Accurate 3D Body Shape Regression using Metric and Semantic Attributes

Vasileios Choutas*[1], Lea Müller*[1], Chun-Hao P. Huang[1], Siyu Tang[2], Dimitrios Tzionas[1], Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]ETH Zürich
{vchoutas, lea.mueller, paul.huang, stang, dtzionas, black}@tuebingen.mpg.de
* Equal contribution, alphabetical order

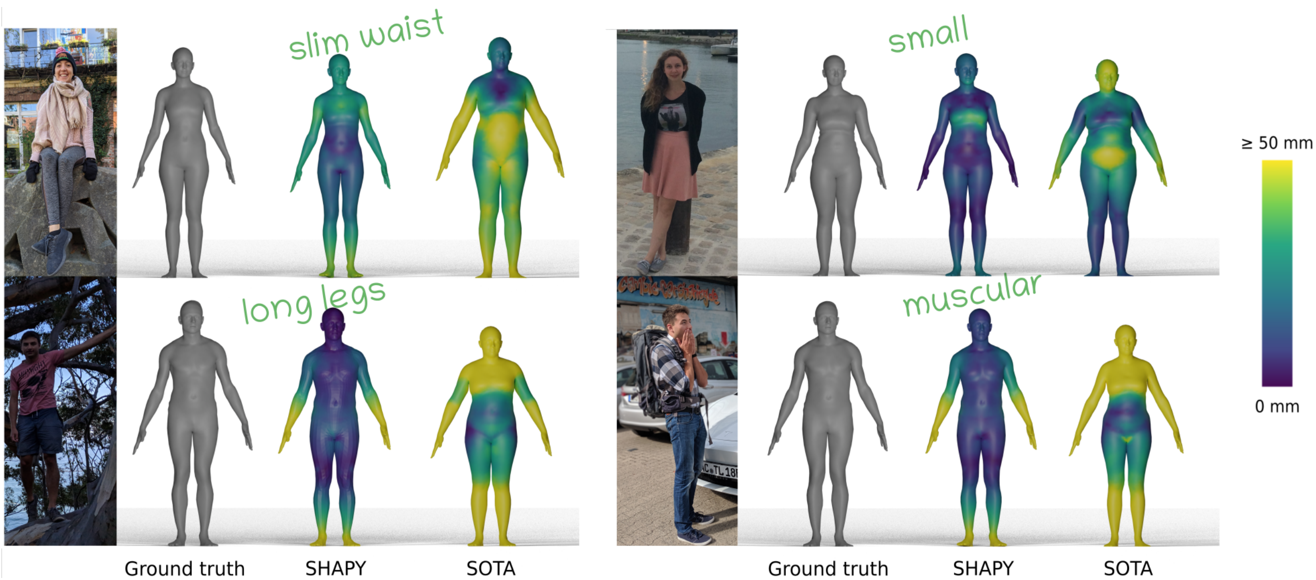

Figure 1. Existing work on 3D human reconstruction from a color image focuses mainly on *pose*. We present SHAPY, a model that focuses on body *shape* and learns to predict dense 3D shape from a color image, using crowd-sourced *linguistic shape attributes*. Even with this weak supervision, SHAPY outperforms the state of the art (SOTA) [52] on in-the-wild images with varied clothing.

Abstract

While methods that regress 3D human meshes from images have progressed rapidly, the estimated body shapes often do not capture the true human shape. This is problematic since, for many applications, accurate body shape is as important as pose. The key reason that body shape accuracy lags pose accuracy is the lack of data. While humans can label 2D joints, and these constrain 3D pose, it is not so easy to "label" 3D body shape. Since paired data with images and 3D body shape are rare, we exploit two sources of information: (1) we collect internet images of diverse "fashion" models together with a small set of anthropometric measurements; (2) we collect linguistic shape attributes for a wide range of 3D body meshes and the model images. Taken together, these datasets provide sufficient constraints to infer dense 3D shape. We exploit the anthropometric measurements and linguistic shape attributes in several novel ways to train a neural network, called SHAPY, that regresses 3D human pose and shape from an RGB image. We evaluate SHAPY on public benchmarks, but note that they either lack significant body shape variation, ground-truth shape, or clothing variation. Thus, we collect a new dataset for evaluating 3D human shape estimation, called HBW, containing photos of "Human Bodies in the Wild" for which we have ground-truth 3D body scans. On this new benchmark, SHAPY significantly outperforms state-of-the-art methods on the task of 3D body shape estimation. This is the first demonstration that 3D body shape regression from images can be trained from easy-to-obtain anthropometric measurements and linguistic shape attributes. Our model and data are available at: shapy.is.tue.mpg.de

1. Introduction

The field of 3D human pose and shape (HPS) estimation is progressing rapidly and methods now regress accurate 3D pose from a single image [7, 26, 28, 30–33, 43, 65, 67]. Unfortunately, less attention has been paid to body shape and many methods produce body shapes that clearly do not represent the person in the image (Fig. 1, top right). There are several reasons behind this. Current evaluation datasets focus on pose and not shape. Training datasets of images with 3D ground-truth shape are lacking. Additionally, humans appear in images wearing clothing that obscures the body, making the problem challenging. Finally, the fundamental scale ambiguity in 2D images, makes 3D shape difficult to estimate. For many applications, however, realistic body shape is critical. These include AR/VR, apparel design, virtual try-on, and fitness. To democratize avatars, it is important to represent and estimate all possible 3D body shapes; we make a step in that direction.

Note that commercial solutions to this problem require users to wear tight fitting clothing and capture multiple images or a video sequence using constrained poses. In contrast, we tackle the unconstrained problem of 3D body shape estimation in the wild from a single RGB image of a person in an arbitrary pose and standard clothing.

Most current approaches to HPS estimation learn to regress a parametric 3D body model like SMPL [37] from images using 2D joint locations as training data. Such joint locations are easy for human annotators to label in images. Supervising the training with joints, however, is not sufficient to learn shape since an infinite number of body shapes can share the same joints. For example, consider someone who puts on weight. Their body shape changes but their joints stay the same. Several recent methods employ additional 2D cues, such as the silhouette, to provide additional shape cues [51,52]. Silhouettes, however, are influenced by clothing and do not provide explicit 3D supervision. Synthetic approaches [35], on the other hand, drape SMPL 3D bodies in virtual clothing and render them in images. While this provides ground-truth 3D shape, realistic synthesis of clothed humans is challenging, resulting in a domain gap.

