min}}$. For the binary mask (Figs. 6c and 6d), we binarized the distance with a threshold equal to 0.5.

B.2. Generative Performance

This section details the generation of qualitative results presented Fig. 4 (a, b, c) in the core paper. For each sub figure, two human measurement vectors A_1, A_2 are samples such that they posses desired physical attributes. In the case of Fig. 4a, A_1 corresponds to dense measurements of a female with lean body, and A_2 corresponds to female with

the similar height but large circumferential measurements. Tab. 8 presents the (A_1, A_2) pairs for the figures a, b, c respectively. Then, three intermediate measurements between A_1, A_2 are obtained by linear interpolation. It can be seen that model is able to generate smooth interpolation in mesh space that respects the prescribed measurements.

Frequencies	P2P	chest c.	hip c.	arm c.	calf c.	thigh c.	waist c.
8	3.83	46.9	9.79	6.22	4.53	9.93	12.3
16	5.04	60.4	14.9	14.5	9.60	7.08	21.6
32	4.87	57.7	13.1	12.4	8.40	7.93	18.0

Table 7. **Effect of Fourier Frequencies used in encoding** P2P loss is the mean vertex-to-vertex distance in mm on held out test dataset. Similarly chest, hip, arm, calf, thigh, waist represents error in measurements of the circumferences in mm.

Figure	Fig. 4 a		Fig. 4 b		Fig. 4 c	
Notation	A1	A2	A1	A2	A1	A2
sex	F	F	M	M	M	F
waist circumference	0.61	1.13	0.91	0.75	0.73	0.73
chest circumference	0.90	1.36	1.11	0.94	0.93	1.07
hip circumference	0.85	1.22	1.01	0.85	0.84	1.00
height	1.60	1.62	1.99	1.50	1.97	1.98
shoulder width	0.27	0.28	0.31	0.26	0.30	0.32
torso height from back	0.34	0.35	0.41	0.31	0.42	0.42
torso height from front	0.29	0.32	0.35	0.26	0.34	0.36
head circumference	0.53	0.54	0.55	0.53	0.55	0.55
neck circumference	0.37	0.47	0.48	0.40	0.42	0.45
head height	0.24	0.24	0.25	0.24	0.25	0.25
midline neck length	0.09	0.07	0.10	0.08	0.11	0.11
lateral neck length	0.11	0.10	0.13	0.10	0.14	0.14
hand size	0.16	0.17	0.19	0.16	0.19	0.19
arm circumference	0.23	0.34	0.33	0.27	0.26	0.26
arm length	0.29	0.30	0.39	0.29	0.38	0.36
forearm circumference	0.22	0.33	0.29	0.24	0.25	0.26
forearm length	0.23	0.23	0.30	0.20	0.30	0.29
thigh circumference	0.48	0.63	0.59	0.50	0.50	0.57
thigh length	0.48	0.47	0.58	0.42	0.57	0.59
calf circumference	0.31	0.38	0.38	0.33	0.33	0.37
calf length	0.38	0.38	0.50	0.37	0.49	0.49
foot width	0.09	0.09	0.11	0.08	0.10	0.11
heel to ball length	0.18	0.18	0.22	0.17	0.21	0.22
heel to toe length	0.25	0.25	0.29	0.23	0.29	0.30

Table 8. **Measurement data (in mm) for generation of Figure 4.** The measurement pairs (A_1, A_2) for the figures 4.a, 4.b, 4.c are presented respectively. If there exists a significant difference between certain measurement, it is highlighted in bold. For example, in Fig. 4c, A_1, A_2 pair corresponds to male and female with similar physical characteristics that only differs in terms of gender and hip circumference.

B.3. Regression from SMPLX parameters

In this section, we describe a regression module, referred to as Beta-to-Anthropometrics (B2A), which allows us to regress the set of dense anthropometric measures from SMPL-X shape parameters (β). In other words, this module allows us to use any pretrained image regression module that provides shape parameters, either in SMPL or SMPL-X formats, to generate AnthroNet meshes. Due to the numerous capabilities of AnthroNet meshes in downstream tasks, e.g., dense anthropometry, rigging, skinning, and posing, the B2A module bridges the gap between any human body model in SMPL or SMPL-X formats and the AnthroNet sys-

Frequencies	Training loss	Validation loss	Test loss
0	0.015	0.014	0.011
8	0.016	0.020	0.016
32	0.014	0.022	0.020

Table 9. **Effect of Fourier encoding in training of B2A module** Please note that loss is the mean squared error between the predictions and the target anthropometric measures. It is worth highlighting that the training, validation, and test sets in this table are different splits of the Model Agency dataset.

tem. As shown in Fig. 7, B2A regresses the set of anthropometric measures required by our generative model from SMPL-X shape parameters.

