Abstract

Accurate human shape recovery from a monocular RGB image is a challenging task because humans come in different shapes and sizes and wear different clothes. In this paper, we propose ShapeBoost, a new human shape recovery framework that achieves pixel-level alignment even for rare body shapes and high accuracy for people wearing different types of clothes. Unlike previous approaches that rely on the use of PCA-based shape coefficients, we adopt a new human shape parameterization that decomposes the human shape into bone lengths and the mean width of each part slice. This part-based parameterization technique achieves a balance between flexibility and validity using a semi-analytical shape reconstruction algorithm. Based on this new parameterization, a clothing-preserving data augmentation module is proposed to generate realistic images with diverse body shapes and accurate annotations. Experimental results show that our method outperforms other state-of-the-art methods in diverse body shape situations as well as in varied clothing situations.

1 Introduction

Human pose and shape (HPS) recovery from monocular RGB images is an essential task of computer vision. It serves as a basis for human behavior understanding and has applications in various fields such as Virtual Reality, Augmented Reality, and Autopilot. Recent methods (Zhang et al. 2022; Li et al. 2022b,a, 2021) achieve high accuracy in human pose estimation, but their results of human shape estimation are often suboptimal.

Due to the scarcity of image datasets featuring diverse body shapes, many existing methods for recovering human pose and shape suffer from overfitting on body shape estimation. Their results are particularly unsatisfactory for very thin or plump people. Previous approaches have attempted to solve the overfitting issue through two main strategies. The first kind of methods (Varol et al. 2017; Sengupta, Budvytis, and Cipolla 2020, 2021b,a) train on synthetic data and exploit proxy representations to reduce the domain gap, while the second kind of methods (Dwivedi et al. 2021; Omran et al. 2018; Agarwal and Triggs 2005) exploit shape cues which are easy to annotate as weak supervision. However, for the first kind of methods, the synthetic images are unnatural with unrealistic texture and clothing, and the extracted proxy representations may be ambiguous and inaccurate. The situation is especially severe when the individual is wearing thick garments or is occluded in the image. For the second kind of methods, since 2D clues such as segmentations and silhouettes are highly correlated with the human pose and clothing, supervising with 2D clues may give wrong guidance of human shape in the case of inaccurate pose estimation or thick clothing. Moreover, the real-world images of extreme shapes are still insufficient. SHAPY (Choutas et al. 2022) improves the second kind of methods by using linguistic attributes and body measurements as supervision, which allows it making better estimates for clothed people. However, similar to other models trained on real-world datasets, it still performs poorly on images of people with extreme body shapes because of the lack of extreme body shapes in the training datasets. To sum up, just as shown in Fig. 1, the first kind of methods often fail on images with people in occlusion or thick clothing, while the second kind of methods often fail on images containing people with extreme body shapes.

To overcome the above limitations, we propose ShapeBoost, a new shape recovery framework based on a novel part-based shape parameterization. The new shape parameters are composed of bone lengths and mean widths of body part slices. Using a novel semi-analytical algorithm, the body shape can be accurately and robustly recovered from these parameters. During training, the bone lengths can be calculated from human keypoints, and the part widths are regressed by the neural network. Compared to the original shape parameters derived from PCA coefficients, our new part-based parameterization has a clear local semantic meaning, making it easier to regress and more flexible in application. During training, ShapeBoost augments new image-shape pairs by randomly transforming the raw image and calculating the corresponding part-based parameters. For image transformation, a clothing-preserving augmentation method is proposed: we first segment the human body out of the image and randomly transform it into a different shape. Then, the human segmentation is pasted back onto the inpainted background image with the guidance of the appearance consistency heatmap (Fang et al. 2019). The corresponding shape parameters can be analytically retrieved by
applying the equivalent transformation since each component in the part-based representation is clearly defined.

Compared to previous approaches, ShapeBoost generates realistic images of diverse human shapes in natural clothing together with the corresponding faithful annotations. Moreover, our new parameterization accurately describes the extreme body shapes and encourages pixel-level alignment. As a result, our method overcomes the disadvantages of existing methods and achieves high accuracy on images of people in thick clothes as well as on images of people with extreme body shapes. We benchmark our method on SSP-3D (Sengupta, Budvytis, and Cipolla 2020) and HBW (Choutas et al. 2022) datasets. The results show that our method achieves state-of-the-art performance in both thick clothes situations and extreme body shape situations.

