ShapeBoost: Boosting Human Shape Estimation with Part-Based Parameterization and Clothing-Preserving Augmentation

Siyuan Bian[1], Jiefeng Li[1], Jiasheng Tang[3,4], Cewu Lu[1,2]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
[3]DAMO Academy, Alibaba group, Hangzhou, China
[4]Hupan Lab, Hangzhou, China
{biansiyuan, ljf_likit, lucewu}@sjtu.edu.cn, jiasheng.tjs@alibaba-inc.com

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

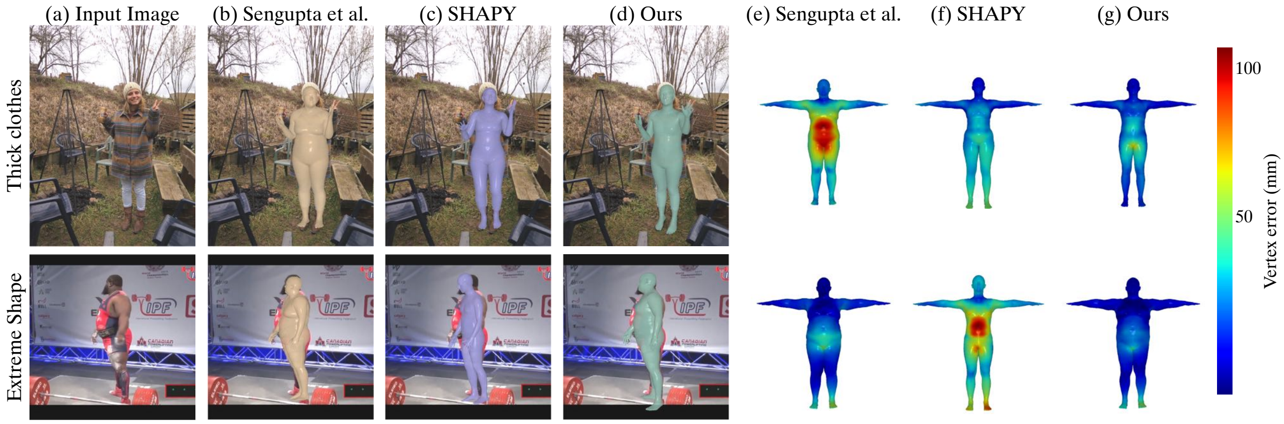

Figure 1: Previous SOTA methods for human shape estimation (Sengupta, Budvytis, and Cipolla 2021a; Choutas et al. 2022) (b, c) either fail on images of people wearing thick clothes or fail on images of people with extreme body shapes, while our method (d) achieves pixel-aligned results with high accuracy in both situations. Warmer colors on the human mesh represent higher per-vertex error.

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 a 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 medthods 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 commmonly-used 3D human datasets (von Marcard et al. 2018; Ionescu et al. 2013) con-

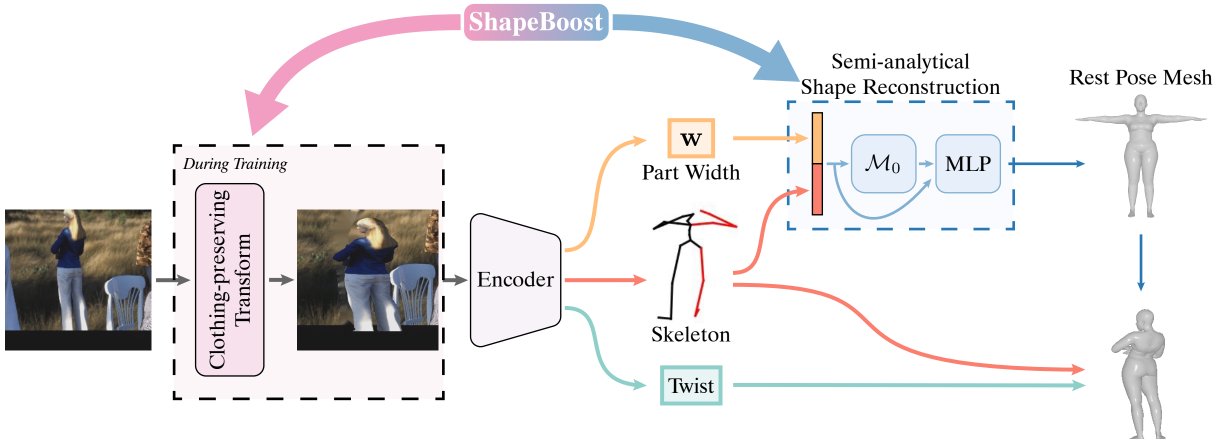

Figure 2: The overall pipeline. First, the input image is randomly transformed with the clothing-preserving image transformation, and a convolutional neural network (CNN) is employed to extract skeleton, part widths and twist rotations. Then, the pose is obtained using inverse kinematics and the shape 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.