To train B2A, we employed the image-regression module from Shapely framework to obtain SMPL-X shape parameters of the subjects in the Model Agency dataset. Then, AnthroNet equivalent of these SMPL-X meshes are obtained through the SMPLX-to-AnthroNet registration pipeline, which allows us to compute the set of dense anthropometrics for each subject in the Model Agency dataset. This process provides us the training data for the B2A module ($B2A() : \beta \rightarrow A$). B2A is a three-layer MLP, which uses mean squared error as the loss function and Adam optimizer as the optimization tool. To train the B2A module, we investigated the effect of Fourier encoding of anthropometric measures on the accuracy of regression. As the results in Tab. 9 show, the best results are obtained when no Fourier encoding is used.

To evaluate the efficacy of B2A module in translating SMPL(-X) mesh system to AnthroNet by generating meshes with high fidelity to their SMPL(-X) counterparts, we deploy the Human Bodies in the Wild (HBW) dataset. As shown in Fig. 7, the process here is to go from SMPL-X shape parameters (obtained from Shapely image regressor) to dense anthropometric measurements and then to AnthroNet mesh. Finally, the AnthroNet mesh is registered to SMPL format to calculate the P2P error with the ground truth mesh. This pipeline not only evaluates the efficacy of the B2A module but also assesses the fidelity of both AnthroNet generator and the AnthroNet-to-SMPL registration pipeline. The results of this evaluation are reported in Tab. 10. It is worth highlighting that due to the unavailability of ground truth meshes for the test set of HBW, we have evaluated the AnthroNet on the validation set of HBW. Please note that HBW dataset has never been used in any development phases of AnthroNet, and thus, the validation split of HBW is indeed as unseen data for AnthroNet.

Due to the fact that we are using the image-regression module from Shapely framework to regress shape parameters from images, it is expected that our performance is lower-bounded by the performance of the Shapely framework. However, given the fact that our pipeline encom-

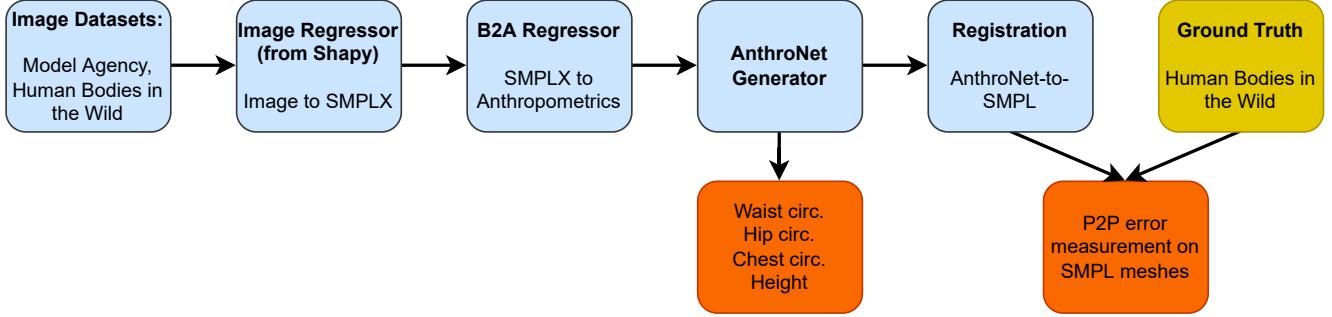


Figure 7. The proposed pipeline to generate AnthroNet meshes by regressing the set of dense anthropometric measures from SMPLX shape parameters (β). For evaluation, the AnthroNet mesh is registered to SMPL mesh, and then it is compared with the ground truth mesh.

Method	Model	Height	Chest	Waist	Hips	P2P
SMPLR [34]	SMPL	182	267	309	305	69
STRAPS [50]	SMPL	135	167	145	102	47
SPIN [29]	SMPL	59	92	78	101	29
TUCH [44]	SMPL	58	89	75	57	26
Sengupta et al. [51]	SMPL	82	133	107	63	32
ExPose [12]	SMPL-X	85	99	92	94	35
SHAPY [11]	SMPL-X	51	65	69	57	21
AnthroNet (ours)	SMPL-X	86	79	90	86	33

Table 10. **Performance comparison of AnthroNet when coupled with a pretrained image regressor.** Values are the error in millimeters and evaluated over the HBW dataset [11]. Please note that due to the absence of ground truth, AnthroNet is evaluated on the validation split of the HBW dataset while its counterparts are evaluated on the test split.

passes several modules that are trained on different datasets and operate based non-deterministic functions such as optimization procedures, the reported performance for AnthroNet in Tab. 10 suggests a high degree of fidelity to the predictions of the image-regressor and a high degree of robustness/generalization in its entire pipeline. In other words, these results suggest that AnthroNet can be coupled with any SMPL(-X)-based image-regression module while imposing minimal loss on the reconstructed meshes (compared to ground truth) and providing several downstream benefits through the AnthroNet system.