The main contributions of this paper are summarized as follows:

- We present an accurate and robust human shape parameterization together with a semi-analytical shape recovery algorithm, which is flexible and interpretable.
- We propose ShapeBoost, a human shape recovery framework consisting of the clothing-preserving data augmentation module and a shape reconstruction module.
- Our approach outperforms previous approaches and can handle diverse clothing as well as extreme body shapes.

2 Related Work

2.1 3D Human Pose and Shape (HPS)

Many algorithms have been proposed for reconstructing human pose and shape from RGB images, which are broadly categorized into two types. Firstly, **model-based methods** estimate parameters of a parameterized human model. Some methods (Bogo et al. 2016; Pavlakos et al. 2019; Guan et al. 2009) estimate human pose and shape parameters by optimization. Regression-based methods (Kanazawa et al. 2018; Kocabas, Athanasiou, and Black 2020; Kocabas et al. 2021; Li et al. 2022b, 2021), on the contrary, employ neural networks to estimate the parameters. To reduce the difficulty of regression, many regression-based methods employ intermediate representations, including keypoints (Kanazawa et al. 2018), silhouettes (Pavlakos et al. 2018), segmentation (Omran et al. 2018) and 2D/3D heatmaps (Tung et al. 2017), keypoints (Li et al. 2021, 2023b,a) etc. Some approaches (Kolotouros et al. 2019; Muller et al. 2021; Joo, Neverova, and Vedaldi 2021) combine optimization and regression. Secondly, **model-free methods** directly predict free-form representations of the human body, with the position of body model vertices predicted based on image features (Corona et al. 2022; Kolotouros, Pavlakos, and Daniilidis 2019; Varol et al. 2018; Lin, Wang, and Liu 2021a,b; Moon and Lee 2020), keypoints (Choi, Moon, and Lee 2020), or segmentations (Varol et al. 2018). These methods mostly focus on human pose estimation and their results of human shape estimation are often unsatisfactory.

Our work belongs to the model-based category, and we adopt inverse kinematics to estimate the human pose similar to HybrIK (Li et al. 2021) for simplicity. However, instead of directly regressing the shape parameters, we employ a flexible and interpretable parameterization and a new shape reconstruction pipeline to achieve more accurate and robust shape estimation. Our method can also be easily applied to different pose estimation backbones.

2.2 Estimating 3D Body Shape

Most recent HPS estimation methods excel in precise pose estimation but exhibit limitations in accurately estimating the real human body shape under clothing. Some methods have attempted to address this issue, and they mainly focus on novel training datasets and the estimation framework.

**Training datasets for human shape estimation.** Accurately annotating body shapes from 2D human datasets (Lin et al. 2014) is hard, and commonly-used 3D human datasets (von Marcard et al. 2018; Ionescu et al. 2013) con-
Figure 2: The overall pipeline. First, the input image is randomly transformed with the clothing-preserving image transformation, and a convolutional neural network (CNN) is employed to extract skeleton, part widths and twist rotations. Then, the pose is obtained using inverse kinematics and the shape is obtained with our semi-analytical algorithm. The final mesh is retrieved based on the pose and shape parameter. The ShapeBoost framework consists of the image augmentation module and the shape reconstruction module.

In this section, we present our solution for human shape recovery (Fig. 2). First, we give background knowledge of the parameterization of SMPL model in Sec. 3.1. Considering its drawbacks, a flexible and interpretable part-based human shape parameterization is proposed in Sec. 3.2. Based on this new parameterization, in Sec. 3.3, we design a new human shape recovery framework called ShapeBoost. The training pipeline and loss functions are described in Sec. 3.4.

3 Method

3.1 Preliminary

SMPL Model. In this work, SMPL model (Loper et al. 2015) is employed to represent human body pose and shape. SMPL provides a differentiable function \( V(\theta, \beta) \) that maps pose \( \theta \in \mathbb{R}^{10} \) and shape parameters \( \beta \in \mathbb{R}^{10} \) to a human mesh \( V \), where \( J \) is the number of joints. The pose parameters \( \theta \) represent the relative rotation of body joints, and the shape parameters \( \beta \) are coefficients of a PCA body shape basis. SMPL model is driven in two steps:

\[
T = S(\beta), \quad (1)
\]

\[
V = V(\theta, \beta) = P(\theta, S(\beta)). \quad (2)
\]

First, a rest-pose mesh \( T \) is constructed using function \( S \). Second, the rest-pose mesh is driven to the target pose by function \( P \). The shape of the mesh is determined only by \( \beta \), and the posing procedure does not change the body shape.