tain limited number of people. To overcome this limitation, some researchers have created synthetic image datasets by rendering the mesh generated by parameterized human models (Hoffmann et al. 2019; Sengupta, Budvytis, and Cipolla 2020; Varol et al. 2017; Weitz et al. 2021). However, it is difficult to obtain images with natural clothing and realistic scenes using the naive rendering. Recently, more realistic synthetic datasets (Bertiche, Madadi, and Escalera 2020; Pumarola et al. 2019; Liang and Lin 2019; Patel et al. 2021; Black et al. 2023) have been proposed, which contain people in different clothing with the help of human scans, simulation or deep generative networks. Choutas et al. (Choutas et al. 2022) have proposed the Model-Agency dataset, which uses images from model agency websites labeled with linguistic attributes and measurements. Although these new datasets contain more diverse body shapes, most datasets still lack people with extreme body shapes, and the authenticity of synthetic images remains insufficient.

Estimation Framework. Several methods (Sengupta, Budvytis, and Cipolla 2020, 2021b,a) train the network directly on synthetic data. To reduce the domain gap, they use proxy representations (PRs) as input, such as part segmentation masks (Varol et al. 2017), silhouettes (Sengupta, Budvytis, and Cipolla 2020; Ruiz et al. 2022), Canny edge detection results (Sengupta, Budvytis, and Cipolla 2021b,a) or 2D keypoint heatmaps (Sengupta, Budvytis, and Cipolla 2020, 2021b,a). Other work (Dwivedi et al. 2021; Omran et al. 2018; Agarwal and Triggs 2005) uses real-world data for training and exploits 2D shape cues as supervision. Body-part segmentation masks (Dwivedi et al. 2021; Omran et al. 2018) and silhouettes (Agarwal and Triggs 2005) are widely used among them. LVD (Corona et al. 2022) learns the vertex descent direction based on image-aligned features, and SHAPY (Choutas et al. 2022) uses linguistic attributes and body measurements as supervision.

Unlike previous work, our method generates images with diverse human body shapes without altering clothing, lighting, and background details. Therefore, the diversity is rich and the domain gap is small. Since our framework utilizes our new parameterization, there is no ambiguity even when the human is in thick clothing and our method will not enlarge error even when the pose estimation is inaccurate.

3 Method

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.

3.1 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 $\mathcal{V}(\boldsymbol{\theta}, \boldsymbol{\beta})$ that maps pose $\boldsymbol{\theta} \in \mathbb{R}^{3J}$ and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ to a human mesh $\mathbf{V}$, where J is the number of joints. The pose parameters $\boldsymbol{\theta}$ represent the relative rotation of body joints, and the shape parameters $\boldsymbol{\beta}$ are coefficients of a PCA body shape basis. SMPL model is drived in two steps:

$$\mathbf{T} = \mathcal{S}(\boldsymbol{\beta}), \tag{1}$$

$$\mathbf{V} = \mathcal{V}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \mathcal{P}(\boldsymbol{\theta}, \mathcal{S}(\boldsymbol{\beta})). \tag{2}$$

First, a rest-pose mesh $\mathbf{T}$ is constructed using function $\mathcal{S}$. Second, the rest-pose mesh is driven to the target pose by function $\mathcal{P}$. The shape of the mesh is determined only by $\boldsymbol{\beta}$, and the posing procedure does not change the body shape. Most current methods regress shape parameters $\boldsymbol{\beta}$ 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.

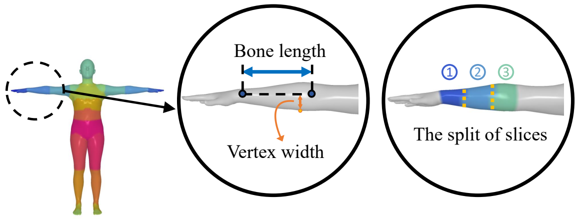

Figure 3: Illustration of the shape decomposition procedure. From left to right, the figure shows the part segmentation, the definition of bone length and vertex width, and the slicing of one body part.