To address these issues, we present SHAPY, a new deep neural network that accurately regresses 3D body shape and pose from a single RGB image. To train SHAPY, we first need to address the lack of paired training data with real images and ground-truth shape. Without access to such data, we need alternatives that are easier to acquire, analogous to 2D joints used in pose estimation. To do so, we introduce two novel datasets and corresponding training methods.

First, in lieu of full 3D body scans, we use images of people with diverse body shapes for which we have anthropometric measurements such as height as well as chest, waist, and hip circumference. While many 3D human shapes can share the same measurements, they do constrain the space of possible shapes. Additionally, these are important measurements for applications in clothing and health. Accurate anthropometric measurements like these are difficult for individuals to take themselves but they are often captured for

Figure 2. Model-agency websites contain multiple images of models together with anthropometric measurements. A wide range of body shapes are represented; example from pexels.com.

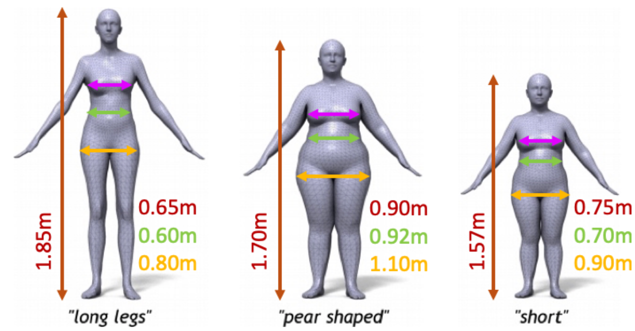

Figure 3. We crowd-source scores for linguistic body-shape attributes [57] and compute anthropometric measurements for CAESAR [47] body meshes. We also crowd-source linguistic shape attribute scores for model images, like those in Fig. 2

different applications. Specifically, modeling agencies provide such information about their models; accuracy is a requirement for modeling clothing. Thus, we collect a diverse set of such model images (with varied ethnicity, clothing, and body shape) with associated measurements; see Fig. 2.

Since sparse anthropometric measurements do not fully constrain body shape, we exploit a novel approach and also use *linguistic shape attributes*. Prior work has shown that people can rate images of others according to shape attributes such as "short/tall", "long legs" or "pear shaped" [57]; see Fig. 3. Using the average scores from several raters, Streuber et al. [57] (BodyTalk) regress metrically accurate 3D body shape. This approach gives us a way to easily label images of people and use these labels to constrain 3D shape. To our knowledge, this sort of linguistic shape attribute data has not previously been exploited to train a neural network to infer 3D body shape from images.

We exploit these new datasets to train SHAPY with three novel *losses*, which can be exploited by any 3D human body reconstruction method: (1) We define functions of the SMPL body mesh that return a sparse set of anthropometric measurements. When measurements are available for an image we use a loss that penalizes mesh measurements that differ from the ground-truth (GT). (2) We learn a "Shape to

Attribute" (S2A) function that maps 3D bodies to linguistic attribute scores. During training, we map meshes to attribute scores and penalize differences from the GT scores. (3) We similarly learn a function that maps "Attributes to Shape" (A2S). We then penalize body shape parameters that deviate from the prediction.

We study each term in detail to arrive at the final method. Evaluation is challenging because existing benchmarks with GT shape either contain too few subjects [61] or have limited clothing complexity and only pseudo-GT shape [51]. We fill this gap with a new dataset, named "Human Bodies in the Wild" (HBW), that contains a ground-truth 3D body scan and several in-the-wild photos of 35 subjects, for a total of 2543 photos. Evaluation on this shows that SHAPY estimates much more accurate 3D shape.

Models, data and code are available at shapy.is.tue.mpg.de.

2. Related Work

3D human pose and shape (HPS): Methods that reconstruct 3D human bodies from one or more RGB images can be split into two broad categories: (1) **parametric methods** that predict parameters of a statistical 3D body model, such as SCAPE [3], SMPL [37], SMPL-X [43], Adam [26], GHUM [65], and (2) **non-parametric methods** that predict a free-form representation of the human body [24, 50, 59, 64]. Parametric approaches lack details w.r.t. non-parametric ones, e.g., clothing or hair. However, parametric models disentangle the effects of identity and pose on the overall shape. Therefore, their parameters provide control for re-shaping and re-posing. Moreover, pose can be factored out to bring meshes in a canonical pose; this is important for evaluating estimates of an individual's shape. Finally, since topology is fixed, meshes can be compared easily. For these reasons, we use a SMPL-X body model.

Parametric methods follow two main paradigms, and are based on optimization or regression. **Optimization-based methods** [5, 7, 16, 43] search for model configurations that best explain image evidence, usually 2D landmarks [8], subject to model priors that usually encourage parameters to be close to the mean of the model space. Numerous methods penalize the discrepancy between the projected and ground-truth silhouettes [22, 34] to estimate shape. However, this needs special care to handle clothing [4]; without this, erroneous solutions emerge that "inflate" body shape to explain the "clothed" silhouette. **Regression-based methods** [9, 14, 25, 27, 30, 33, 35, 40, 66] are currently based on deep neural networks that directly regress model parameters from image pixels. Their training sets are a mixture of data captured in laboratory settings [23, 56], with model parameters estimated from MoCap markers [39], and in-the-wild image collections, such as COCO [36], that contain 2D keypoint annotations. Optimization and regression can be combined, for example via in-the-network model fitting [33,40].

Estimating 3D body shape: State-of-the-art methods are effective for estimating 3D pose, but *struggle* with estimating *body shape* under clothing. There are several reasons for this. First, 2D keypoints alone are not sufficient to fully constrain 3D body shape. Second, shape priors address the lack of constraints, but bias solutions towards "average" shapes [7,33,40,43]. Third, datasets with in-the-wild images have noisy 3D bodies, recovered by fitting a model to 2D keypoints [7, 43]. Fourth, datasets captured in laboratory settings have a small number of subjects, who do not represent the full spectrum of body shapes. Thus, there is a scarcity of images with known, *accurate*, 3D body shape. Existing methods deal with this in two ways.