Most current methods regress shape parameters \( \beta \) directly. However, since most available training datasets lack people with diverse body shapes, these methods often overfit and fail to generalize to unseen body shapes.
In this work, we propose a novel parameterization of human shape using bone lengths and widths of part slices. Compared to the $\beta$ representation which uses a global descriptor of the body shape, this new representation allocates shape descriptors to local body parts. This allows the network to learn from local image features and thus alleviates the overfitting problem. Furthermore, our parameterization is more flexible and interpretable, allowing compatibility with the our data augmentation procedure discussed in Sec. 3.3.

In our parameterization, the SMPL mesh is divided into $J = 24$ segments according to the linear blending weight, and each segment has a corresponding central bone ended with two joints. The distance of one vertex from its corresponding bone is called the “width” of this vertext for short. Each body part is further sliced into $n$ components along the bone, and the mean widths of the vertices in these $n$ slices are used to represent the thickness of that part. The segmenting and slicing technique is visually illustrated in Fig. 3. In this way, the formula of SMPL model is converted to:

$$
T = M(I, w),
$$

$$
V = P(\theta, M(I, w)),
$$

where $I \in R^{J-1}$ represents the bone lengths of the body skeleton and $w \in R^{nJ}$ represents the mean widths of all part slices. Under our new representation, the SMPL model first derives a rest-pose mesh using $M(I, w)$, and then uses function $P$ to drive the mesh to the target pose just like the original SMPL model.

Deriving the function $M$ directly by a neural network is untrivial and can lead to overfitting. Therefore, a semi-analytical algorithm is proposed that first solves a roughly correct mesh using analytical methods and then uses a multi-layer perceptron (MLP) to correct the result using error feedback techniques.

We can analytically retrieve a body shape that roughly conforms to the target bone lengths and part slice widths by (1) stretching the bones and broadening each part slice of the template mesh according to the target values. (2) using linear blend weights (LBS weights) to assemble these adjusted parts. (3) using the PCA-coefficients of SMPL to retrieve the shape parameters from the deformed template mesh. This mapping is referred to as $M_0$.

Since the input bone lengths and part widths often contain noise, the analytical algorithm sometimes produces suboptimal body shapes. Therefore, we use a 4-layer MLP to modify the analytically-retrieved shape parameters. The final formula of $M$ can be written as

$$
T = M(I, w) = MLP(M_0(I, w), l, w, \Delta l, \Delta w),
$$

where $\Delta l$ and $\Delta w$ are the difference between the target bone lengths and part slice widths and the corresponding values obtained by $M_0$. In practice, instead of regressing the bone lengths directly, we extract the bone lengths from human keypoints. This setting further encourages the network to only focus on local, per-part image features and thus alleviate overfitting.

### 3.3 ShapeBoost

Armed with the part-based parameterization discussed in Sec 3.2, we can manipulate the body shape in an intuitive way by stretching the bone lengths and broadening the part slice widths. These manipulations enable us to augment the raw human images and retrieve the new ground truth body shape which accurately explains the figure in the image after the transformation. This framework, named ShapeBoost, generates diverse body shapes while preserving clothing, lighting, and background details, and then takes use of our new parameterization to reconstruct the body shape.

**Clothing-preserving Image Transformation.** An intuitive way to change the human shape in an image is to apply the affine transformation to the input image. For example, scaling an image with an aspect ratio unequal to 1 results in a visually thinner or ampler human figure.

However, applying the affine transform to the entire image results in a stretched background, which may leak the scaling information and thus incur overfitting. To alleviate this problem, we propose a silhouette-based augmentation method inspired by Instaboost (Fang et al. 2019). Instead of affine transforming the whole image, we first segment the human body out using the ground truth segmentation. Then we inpaint the background image, affine transform the segmented human body, and paste the transformed human body back onto the inpainted background image with the guidance of the appearance consistency heatmap (Fang et al. 2019). This method effectively avoids background stretching and produces more natural-looking images. The process is visually illustrated in Fig. 4.
To simplify the discussion, we assume that the affine transformation consists of a rotation matrix and a scaling matrix, which is written as

\[
T = SR = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}.
\] (6)

**Shape-parameter Derivation.** People in different poses are affected by the image transformation in different ways, which poses a great challenge for the derivation of the PCA-based shape parameters after the image transformation. However, with the part-based parameterization, we can still accurately explain the new body shape by estimating the widths and bone lengths of each body part. We use orthographic projection in our derivation.