3.2 Part-based Parameterization

In this work, we propose a novel parameterization of human shape using bone lengths and widths of part slices. Compared to the $\boldsymbol{\beta}$ 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 vertex 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:

$$\mathbf{T} = \mathcal{M}(\mathbf{l}, \mathbf{w}), \tag{3}$$

$$\mathbf{V} = \mathcal{P}(\boldsymbol{\theta}, \mathcal{M}(\mathbf{l}, \mathbf{w})), \tag{4}$$

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

Deriving the function $\mathcal{M}$ 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 $\mathcal{M}_0$.

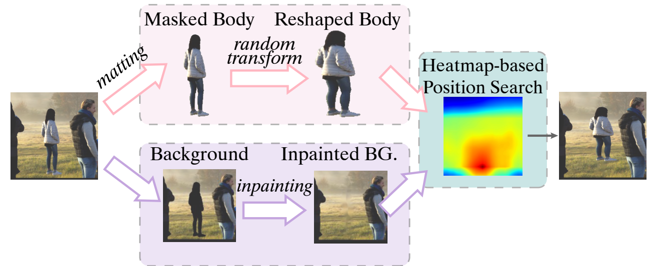

Figure 4: Illustration of clothing-preserving transformation.

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 $\mathcal{M}$ can be written as

$$\mathbf{T} = \mathcal{M}(\mathbf{l}, \mathbf{w}) = \mathrm{MLP}(\mathcal{M}_0(\mathbf{l}, \mathbf{w}), \mathbf{l}, \mathbf{w}, \boldsymbol{\Delta}\mathbf{l}, \boldsymbol{\Delta}\mathbf{w}), \tag{5}$$

where $\boldsymbol{\Delta}\mathbf{l}$ and $\boldsymbol{\Delta}\mathbf{w}$ are the difference between the target bone lengths and part slice widths and the corresponding values obtained by $\mathcal{M}_0$. 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

$$T = SR = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{6}$$

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:

$$\bar{w}_k^{2D} = \frac{ab \cdot l_j^{2D}}{\bar{l}_j^{2D}} w_k^{2D}, \tag{7}$$

where l_j^{2D} and $\bar{l}_j^{2D}$ 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

$$\bar{w}_j = \frac{\bar{s}}{s} \times \frac{ab \cdot l_j^{2D}}{\bar{l}_j^{2D}} \times w_j. \tag{8}$$

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 $\bar{s}$ and $\bar{w}_j$ are the corresponding values after the transformation. Due to scale ambiguity, $\bar{s}$ 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 $\bar{s}$ 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 estimate 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 module 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.

$$L_{shape} = \sum_{j}^{J} \|\hat{w}_j^{2D} - \bar{w}_j^{2D}\|_2^2 \; + \; \mu_0 \sum_{k}^{K} \|\hat{w}_k^{2D} - \bar{w}_k^{2D}\|_2^2. \tag{9}$$

Pose Loss. Pose loss is designed 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_{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

$$L_{decomp} = L_{bone} + L_{width} + \mu_1 L_{reg}, \tag{10}$$

where

$$L_{bone} = \sum_{j}^{J} \left(\|\tilde{\mathbf{x}}_j - \hat{\mathbf{x}}_j\|_1 \; + \; \|\tilde{l}_j - \hat{l}_j\|_1 \right), \tag{11}$$

$$L_{width} = \sum_{j}^{J} \left(\|\tilde{w}_j - \hat{w}_j\|_2^2 \; + \; \|\frac{\tilde{w}_j}{\tilde{l}_j} - \frac{\hat{w}_j}{\hat{l}_j}\|_2^2 \right), \tag{12}$$

$$L_{reg} = \|\tilde{\boldsymbol{\beta}}\|_2^2. \tag{13}$$

In the equations, $\tilde{\mathbf{x}}_j$, $\tilde{l}_j$, $\tilde{w}_j$ 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_{bone} and L_{width} supervise the preservation of the bone length and part slice widths respectively, and L_{reg} regularizes $\tilde{\boldsymbol{\beta}}$ parameter.