First, rendering *synthetic images* is attractive since it gives automatic and precise ground-truth annotation. This involves shaping, posing, dressing and texturing a 3D body model [20,51,53,60,62], then lighting it and rendering it in a scene. Doing this realistically and with natural clothing is expensive, hence, current datasets suffer from a domain gap. Alternative methods use artist-curated 3D scans [42,49,50], which are realistic but limited in variety.

Second, *2D shape cues* for in-the-wild images, (body-part segmentation masks [12,41,48], silhouettes [1,22,44]) are attractive, as these can be manually annotated or automatically detected [15, 18]. However, fitting to such cues often gives unrealistic body shapes, by inflating the body to "explain" the clothing "baked" into silhouettes and masks.

Most related to our work is the work of Sengupta et al. [51–53] who estimate body shape using a probabilistic learning approach, trained on edge-filtered synthetic images. They evaluate on the SSP-3D dataset of real images with pseudo-GT 3D bodies, estimated by fitting SMPL to multiple video frames. SSP-3D is biased to people with tight-fitting clothing. Their silhouette-based method works well on SSP-3D but does not generalize to people in normal clothing, tending to over-estimate body shape; see Fig. 1.

In contrast to previous work, SHAPY is trained with in-the-wild images paired with linguistic shape attributes, which are annotations that can be easily crowd-sourced for weak shape supervision. We also go beyond SSP-3D to provide HBW, a new dataset with in-the-wild images, varied clothing, and precise GT from 3D scans.

Shape, measurements and attributes: Body shapes can be generated from anthropometric measurements [2, 54, 55]. Tsoli et al. [58] register a body model to multiple high-resolution body scans to extract body measurements. The "Virtual Caliper" [46] allows users to build metrically accurate avatars of themselves using measurements or VR game controllers. ViBE [21] collects images, measurements (bust, waist, hip circumference, height) and the dress-size of models from clothing websites to train a clothing recommendation network. We draw inspiration from these approaches for data collection and supervision.

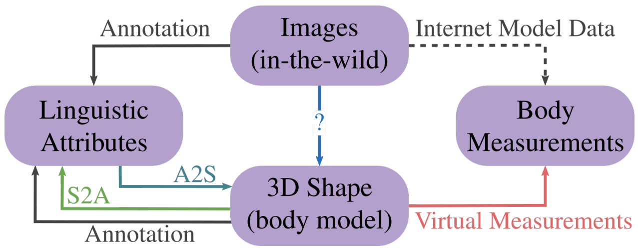

Figure 4. Shape representations and data collection. Our goal is 3D body shape estimation from in-the-wild images. Collecting data for direct supervision is difficult and does not scale. We explore two alternatives. **Linguistic Shape Attributes**: We annotate attributes ("A") for CAESAR meshes, for which we have accurate shape ("S") parameters, and learn the "A2S" and "S2A" models, to map between these representations. Attribute annotations for images can be easily crowd-sourced, making these scalable. **Anthropometric Measurements**: We collect images with sparse body measurements from model-agency websites. A virtual measurement module [46] computes the measurements from 3D meshes. **Training**: We combine these sources to learn a regressor with weak supervision that infers 3D shape from an image.

Streuber et al. [57] learn BodyTalk, a model that generates 3D body shapes from linguistic attributes. For this, they select attributes that describe human shape and ask annotators to rate how much each attribute applies to a body. They fit a linear model that maps attribute ratings to SMPL shape parameters. Inspired by this, we collect attribute ratings for CAESAR meshes [47] and in-the-wild data as proxy shape supervision to train a HPS regressor. Unlike BodyTalk, SHAPY automatically infers shape from images.

Anthropometry from images: Single-View metrology [10] estimates the height of a person in an image, using horizontal and vertical vanishing points and the height of a reference object. Günel et al. [17] introduce the IMDB-23K dataset by gathering publicly available celebrity images and their height information. Zhu et al. [68] use this dataset to learn to predict the height of people in images. Dey et al. [11] estimate the height of users in a photo collection by computing height differences between people in an image, creating a graph that links people across photos, and solving a maximum likelihood estimation problem. Bieler et al. [6] use gravity as a prior to convert pixel measurements extracted from a video to metric height. These methods do not address body shape.

3. Representations & Data for Body Shape

We use linguistic shape attributes and anthropometric measurements as a connecting component between in-the-wild images and ground-truth body shapes; see Fig. 4. To that end, we annotate linguistic shape attributes for 3D meshes and in-the-wild images, the latter from fashion-model agencies, labeled via Amazon Mechanical Turk.

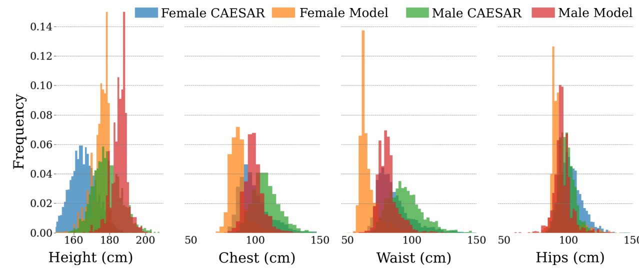

Figure 5. Histogram of height and chest/waist/hips circumference for data from model-agency websites (Sec. 3.2) and CAESAR. Model-agency data is diverse, yet not as much as CAESAR data.

3.1. SMPL-X Body Model

We use SMPL-X [43], a differentiable model that maps shape, $\boldsymbol{\beta}$, pose, $\boldsymbol{\theta}$, and expression, $\boldsymbol{\psi}$, parameters to a 3D mesh, M, with $N = 10,475$ vertices, V. The shape vector $\boldsymbol{\beta} \in \mathbb{R}^B$ ($B \leq 300$) has coefficients of a low-dimensional PCA space. The vertices are posed with linear blend skinning with a learned rigged skeleton, $X \in \mathbb{R}^{55\times3}$.

3.2. Model-Agency Images

Model agencies typically provide multiple color images of each model, in various poses, outfits, hairstyles, scenes, and with a varying camera framing, together with anthropometric measurements and clothing size. We collect training data from multiple model-agency websites, focusing on under-represented body types, namely: curve-models.com, cocainemodels.com, nemesismodels.com, jayjay-models.de, kultmodels.com, modelwerk.de, models1.co.uk. showcast.de, the-models.de, and ullamodels.com. In addition to photos, we store gender and four anthropometric measurements, i.e. height, chest, waist and hip circumference, when available. To avoid having the same subject in both the training and test set, we match model identities across websites to identify models that work for several agencies. For details, see Sup. Mat.