Given the camera and pose parameters, the bone lengths after transformation can be easily obtained by stretching the bones to ensure a consistent 2D joint projection. Compared to the derivation of bone lengths, the derivation of the part slice widths after transformation is more complex. Suppose a vertex indexed by \(k\) belongs to the \(j\)-th part. The distance of the vertex from the part bone on the 2D image plane, denoted by \(w_{k}^{2D}\), is affected by the transformation according to the following equations:

\[
w_{k}^{2D} = \frac{ab}{\bar{s}j} \cdot \frac{\bar{l}_{j}^{2D}}{l_{j}^{2D}} \cdot w_{j},
\] (7)

where \(\bar{l}_{j}^{2D}\) and \(l_{j}^{2D}\) represent the bone lengths of part \(j\) on the 2D image plane before and after the transformation, respectively; \(a\) and \(b\) are scaling factors mentioned in Eq. 6. A detailed derivation is available in the supplementary materials. It is noteworthy that Eq. 7 implies the 2D widths of vertices on the same part are scaled by the same factor. Therefore, the underlining 3D part width of part \(j\) is changed by

\[
\tilde{w}_{j} = \frac{s}{\tilde{s}} \times \frac{ab}{l_{j}^{2D}} \cdot \tilde{l}_{j} \times w_{j}.
\] (8)

In the equation, \(s\) and \(w_{j}\) are the scale factor of the orthographic projection and the 3D part width of part \(j\) before the image transformation, whereas \(\tilde{s}\) and \(\tilde{w}_{j}\) are the corresponding values after the transformation. Due to scale ambiguity, \(\tilde{s}\) is an ambiguous scaling factor that is difficult to directly derive. Therefore, in our training, we only supervise the projected results of the predicted part slice widths on the 2D image plane, without directly supervising their actual values. We hypothesize that the network can learn the best scaling factor \(\tilde{s}\) using the prior knowledge of human body shape.

### 3.4 Training Pipeline and Loss Function

The overall training pipeline is illustrated in Fig. 2. First, the input image is transformed using the clothing-preserving image transformation, and the convolutional neural network (CNN) backbone is utilized to process the augmented image and find the skeleton (3D keypoints extracted from heatmaps), twist angles and part slice widths. Second, we use these estimated values to reconstruct the pose and shape of the individual. The pose parameters are obtained with inverse kinematics similar to HybrIK (Li et al. 2021), while the shape parameters are retrieved using the semi-analytical algorithm discussed in Sec. 3.2. The final mesh is obtained based on the pose and refined shape parameters.

We employ end-to-end training for the pipeline, and the loss function consists of three components: shape loss, pose loss, and shape-decompose loss. The CNN backbone is supervised by shape loss and pose loss, while the MLP used in the shape reconstruction part is supervised by shape-decompose loss.

**Shape Loss.** In shape loss, we supervise the predicted part widths predicted by the CNN backbone. Specifically, we require the projection results of the part slice widths and the vertex widths to be close to the target value after data augmentation. \(K\) represents the number of vertices in the human mesh model and \(J\) represents the number of joints.

\[
L_{\text{shape}} = \sum_{j} \left[ \frac{1}{2} \left( \tilde{w}_{j}^{2D} - \bar{w}_{j}^{2D} \right)^{2} + \mu \sum_{k} \left( \frac{1}{2} \left( \tilde{w}_{k}^{2D} - \bar{w}_{k}^{2D} \right)^{2} \right) \right].
\] (9)

**Pose Loss.** Pose loss is defined to supervise the predicted skeleton and twist angle. We adopt the same loss function as HybrIK (Li et al. 2021) and denote it as \(L_{\text{pose}}\).

