Accurate 3D Body Shape Regression using Metric and Semantic Attributes

Vasileios Choutas\textsuperscript{*1}, Lea Müller\textsuperscript{*1}, Chun-Hao P. Huang\textsuperscript{1}, Siyu Tang\textsuperscript{2}, Dimitrios Tzionas\textsuperscript{1}, Michael J. Black\textsuperscript{1}
\textsuperscript{1}Max Planck Institute for Intelligent Systems, Tübingen, Germany \textsuperscript{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.mp.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]. Un-
fortunately, 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 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
2. Related Work

3D human pose and shape (HPS): Methods that re-construct 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.

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

### 3.1. SMPL-X Body Model

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

### 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
of subjects. In the following, their task is to “indicate how strongly [they] agree” with a 5-point Likert scale (Sec. 3.3). Here we learn mappings between metric measurements (Sec. 3.2), and (3) linguistic shape attributes (Sec. 3.1), (2) anthropometric measurements, and (1) SMPL-X’s PCA shape space (Sec. 3.1), (2) anthropometric measurements, and (3) linguistic shape attributes, applies to both genders, but others are gender specific.

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, \( \beta \). 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 attribute scores 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 \( \bar{X} \in \{1, 2, 3, 4, 5\}^{N \times A \times K} \), where \( N \) is the number of subjects. In the following, \( a_{i,j,k} \) denotes an element of \( A \).

### 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, \( H(\beta) \), weight, \( W(\beta) \), and chest, waist and hip circumferences, \( C_c(\beta), C_w(\beta), \) and \( C_h(\beta) \), 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 \( \beta_i \), \( i = 1 \ldots N \), we create a feature vector, \( x_i^{\text{A2S}} \), by averaging for each of the \( A \) attributes the corresponding \( K \) scores:

\[
x_i^{\text{A2S}} = [\bar{a}_{i,1}, \ldots, \bar{a}_{i,A}], \quad \bar{a}_{i,j} = \frac{1}{K} \sum_{k=1}^{K} a_{i,j,k},
\]

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:

\[
X^{\text{A2S}} = [\phi(x_1^{\text{A2S}}), \ldots, \phi(x_N^{\text{A2S}})]^T,
\]

where \( \phi(x_i^{\text{A2S}}) \) maps \( x_i \) to 2nd order polynomial features. The target matrix \( Y = [\beta_1, \ldots, \beta_N]^T \) contains the shape parameters \( \beta_i = [\beta_{i,1}, \ldots, \beta_{i,B}]^T \). We compute the polynomial model’s coefficients \( W \) via least-squares fitting:

\[
Y = XW + \epsilon.
\]

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, \( \beta \). Again, we fit a second-degree polynomial regression model. S2A has “swapped” inputs and outputs w.r.t. A2S:

\[
x_i^{\text{S2A}} = [\beta_{i,1}, \ldots, \beta_{i,B}],
\]

\[
y_i = [\bar{a}_{i,1}, \ldots, \bar{a}_{i,A}]^T.
\]

**Attributes & Measurements to Shape (AHWC2S):**

Given a sparse set of anthropometric measurements, we predict SMPL-X shape parameters, \( \beta \). The input vector is:

\[
x_i^{\text{AHWC2S}} = [h_i, w_i, c_{c_i}, c_{w_i}, c_{h_i}],
\]

where \( c_{c_i}, c_{w_i}, c_{h_i} \) 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, \( y_i \).

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

\[
x_i^{\text{AHWC2S}} = [\bar{a}_{i,1}, \ldots, \bar{a}_{i,A}, h_i, w_i, c_{c_i}, c_{w_i}, c_{h_i}].
\]
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 \( \beta \) to height and circumferences values \( \{H, C_c, C_w, C_h\} = \{H(\hat{\beta}), C_c(\hat{\beta}), C_w(\hat{\beta}), C_h(\hat{\beta)}\) by applying our virtual measurement tool (Sec. 4.1) to the mesh \( \hat{M}(\hat{\beta}) \) in the canonical T-pose. We also convert shape \( \beta \) to linguistic attribute scores, with \( \hat{A} = \text{S2A}(\hat{\beta}) \).

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

\[
\begin{align*}
L_{\text{attr}} &= ||A - \hat{A}||^2_2, \\
L_{\text{height}} &= ||H - \hat{H}||^2_2, \\
L_{\text{circ}} &= \sum_{i \in \{c,w,h\}} ||C_i - \hat{C}_i||^2_2.
\end{align*}
\]

These are optionally added to a base loss, \( L_{\text{base}} \), 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., \( L_{\text{SHAPE-AH2S}} = L_{\text{base}} + L_{\text{attr}} + L_{\text{height}} \).