After identity filtering, we have $94,620$ images of $4,419$ models along with their anthropometric measurements. However, the distributions of these measurements, shown in Fig. 5, reveal a bias for "fashion model" body shapes, while other body types are under-represented in comparison to CAESAR [47]. To enhance diversity in body-shapes and avoid strong biases and log tails, we compute the quantized 2D-distribution for height and weight and sample up to 3 models per bin. This results in $N = 1,185$ models (714 females, 471 males) and $20,635$ images.

3.3. Linguistic Shape Attributes

Human body shape can be described by linguistic shape attributes [19]. We draw inspiration from Streuber et al. [57] who collect scores for 30 linguistic attributes for

Male & Female		Male only	Female only
short	long neck	skinny arms	pear shaped
big	long legs	average	petite
tall	long torso	rectangular	slim waist
muscular	short arms	delicate build	large breasts
	broad shoulders	soft body	skinny legs
		masculine	feminine

Table 1. Linguistic shape attributes for human bodies. Some attributes apply to both genders, but others are gender specific.

256 3D body meshes, generated by sampling SMPL's shape space, to train a linear "attribute to shape" regressor. In contrast, we train a model that takes as input an image, instead of attributes, and outputs an accurate 3D shape (and pose).

We crowd-source linguistic attribute scores for a variety of body shapes, using images from the following sources:

Rendered CAESAR images: We use CAESAR [47] bodies to learn mappings between linguistic shape attributes, anthropometric measurements, and SMPL-X shape parameters, $\boldsymbol{\beta}$. Specifically, we register a "gendered" SMPL-X model with 100 shape components to $1,700$ male and $2,102$ female 3D scans, pose all meshes in an A-pose, and render synthetic images with the same virtual camera.

Model-agency photos: Each annotator is shown 3 body images per subject, sampled from the image pool of Sec. 3.2.

Annotation: To keep annotation tractable, we use $A = 15$ linguistic shape attributes per gender (subset of BodyTalk's [57] attributes); see Tab. 1. Each image is annotated by $K = 15$ annotators on Amazon Mechanical Turk. Their task is to *"indicate how strongly [they] agree or disagree that the [listed] words describe the shape of the [depicted] person's body"*; for an example, see Sup. Mat. Annotations range on a discrete 5-level Likert scale from 1 (strongly disagree) to 5 (strongly agree). We get a rating matrix $\mathbf{A} \in \{1, 2, 3, 4, 5\}^{N \times A \times K}$, where N is the number of subjects. In the following, a_{ijk} denotes an element of $\mathbf{A}$.

4. Mapping Shape Representations

In Sec. 3 we introduce three body-shape representations: (1) SMPL-X's PCA shape space (Sec. 3.1), (2) anthropometric measurements (Sec. 3.2), and (3) linguistic shape attribute scores (Sec. 3.3). Here we learn mappings between these, so that in Sec. 5 we can define new losses for training body shape regressors using multiple data sources.

4.1. Virtual Measurements (VM)

We obtain anthropometric measurements from a 3D body mesh in a T-pose, namely height, $H(\boldsymbol{\beta})$, weight, $W(\boldsymbol{\beta})$, and chest, waist and hip circumferences, $C_{\mathrm{c}}(\boldsymbol{\beta})$, $C_{\mathrm{w}}(\boldsymbol{\beta})$, and $C_{\mathrm{h}}(\boldsymbol{\beta})$, respectively, by following Wuhrer et al. [63] and the "Virtual Caliper" [46]. For details on how we compute these measurements, see Sup. Mat.

4.2. Attributes and 3D Shape

Attributes to Shape (A2S): We predict SMPL-X shape coefficients from linguistic attribute scores with a second-degree polynomial regression model. For each shape $\boldsymbol{\beta}_i$, $i = 1 \dots N$, we create a feature vector, $\mathbf{x}_i^{\mathrm{A2S}}$, by averaging for each of the A attributes the corresponding K scores:

$$\mathbf{x}_i^{\mathrm{A2S}} = [\bar{a}_{i,1}, \dots, \bar{a}_{i,A}], \qquad \bar{a}_{i,j} = \frac{1}{K} \sum_{k=1}^{K} a_{ijk}, \tag{1}$$

where i is the shape index (list of "fashion" or CAESAR bodies), j is the attribute index, and k the annotation index. We then define the full feature matrix for all N shapes as:

$$\mathbf{X}^{\mathrm{A2S}} = [\phi(\mathbf{x}_1^{\mathrm{A2S}}), \quad \dots, \quad \phi(\mathbf{x}_N^{\mathrm{A2S}})]^\top, \tag{2}$$

where $\phi(\mathbf{x}_i^{\mathrm{A2S}})$ maps $\mathbf{x}_i$ to 2nd order polynomial features. The target matrix $\mathbf{Y} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N]^\top$ contains the shape parameters $\boldsymbol{\beta}_i = [\boldsymbol{\beta}_{i,1}, \dots, \boldsymbol{\beta}_{i,B}]^\top$. We compute the polynomial model's coefficients $\boldsymbol{W}$ via least-squares fitting:

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \epsilon. \tag{3}$$

Empirically, the polynomial model performs better than several models that we evaluated; for details, see Sup. Mat.

Shape to Attributes (S2A): We predict linguistic attribute scores, A, from SMPL-X shape parameters, $\boldsymbol{\beta}$. Again, we fit a second-degree polynomial regression model. S2A has "swapped" inputs and outputs w.r.t. A2S:

$$\mathbf{x}_i^{\mathrm{S2A}} = [\boldsymbol{\beta}_{i,1}, \dots, \boldsymbol{\beta}_{i,B}], \tag{4}$$

$$\mathbf{y}_i = [\bar{a}_{i,1}, \dots, \bar{a}_{i,A}]^\top. \tag{5}$$

Attributes & Measurements to Shape (AHWC2S): Given a sparse set of anthropometric measurements, we predict SMPL-X shape parameters, $\boldsymbol{\beta}$. The input vector is:

$$\mathbf{x}_i^{\mathrm{HWC2S}} = [h_i, w_i, c_{c_i}, c_{w_i}, c_{h_i}], \tag{6}$$

where c_c, c_w, c_h is the chest, waist, and hip circumference, respectively, h and w are the height and weight, and HWC2S means *Height* + *Weight* + *Circumference to Shape*. The regression target is the SMPL-X shape parameters, $\mathbf{y}_i$.