**Shape-decompose Loss.** Shape-decompose loss ensures that the shape reconstruction module predicts a valid human mesh while best preserving the part slice widths and bone lengths predicted by the CNN backbone. It consists of three loss functions

\[
L_{\text{decomp}} = L_{\text{bone}} + L_{\text{width}} + \mu L_{\text{reg}},
\] (10)

where

\[
L_{\text{bone}} = \sum_{j} \left( \left\| \tilde{x}_{j} - \tilde{x}_{j} \right\|_{1} + \left\| \tilde{l}_{j} - \tilde{l}_{j} \right\|_{1} \right),
\] (11)

\[
L_{\text{width}} = \sum_{j} \left( \left\| \tilde{w}_{j} - \bar{w}_{j} \right\|_{2} + \left\| \tilde{w}_{j} - \bar{w}_{j} \right\|_{2} \right),
\] (12)

\[
L_{\text{reg}} = \left\| \beta \right\|_{2}^{2}.
\] (13)

In the equations, \(\tilde{x}_{j}, \tilde{l}_{j}, \tilde{w}_{j}\) are the keypoint coordinates, the bone length and the part slice widths of part \(j\) refined by the MLP in the shape reconstruction module. \(L_{\text{bone}}\) and \(L_{\text{width}}\) supervise the preservation of the bone length and part slice widths respectively, and \(L_{\text{reg}}\) regularizes \(\beta\) parameter.

**Overall Loss.** The overall loss of our pipeline is formulated as

\[
L = L_{\text{pose}} + \mu L_{\text{decomp}} + \mu L_{\text{shape}}.
\] (14)

### 4 Experiments

#### 4.1 Datasets

We use 3DPW (von Marcard et al. 2018), Human3.6M (Ionescu et al. 2013), COCO (Lin et al. 2014), AGORA (Patel et al. 2021) and Model Agency Dataset (Choutas et al. 2022) for training. The original Model Agency Dataset contains 94,620 images of 4,419
models, but we only use about one-third of these images in our training due to the unavailability of many images on the Internet. To avoid data bias, the images are sampled following previous work (Choutas et al. 2022). We also follow previous work and use synthetic data to assist network training. The rendering settings are identical to (Sengupta, Budvytis, and Cipolla 2021a).

We evaluate our model on SSP-3D (Sengupta, Budvytis, and Cipolla 2020) and HBW datasets (Choutas et al. 2022). The results on SSP-3D dataset show the model’s performance on diverse human body shapes, while the results on HBW dataset indicate the model’s performance on images of people wearing different clothing.

4.2 Comparison with the State-of-the-art

We evaluate the performance of different methods on SSP-3D and HBW test and validation datasets. Following previous work, on SSP-3D dataset, we use PVE-T-SC, a scale-normalized per-vertex error metric to evaluate the model performance. On HBW dataset, we report the predicted height (H), chest (C), waist (W), and hip circumference (HC) errors, and P2P_{20K} errors of different models. All the experiments of our method use part slicing number \( n = 1 \) by default unless otherwise stated. For a fair comparison, we also retrain two best-performing networks (Sengupta, Budvytis, and Cipolla 2021a; Choutas et al. 2022) with the same datasets and settings as our method.

Tab. 1 shows that our method surpasses previous works on SSP-3D dataset, which shows that our method can deal with the diverse human body shape much better than previous methods. Tab. 3 and 2 shows the performance on HBW validation and test dataset. On HBW test dataset, our method achieves comparable results with SOTA methods and predicts more accurate waist and hip circumferences. On HBW validation set, our method outperforms previous SOTA methods. These results prove that our method can deal with diverse human clothing better than previous methods. Qualitative results are provided in Fig. 5.

### 4.3 Ablation Study

To demonstrate the effectiveness of different components in our method, we conduct ablation studies on SSP-3D dataset and HBW validation set.

**Shape reconstruction.** To analyze the effectiveness and robustness of our new human shape parameterization, we reconstruct body shapes using bone lengths and part slice widths with different reconstruction algorithms under different noise ratios. The results are shown in Tab. 4. All the model are trained on shape parameters sampled from Gaussian distributions and tested on 500 different body shapes obtained from AMASS dataset (Mahmood et al. 2019). “Hybrid” algorithm means using the semi-analytical algorithm, “Analytical” algorithm means solely employing the analytical algorithm, and “NN” algorithm means directly using the neural network without analytical steps. From the first three lines in Tab. 4, we observe that our proposed semi-analytical algorithm achieves the lowest error especially when the noise ratio is small. Additionally, when the noise is subtle, the parameterizations using different part slicing number \( (n = 1, 2, 3) \) all achieve an acceptable low error. When the noise ratio is large, the error ratio decreases with larger \( n \). Thus, we can conclude that our semi-analytically method accurately reconstructs human shape, and a larger \( n \) makes