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:

\[
L_{\text{base}} = L_{\text{pose}} + L_{\text{shape}},
\]

where \( L_{\text{pose}} \) and \( L_{\text{shape}} \) are 2D and 3D joint losses:

\[
L_{\text{pose}} = L_{\text{ joints}}^2 + L_{\text{ joints}}^3 + L_{\theta},
\]

\[
L_{\text{shape}} = L_{\beta} + L_{\beta}^\text{prior},
\]

\( L_{\theta} \) and \( L_{\beta} \) are losses on pose and shape parameters, and \( L_{\beta}^\text{prior} \) 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
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 (P2P20K): 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 (P2P20K) 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 P2P20K-errors of 5.8 ± 2.0 mm (males) and 6.2 ± 2.4 mm (females).

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>Height (mm)</th>
<th>Chest (mm)</th>
<th>Waist (mm)</th>
<th>Hips (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2S</td>
<td>SMPL-X</td>
<td>11.1 ± 5.2</td>
<td>29 ± 21</td>
<td>5 ± 4</td>
<td>30 ± 22</td>
</tr>
<tr>
<td>H2S</td>
<td>SMPL-X</td>
<td>12.1 ± 6.1</td>
<td>5 ± 4</td>
<td>11 ± 11</td>
<td>81 ± 66</td>
</tr>
<tr>
<td>AH2S</td>
<td>SMPL-X</td>
<td>6.8 ± 2.3</td>
<td>4 ± 3</td>
<td>3 ± 3</td>
<td>27 ± 21</td>
</tr>
<tr>
<td>HW2S</td>
<td>SMPL-X</td>
<td>8.1 ± 2.7</td>
<td>5 ± 4</td>
<td>1 ± 1</td>
<td>24 ± 17</td>
</tr>
<tr>
<td>AHWC2S</td>
<td>SMPL-X</td>
<td>6.3 ± 2.1</td>
<td>4.3 ± 3</td>
<td>1 ± 1</td>
<td>19 ± 15</td>
</tr>
<tr>
<td>CS2S</td>
<td>SMPL-X</td>
<td>19.7 ± 11.1</td>
<td>59 ± 47</td>
<td>5 ± 8</td>
<td>55 ± 41</td>
</tr>
<tr>
<td>A2S</td>
<td>SMPL-X</td>
<td>9.6 ± 4.4</td>
<td>25 ± 19</td>
<td>3 ± 3</td>
<td>23 ± 19</td>
</tr>
<tr>
<td>H2S</td>
<td>SMPL-X</td>
<td>7.7 ± 2.6</td>
<td>5 ± 4</td>
<td>2 ± 2</td>
<td>28 ± 23</td>
</tr>
<tr>
<td>AH2S</td>
<td>SMPL-X</td>
<td>6.0 ± 2.0</td>
<td>4.3 ± 3</td>
<td>2 ± 2</td>
<td>21 ± 17</td>
</tr>
<tr>
<td>AHWC2S</td>
<td>SMPL-X</td>
<td>7.3 ± 2.6</td>
<td>5 ± 4</td>
<td>1 ± 1</td>
<td>20 ± 15</td>
</tr>
<tr>
<td>CS2S</td>
<td>SMPL-X</td>
<td>5.8 ± 2.0</td>
<td>4 ± 3</td>
<td>1 ± 1</td>
<td>16 ± 13</td>
</tr>
</tbody>
</table>

Table 3. Evaluation on the HBW test set in mm. We compute the measurement and point-to-point (P2P20K) error between predicted and ground-truth SMPL-X meshes.

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>Height (mm)</th>
<th>Chest (mm)</th>
<th>Waist (mm)</th>
<th>Hips (mm)</th>
<th>P2P20K</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMPLR [38]</td>
<td>SMPL</td>
<td>182 ± 267</td>
<td>309 ± 309</td>
<td>305 ± 69</td>
<td>24 ± 14</td>
<td></td>
</tr>
<tr>
<td>STRAPS [51]</td>
<td>SMPL</td>
<td>135 ± 167</td>
<td>145 ± 102</td>
<td>47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPIN [33]</td>
<td>SMPL</td>
<td>59 ± 92</td>
<td>78 ± 101</td>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TUCH [69]</td>
<td>SMPL</td>
<td>58 ± 89</td>
<td>75 ± 57</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sengupta et al. [52]</td>
<td>SMPL</td>
<td>82 ± 133</td>
<td>107 ± 63</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ExPose [9]</td>
<td>SMPL-X</td>
<td>85 ± 99</td>
<td>92 ± 94</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHAPY (ours)</td>
<td>SMPL-X</td>
<td>51 ± 65</td>
<td>69 ± 57</td>
<td>21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Results of A2S variants on MMTS for male subjects, using the male SMPL-X model. For females, see 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 3D shape? S2A’s accuracy on inferring the attribute Likert score is 75%/69% for males/females; details in Sup. Mat.

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.
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 Höschele 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/Col_CVPR_2022.txt
References

[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