When both *Attributes* and measurements are available, we combine them for the AHWC2S model with input:

$$\mathbf{x}_i^{\mathrm{AHWC2S}} = [\bar{a}_{i,1}, \dots, \bar{a}_{i,A}, h_i, w_i, c_{c_i}, c_{w_i}, c_{h_i}]. \tag{7}$$

In practice, depending on which measurements are available, we train and use different regressors. Following the naming convention of AHWC2S, these models are: AH2S, AHW2S, AC2S, and AHC2S, as well as their equivalents without attribute input H2S, HW2S, C2S, and HC2S. For an evaluation of the contribution of linguistic shape attributes on top of each anthropometric measurement, see Sup. Mat.

Training Data: To train the A2S and S2A mappings we use CAESAR data, for which we have SMPL-X shape parameters, anthropometric measurements, and linguistic attribute scores. We train separate gender-specific models.

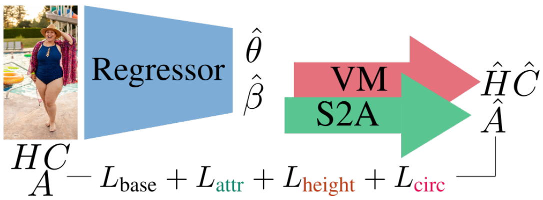

Figure 6. SHAPY first estimates shape, $\hat{\beta}$, and pose, $\hat{\theta}$. Shape is used by: (1) our virtual anthropometric measurement (VM) module to compute height, $\hat{H}$, and circumferences, $\hat{C}$, and (2) our S2A module to infer linguistic attribute scores, $\hat{A}$. There are several SHAPY variations, e.g., SHAPY-H uses only VM to infer $\hat{H}$, while SHAPY-HA uses VM to infer $\hat{H}$ and S2A to infer $\hat{A}$.

5. 3D Shape Regression from an Image

We present SHAPY, a network that predicts SMPL-X parameters from an RGB image with more accurate body shape than existing methods. To improve the realism and accuracy of shape, we explore training losses based on all shape representations discussed above, i.e., SMPL-X meshes (Sec. 3.1), linguistic attribute scores (Sec. 3.3) and anthropometric measurements (Sec. 4.1). In the following, symbols with/-out a hat are regressed/ground-truth values.

We convert shape $\hat{\boldsymbol{\beta}}$ to height and circumferences values $\{\hat{H}, \hat{C}_{\text{c}}, \hat{C}_{\text{w}}, \hat{C}_{\text{h}}\} = \{H(\hat{\boldsymbol{\beta}}), C_{\text{c}}(\hat{\boldsymbol{\beta}}), C_{\text{w}}(\hat{\boldsymbol{\beta}}), C_{\text{h}}(\hat{\boldsymbol{\beta}})\}$, by applying our virtual measurement tool (Sec. 4.1) to the mesh $M(\hat{\boldsymbol{\beta}})$ in the canonical T-pose. We also convert shape $\hat{\boldsymbol{\beta}}$ to linguistic attribute scores, with $\hat{A} = \text{S2A}(\hat{\boldsymbol{\beta}})$.

We train various SHAPY versions with the following "SHAPY losses", using either linguistic shape attributes, or anthropometric measurements, or both:

$$L_{\text{attr}} = ||A - \hat{A}||_2^2, \tag{8}$$

$$L_{\text{height}} = ||H - \hat{H}||_2^2, \tag{9}$$

$$L_{\text{circ}} = \sum\nolimits_{i \in \{c,w,h\}} ||C_i - \hat{C}_i||_2^2 \tag{10}$$

These are optionally added to a base loss, L_{base}, defined below in "training details". The architecture of SHAPY, with all optional components, is shown in Fig. 6. A suffix of color-coded letters describes which of the above losses are used when training a model. For example, SHAPY-AH denotes a model trained with the attribute and height losses, i.e.: $L_{\text{SHAPY-AH2S}} = L_{\text{base}} + L_{\text{attr}} + L_{\text{height}}$.

Training Details: We initialize SHAPY with the ExPose [9] network weights and use curated fits [9], H3.6M [23], the SPIN [33] training data, and our model-agency dataset (Sec. 3.2) for training. In each batch, 50% of the images are sampled from the model-agency images, for which we ensure a gender balance. The "SHAPY losses" of Eqs. (8) to (10) are applied only on the model-agency images. We use these on top of a standard base loss:

$$L_{\text{base}} = L_{\text{pose}} + L_{\text{shape}}, \tag{11}$$

where $L_{\text{joints}}^{\text{2D}}$ and $L_{\text{joints}}^{\text{3D}}$ are 2D and 3D joint losses:

$$L_{\text{pose}} = L_{\text{joints}}^{\text{2D}} + L_{\text{joints}}^{\text{3D}} + L_{\boldsymbol{\theta}}, \tag{12}$$

$$L_{\text{shape}} = L_{\boldsymbol{\beta}} + L_{\boldsymbol{\beta}}^{\text{prior}}, \tag{13}$$

$L_{\boldsymbol{\theta}}$ and $L_{\boldsymbol{\beta}}$ are losses on pose and shape parameters, and $L_{\boldsymbol{\beta}}^{\text{prior}}$ is PIXIE's [13] "gendered" shape prior. All losses are L2, unless otherwise explicitly specified. Losses on SMPL-X parameters are applied only on the pose data [9, 23, 33]. For more implementation details, see Sup. Mat.

6. Experiments

6.1. Evaluation Datasets

3D Poses in the Wild (3DPW) [61]: We use this to evaluate *pose* accuracy. This is widely used, but has only 5 test subjects, i.e., limited shape variation. For results, see Sup. Mat.

Sports Shape and Pose 3D (SSP-3D) [51]: We use this to evaluate 3D body *shape* accuracy from images. It has 62 tightly-clothed subjects in 311 in-the-wild images from Sports-1M [29], with *pseudo* ground-truth SMPL meshes that we convert to SMPL-X for evaluation.