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>PVE-T-SC ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMR (Kanazawa et al. 2018)</td>
<td>SMPL</td>
<td>22.9</td>
</tr>
<tr>
<td>SPIN (Kolotouros et al. 2019)</td>
<td>SMPL</td>
<td>22.2</td>
</tr>
<tr>
<td>SMPL</td>
<td></td>
<td>15.9</td>
</tr>
<tr>
<td>(Sengupta et al. 2020)</td>
<td>SMPL</td>
<td>13.3</td>
</tr>
<tr>
<td>(Sengupta et al. 2021b) †</td>
<td>SMPL</td>
<td>13.6</td>
</tr>
<tr>
<td>(Sengupta et al. 2021a)</td>
<td>SMPL</td>
<td>18.8</td>
</tr>
<tr>
<td>HybIK (Li et al. 2021)</td>
<td>SMPL</td>
<td>22.8</td>
</tr>
<tr>
<td>LVD (Corona et al. 2022)</td>
<td>SMPL</td>
<td>26.1</td>
</tr>
<tr>
<td>CLIFF (Li et al. 2022b)</td>
<td>SMPL</td>
<td>18.4</td>
</tr>
<tr>
<td>SHAPY (Choutas et al. 2022)</td>
<td>SMPL-X</td>
<td>19.2</td>
</tr>
<tr>
<td>SoY (Sarkar et al. 2023)</td>
<td>SMPL</td>
<td>15.8</td>
</tr>
<tr>
<td>(Ma et al. 2023)</td>
<td>SMPL</td>
<td>18.8</td>
</tr>
<tr>
<td>(Sengupta et al. 2021a) †</td>
<td>SMPL</td>
<td>15.4</td>
</tr>
<tr>
<td>SHAPY (Choutas et al. 2022)</td>
<td>SMPL</td>
<td>12.2</td>
</tr>
<tr>
<td>ShapeBoost (Ours)</td>
<td>SMPL</td>
<td>11.4</td>
</tr>
<tr>
<td>ShapeBoost (Ours)</td>
<td>SMPL-X</td>
<td>12.0</td>
</tr>
</tbody>
</table>

Table 1: Quantitative comparisons with state-of-the-art methods on the SSP-3D test set in mm. Symbol † means using multiple images as input, and symbol * means retraining using the same training setting as our method.

<table>
<thead>
<tr>
<th>Method</th>
<th>H</th>
<th>C</th>
<th>W</th>
<th>HC</th>
<th>P2P_{20K}</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPIN</td>
<td>59</td>
<td>92</td>
<td>78</td>
<td>101</td>
<td>29</td>
</tr>
<tr>
<td>Sengupta et al. 2020</td>
<td>135</td>
<td>167</td>
<td>145</td>
<td>102</td>
<td>47</td>
</tr>
<tr>
<td>TUCH</td>
<td>58</td>
<td>89</td>
<td>75</td>
<td>57</td>
<td>26</td>
</tr>
<tr>
<td>Sengupta et al. 2021a</td>
<td>82</td>
<td>133</td>
<td>107</td>
<td>63</td>
<td>32</td>
</tr>
<tr>
<td>CLIFF</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td>SHAPY</td>
<td>51</td>
<td>65</td>
<td>69</td>
<td>57</td>
<td>21</td>
</tr>
</tbody>
</table>

Table 2: Quantitative comparisons with state-of-the-art methods on the HBW test set in mm.