Model Measurements Test Set (MMTS): We use this to evaluate anthropometric measurement accuracy, as a proxy for body *shape* accuracy. To create MMTS, we withhold 2699/1514 images of 143/95 female/male identities from our model-agency data, described in Sec. 3.2

CAESAR Meshes Test Set (CMTS): We use CAESAR to measure the accuracy of SMPL-X body shapes and linguistic shape attributes for the models of Sec. 4. Specifically, we compute: (1) errors for SMPL-X meshes estimated from linguistic shape attributes and/or anthropometric measurements by A2S and its variations, and (2) errors for linguistic shape attributes estimated from SMPL-X meshes by S2A. To create an unseen mesh test set, we withhold 339 male and 410 female CAESAR meshes from the crowd-sourced CAESAR linguistic shape attributes, described in Sec. 3.3.

Human Bodies in the Wild (HBW): The field is missing a dataset with varied bodies, varied clothing, in-the-wild images, and accurate *3D shape ground truth*. We fill this gap by collecting a novel dataset, called *"Human Bodies in the Wild" (HBW)*, with three steps: (1) We collect accurate 3D body scans for 35 subjects (20 female, 15 male), and register a "gendered" SMPL-X model to these to recover 3D SMPL-X ground-truth bodies [45]. (2) We take photos of each subject in "photo-lab" settings, i.e., in front of a white background with controlled lighting, and in various everyday outfits and "fashion" poses. (3) Subjects upload full-body photos of themselves taken in the wild. For each subject we take up to 111 photos in lab settings, and collect up to 126 in-the-wild photos. In total, HBW has 2543 photos, 1,318 in the lab setting and 1,225 in the wild. We split the data into a validation and a test

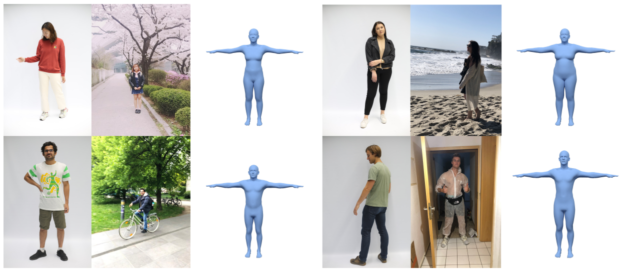

Figure 7. "Human Bodies in the Wild" (HBW) color images, taken in the lab and in the wild, and the SMPL-X ground-truth shape.

set (val/test) with 10/25 subjects (6/14 female 4/11 male) and 781/1,762 images (432/983 female 349/779 male), respectively. Figure 7 shows a few HBW subjects, photos and their SMPL-X ground-truth shapes. All subjects gave prior written informed consent to participate in this study and to release the data. The study was reviewed by the ethics board of the University of Tübingen, without objections.

6.2. Evaluation Metrics

We use standard accuracy metrics for 3D body pose, but also introduce metrics specific to 3D body shape.

Anthropometric Measurements: We report the mean absolute error in mm between ground-truth and estimated measurements, computed as described in Sec. 4.1. When weight is available, we report the mean absolute error in kg.

MPJPE and V2V metrics: We report in Sup. Mat. the mean per-joint point error (MPJPE) and mean vertex-to-vertex error (V2V), when SMPL-X meshes are available. The prefix "PA" denotes metrics after Procrustes alignment.

Mean point-to-point error ($\mathbf{P2P_{20K}}$): SMPL-X has a highly non-uniform vertex distribution across the body, which negatively biases the mean vertex-to-vertex (V2V) error, when comparing estimated and ground-truth SMPL-X meshes. To account for this, we evenly sample 20K points on SMPL-X's surface, and report the mean point-to-point ($P2P_{20K}$) error. For details, see Sup. Mat.

6.3. Shape-Representation Mappings

We evaluate the models A2S and S2A, which map between the various body shape representations (Sec. 4).

A2S and its variations: How well can we infer 3D body shape from just linguistic shape attributes, anthropometric measurements, or both of these together? In Tab. 2, we report reconstruction and measurement errors using many combinations of attributes (A), height (H), weight (W), and circumferences (C). Evaluation on CMTS data shows that attributes improve the overall shape prediction across the board. For example, height+attributes (AH2S) has a lower point-to-point error than height alone. The best performing model, AHWC, uses everything, with $P2P_{20K}$-errors of 5.8 ± 2.0 mm (males) and 6.2 ± 2.4 mm (females).

	Method -	$P2P_{20K}$ (mm)	Height (mm)	Weight (kg)	Chest (mm)	Waist (mm)	Hips (mm)
Male subjects	A2S	11.1 ± 5.2	29 ± 21	5 ± 4	30 ± 22	32 ± 24	28 ± 21
	H2S	12.1 ± 6.1	5 ± 4	11 ± 11	81 ± 66	102 ± 87	40 ± 33
	AH2S	6.8 ± 2.3	4 ± 3	3 ± 3	27 ± 21	29 ± 23	24 ± 18
	HW2S	8.1 ± 2.7	5 ± 4	1 ± 1	24 ± 17	26 ± 20	21 ± 18
	AHW2S	6.3 ± 2.1	4 ± 3	1 ± 1	19 ± 15	19 ± 14	20 ± 16
	C2S	19.7 ± 11.1	59 ± 47	9 ± 8	55 ± 41	63 ± 49	37 ± 28
	AC2S	9.6 ± 4.4	25 ± 19	3 ± 3	23 ± 19	21 ± 17	18 ± 14
	HC2S	7.7 ± 2.6	5 ± 4	2 ± 2	28 ± 23	18 ± 15	13 ± 11
	AHC2S	6.0 ± 2.0	4 ± 3	2 ± 2	21 ± 17	17 ± 14	13 ± 10
	HWC2S	7.3 ± 2.6	5 ± 4	1 ± 1	20 ± 15	14 ± 12	13 ± 11
	AHWC2S	5.8 ± 2.0	4 ± 3	1 ± 1	16 ± 13	13 ± 10	13 ± 10

Table 2. Results of A2S variants on CMTS for male subjects, using the male SMPL-X model. For females, see Sup. Mat.