<table>
<thead>
<tr>
<th>Method</th>
<th>H</th>
<th>C</th>
<th>W</th>
<th>HC</th>
<th>P2P_{20K}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sengupta et al. 2021a</td>
<td>68</td>
<td>89</td>
<td>111</td>
<td>71</td>
<td>30</td>
</tr>
<tr>
<td>HybIK</td>
<td>88</td>
<td>82</td>
<td>74</td>
<td>51</td>
<td>33</td>
</tr>
<tr>
<td>LVD #</td>
<td>-</td>
<td>89</td>
<td>131</td>
<td>87</td>
<td>31</td>
</tr>
<tr>
<td>SHAPY</td>
<td>63</td>
<td>59</td>
<td>85</td>
<td>54</td>
<td>25</td>
</tr>
<tr>
<td>Ma et al. 2023</td>
<td>112</td>
<td>87</td>
<td>133</td>
<td>59</td>
<td>41</td>
</tr>
<tr>
<td>(Sengupta et al. 2021a)</td>
<td>72</td>
<td>66</td>
<td>74</td>
<td>49</td>
<td>29</td>
</tr>
<tr>
<td>SHAPY*</td>
<td>62</td>
<td>52</td>
<td>72</td>
<td>50</td>
<td>26</td>
</tr>
<tr>
<td>ShapeBoost (SMPL)</td>
<td>58</td>
<td>54</td>
<td>72</td>
<td>42</td>
<td>25</td>
</tr>
<tr>
<td>ShapeBoost (SMPL-X)</td>
<td>61</td>
<td>49</td>
<td>71</td>
<td>49</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 3: Quantitative comparisons with state-of-the-art methods on the HBW validation set in mm. Symbol # means using ground truth scale and symbol * means retraining using the same training setting as our method.
Table 4: Ablation experiments of reconstructing shape using our new shape parameterization in mm.

<table>
<thead>
<tr>
<th>n</th>
<th>Algo.</th>
<th>0% noise</th>
<th>1% noise</th>
<th>2% noise</th>
<th>5% noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hybrid</td>
<td>0.69</td>
<td>2.30</td>
<td>5.95</td>
<td>8.83</td>
</tr>
<tr>
<td>1</td>
<td>Analy.</td>
<td>6.14</td>
<td>6.59</td>
<td>8.99</td>
<td>12.34</td>
</tr>
<tr>
<td>1</td>
<td>NN</td>
<td>1.82</td>
<td>2.99</td>
<td>6.20</td>
<td>8.98</td>
</tr>
<tr>
<td>2</td>
<td>Hybrid</td>
<td>0.58</td>
<td>2.01</td>
<td>5.40</td>
<td>8.21</td>
</tr>
<tr>
<td>3</td>
<td>Hybrid</td>
<td>0.65</td>
<td>1.93</td>
<td>5.00</td>
<td>7.63</td>
</tr>
</tbody>
</table>

Table 5: Ablation experiments of shape estimation from RGB images using different shape parameterization on SSP-3D and HBW validation set in mm.

<table>
<thead>
<tr>
<th>Method</th>
<th>PVE-T-SC</th>
<th>P2P20K</th>
</tr>
</thead>
<tbody>
<tr>
<td>β</td>
<td>12.3</td>
<td>26.0</td>
</tr>
<tr>
<td>n = 1</td>
<td>11.4</td>
<td>25.1</td>
</tr>
<tr>
<td>n = 2</td>
<td>11.6</td>
<td>26.2</td>
</tr>
</tbody>
</table>

Table 6: Ablation experiments of data augmentation module on SSP-3D and HBW validation set in mm.

<table>
<thead>
<tr>
<th>Method</th>
<th>PVE-T-SC</th>
<th>P2P20K</th>
</tr>
</thead>
<tbody>
<tr>
<td>ShapeBoost (Ours)</td>
<td>11.4</td>
<td>25.1</td>
</tr>
<tr>
<td>w/o Augment</td>
<td>12.1</td>
<td>26.5</td>
</tr>
<tr>
<td>w/o Augment, w/o Decompose</td>
<td>12.4</td>
<td>27.0</td>
</tr>
</tbody>
</table>

5 Conclusion

In this paper, we present ShapeBoost, a new framework for accurate human shape recovery that outperforms the current state-of-the-art methods. This framework exploits a new human shape parameterization that decomposes human shape into bone lengths and the mean width of each part slice. Compared to the existing representation with PCA coefficients, our new method is more flexible and interpretable. Based on the new shape parameterization, a new clothing-preserving data augmentation module is proposed to generate realistic images of various human shapes and the corresponding accurate annotations. Our method randomly augments the body shape without destructing the clothing details. Experiments show that our method achieves SOTA performance for extreme body shapes as well as achieves high accuracy for people under different types of clothing.
Acknowledgments

Cewu Lu is the corresponding author. He is the member of Qing Yuan Research Institute, Qi Zhi Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China.

This work was supported by the National Key R&D Program of China (No. 2021ZD0110704), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, and Shanghai Science and Technology Commission (21511101200).

References


Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. TPAMI.


Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In ICCV.