Method	Model	Height	Chest	Waist	Hips	$P2P_{20K}$
SMPLR [38]	SMPL	182	267	309	305	69
STRAPS [51]	SMPL	135	167	145	102	47
SPIN [33]	SMPL	59	92	78	101	29
TUCH [40]	SMPL	58	89	75	**57**	26
Sengupta et al. [52]	SMPL	82	133	107	63	32
ExPose [9]	SMPL-X	85	99	92	94	35
SHAPY (ours)	SMPL-X	**51**	**65**	**69**	**57**	**21**

Table 3. Evaluation on the HBW test set in mm. We compute the measurement and point-to-point ($P2P_{20K}$) error between predicted and ground-truth SMPL-X meshes.

S2A: How well can we infer linguistic shape attributes from 3D shape? S2A's accuracy on inferring the attribute Likert score is $75\%/69\%$ for males/females; details in Sup. Mat.

6.4. 3D Shape from an Image

We evaluate all of our model's variations (see Sec. 5) on the HBW validation set and find, perhaps surprisingly, that SHAPY-A outperforms other variants. We refer to this below (and Fig. 1) simply as "SHAPY" and report its performance in Tab. 3 for HBW, Tab. 4 for MMTS, and Tab. 5 for SSP-3D. For images with natural and varied clothing (HBW, MMTS), SHAPY significantly outperforms all other methods (Tabs. 3 and 4) using only weak 3D shape supervision (Attributes). On these images, Sengupta et al.'s method [52] struggles with the natural clothing. In contrast, their method is more accurate than SHAPY on SSP-3D (Tab. 5), which has tight "sports" clothing, in terms of PVE-T-SC, a scale-normalized metric used on this dataset. These results show that silhouettes are good for tight/minimal clothing and that SHAPY struggles with high BMI shapes due to the lack of such shapes in our training data; see Fig. 5. Note that, as HBW has true ground-truth 3D shape, it does not need SSP-3D's scaling for evaluation.

A key observation is that training with linguistic shape attributes alone is sufficient, i.e., without anthropometric measurements. Importantly, this opens up the possibility for significantly larger data collections. For a study of how different measurements or attributes impact accuracy, see Sup. Mat. Figure 8 shows SHAPY's qualitative results.

Figure 8. Qualitative results from HBW. From left to right: RGB, ground-truth shape, SHAPY and Sengupta et al. [52]. For example, in the upper- and lower- right images, SHAPY is less affected by pose variation and loose clothing.

		Mean absolute error (mm) ↓			
Method	Model	Height	Chest	Waist	Hips
Sengupta et al. [52]	SMPL	84	186	263	142
TUCH [40]	SMPL	82	92	129	91
SPIN [33]	SMPL	72	91	129	101
STRAPS [51]	SMPL	207	278	326	145
ExPose [9]	SMPL-X	107	107	136	92
SHAPY (ours)	SMPL-X	**71**	**64**	**98**	**74**

Table 4. Evaluation on MMTS. We report the mean absolute error between ground-truth and estimated measurements.

Method	Model	PVE-T-SC	mIOU
HMR [27]	SMPL	22.9	0.69
SPIN [33]	SMPL	22.2	0.70
STRAPS [51]	SMPL	15.9	**0.80**
Sengupta et al. [52]	SMPL	**13.6**	-
SHAPY (ours)	SMPL-X	19.2	-

Table 5. Evaluation on the SSP-3D test set [51]. We report the scaled mean vertex-to-vertex error in T-pose [51], and mIOU.

7. Conclusion

SHAPY is trained to regress more accurate human body shape from images than previous methods, without explicit 3D shape supervision. To achieve this, we present two different ways to collect proxy annotations for 3D body shape for in-the-wild images. First, we collect sparse anthropometric measurements from online model-agency data. Second, we annotate images with linguistic shape attributes using crowd-sourcing. We learn mappings between body shape, measurements, and attributes, enabling us to supervise a regressor using any combination of these. To evaluate SHAPY, we introduce a new shape estimation benchmark, the "Human Bodies in the Wild" (HBW) dataset. HBW has images of people in natural clothing and natural settings together with ground-truth 3D shape from a body scanner. HBW is more challenging than existing shape benchmarks like SSP-3D, and SHAPY significantly outperforms existing methods on this benchmark. We believe this work will open new directions, since the idea of leveraging linguistic annotations to improve 3D shape has many applications.

Limitations: Our model-agency training dataset (Sec. 3.2) is not representative of the entire human population and this limits SHAPY's ability to predict larger body shapes. To address this, we need to find images of more diverse bodies together with anthropometric measurements and linguistic shape attributes describing them.

Social impact: Knowing the 3D shape of a person has advantages, for example, in the clothing industry to avoid unnecessary returns. If used without consent, 3D shape estimation may invade individuals' privacy. As with all other 3D pose and shape estimation methods, surveillance and deep-fake creation is another important risk. Consequently, SHAPY's license prohibits such uses.

Acknowledgments: This work was supported by the Max Planck ETH Center for Learning Systems and the International Max Planck Research School for Intelligent Systems. We thank Tsvetelina Alexiadis, Galina Henz, Claudia Gallatz, and Taylor McConnell for the data collection, and Markus Höschle for the camera setup. We thank Muhammed Kocabas, Nikos Athanasiou and Maria Alejandra Quiros-Ramirez for the insightful discussions.

Disclosure: https://files.is.tue.mpg.de/black/CoI_CVPR_2022.txt

References

[1] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(1):44–58, 2006. 3

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

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

[4] Alexandru Balan and Michael J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision (ECCV)*, volume 5304, pages 15–29, 2008. 3

[5] Alexandru O. Balan, Leonid Sigal, Michael J. Black, James E. Davis, and Horst W. Haussecker. Detailed human shape and pose from images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 3

[6] Didier Bieler, Semih Gunel, Pascal Fua, and Helge Rhodin. Gravity as a reference for estimating a person's height from video. In *International Conference on Computer Vision (ICCV)*, pages 8568–8576, 2019. 4

[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 *European Conference on Computer Vision (ECCV)*, volume 9909, pages 561–578, 2016. 1, 3

[8] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2019. 3

[9] 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 (ECCV)*, volume 12355, pages 20–40, 2020. 3, 6, 7, 8

[10] Antonio Criminisi, Ian Reid, and Andrew Zisserman. Single view metrology. *International Journal of Computer Vision (IJCV)*, 40(2):123–148, 2000. 4

[11] Ratan Dey, Madhurya Nangia, Keith W. Ross, and Yong Liu. Estimating heights from photo collections: A data-driven approach. In *Conference on Online Social Networks (COSN)*, page 227–238, 2014. 4

[12] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to regress bodies from images using differentiable semantic rendering. In *International Conference on Computer Vision (ICCV)*, pages 11250–11259, 2021. 3

[13] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pages 792–804, 2021. 6

[14] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *European Conference on Computer Vision (ECCV)*, volume 12362, pages 768–784, 2020. 3

[15] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7450–7459, 2019. 3

[16] Peng Guan, Alexander Weiss, Alexandru Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *International Conference on Computer Vision (ICCV)*, pages 1381–1388, 2009. 3

[17] Semih Gunel, Helge Rhodin, and Pascal Fua. What face and body shapes can tell us about height. In *International Conference on Computer Vision Workshops (ICCVw)*, pages 1819–1827, 2019. 4

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(2):386–397, 2020. 3

[19] Matthew Hill, Stephan Streuber, Carina Hahn, Michael Black, and Alice O'Toole. Exploring the relationship between body shapes and descriptions by linking similarity spaces. *Journal of Vision (JOV)*, 15(12):931–931, 2015. 4

[20] David T. Hoffmann, Dimitrios Tzionas, Michael J. Black, and Siyu Tang. Learning to train with synthetic humans. In *German Conference on Pattern Recognition (GCPR)*, pages 609–623, 2019. 3

[21] Wei-Lin Hsiao and Kristen Grauman. ViBE: Dressing for diverse body shapes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11056–11066, 2020. 3

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

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2013. 3, 6

[24] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12753–12762, 2021. 3

[25] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5578–5587, 2020. 3

[26] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 1, 3

[27] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 3, 8

[28] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5614–5623, 2019. 1

[29] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014. 6

[30] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2020. 1, 3

[31] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 1

[32] 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 *International Conference on Computer Vision (ICCV)*, pages 11035–11045, 2021. 1

[33] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 1, 3, 6, 7, 8

[34] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 6050–6059, 2017. 3

[35] Junbang Liang and Ming C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *International Conference on Computer Vision (ICCV)*, pages 4351–4361, 2019. 2, 3

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755, 2014. 3

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 2, 3

[38] Meysam Madadi, Hugo Bertiche, and Sergio Escalera. SMPLR: Deep learning based SMPL reverse for 3D human pose and shape recovery. *Pattern Recognition (PR)*, 106:107472, 2020. 7

[39] 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 (ICCV)*, pages 5442–5451, 2019. 3

[40] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, 2021. 3, 7, 8

[41] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, pages 484–494, 2018. 3

[42] Priyanka Patel, Chun-Hao Paul Huang, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021. 3

[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 *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 3, 4

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

[45] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *Transactions on Graphics (TOG)*, 34(4):120:1–120:14, 2015. 6

[46] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H Bülthoff, and Michael J. Black. The virtual caliper: Rapid creation of metrically accurate avatars from 3D measurements. *Transactions on Visualization and Computer Graphics (TVCG)*, 25(5):1887–1897, 2019. 3, 4, 5

[47] Kathleen M. Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoeferlin, and Dennis Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Technical Report AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory, 2002. 2, 4, 5

[48] Nadine Rueegg, Christoph Lassner, Michael J. Black, and Konrad Schindler. Chained representation cycling: Learning to estimate 3D human pose and shape by cycling between representations. In *Conference on Artificial Intelligence (AAAI)*, pages 5561–5569, 2020. 3

[49] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 3

[50] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 81–90, 2020. 3

[51] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3D human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, 2020. 2, 3, 6, 7, 8

[52] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3D human

shape and pose estimation from images in the wild. In *International Conference on Computer Vision (ICCV)*, pages 11219–11229, 2021. 1, 2, 3, 7, 8

[53] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 16094–16104, 2021. 3

[54] Hyewon Seo, Frederic Cordier, and Nadia Magnenat-Thalmann. Synthesizing animatable body models with parameterized shape modifications. In *Symposium on Computer Animation (SCA)*, pages 120–125, 2003. 3

[55] Hyewon Seo and Nadia Magnenat-Thalmann. An automatic modeling of human bodies from sizing parameters. In *Symposium on Interactive 3D Graphics (SI3D)*, pages 19–26, 2003. 3

[56] Leonid Sigal, Alexandru Balan, and Michael J Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87(1):4–27, 2010. 3

[57] Stephan Streuber, M. Alejandra Quiros-Ramirez, Matthew Q. Hill, Carina A. Hahn, Silvia Zuffi, Alice O'Toole, and Michael J. Black. Body Talk: Crowdshaping realistic 3D avatars with words. *Transactions on Graphics (TOG)*, 35(4):54:1–54:14, 2016. 2, 4, 5

[58] Aggeliki Tsoli, Matthew Loper, and Michael J. Black. Model-based anthropometry: Predicting measurements from 3D human scans in multiple poses. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 83–90, 2014. 3

[59] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision (ECCV)*, volume 11211, pages 20–38, 2018. 3

[60] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017. 3

[61] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, volume 11214, pages 614–631, 2018. 3, 6

[62] Andrew Weitz, Lina Colucci, Sidney Primas, and Brinnae Bent. InfiniteForm: A synthetic, minimal bias dataset for fitness applications. *arXiv:2110.01330*, 2021. 3

[63] Stefanie Wuhrer and Chang Shu. Estimating 3D human shapes from measurements. *Machine Vision and Applications (MVA)*, 24(6):1133–1147, 2013. 5

[64] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[65] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020. 1, 3

[66] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3D human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision (ECCV)*, volume 12351, pages 465–481, 2020. 3

[67] 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 *International Conference on Computer Vision (ICCV)*, pages 11446–11456, 2021. 1

[68] Rui Zhu, Xingyi Yang, Yannick Hold-Geoffroy, Federico Perazzi, Jonathan Eisenmann, Kalyan Sunkavalli, and Manmohan Chandraker. Single view metrology in the wild. In *European Conference on Computer Vision (ECCV)*, volume 12356, pages 316–333, 2020. 4