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

$$L = L_{pose} + \mu_2 L_{decomp} + \mu_3 L_{shape}. \tag{14}$$

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 $4,419$

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

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.

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 $\mathrm{P2P}_{20\mathrm{K}}$ 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 previous 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.

Method	H	C	W	HC	$\mathrm{P2P}_{20\mathrm{K}}$
SPIN	59	92	78	101	29
Sengupta et al. 2020	135	167	145	102	47
TUCH	58	89	75	57	26
Sengupta et al. 2021a	82	133	107	63	32
CLIFF	-	-	-	-	27
SHAPY	**51**	65	69	57	**21**
ShapeBoost (SMPL)	66	**63**	58	**47**	25
ShapeBoost (SMPL-X)	68	69	**56**	49	22

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

Method	H	C	W	HC	$\mathrm{P2P}_{20\mathrm{K}}$
Sengupta et al. 2021a	68	89	111	71	30
HybrIK	88	82	74	51	33
LVD #	-	89	131	87	31
SHAPY	63	59	85	54	25
Ma et al. 2023	112	87	133	59	41
Sengupta et al. 2021a*	72	66	74	49	29
SHAPY*	62	52	72	50	26
ShapeBoost (SMPL)	**58**	54	72	**42**	25
ShapeBoost (SMPL-X)	61	**49**	**71**	49	**23**

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.

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

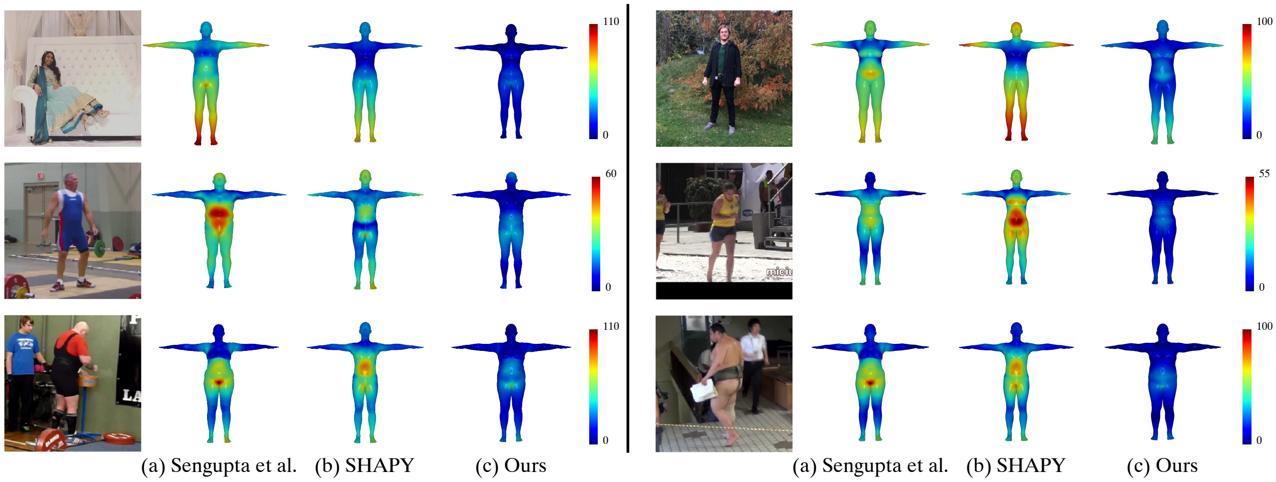

Figure 5: Qualitative results on SSP-3D and HBW datasets. From left to right: Input image, (a) Sengupta et al. (Sengupta, Budvytis, and Cipolla 2021a) results, (b) SHAPY (Choutas et al. 2022) results, and (c) Our results. Warmer colors mean higher per-vertex error. Experiments on SSP-3D dataset use PVE-T-SC metric, and experiments on HBW dataset use $P2P_{20K}$ metric.

		V2V Error (mm) ↓			
n	**Algo.**	**0%noise**	**1%noise**	**2%noise**	**5%noise**
1	Hybrid	0.69	2.30	5.95	8.83
1	Analy.	6.14	6.59	8.99	12.34
1	NN	1.82	2.99	6.20	8.98
2	Hybrid	**0.58**	2.01	5.40	8.21
3	Hybrid	0.65	**1.93**	**5.00**	**7.63**

Table 4: Ablation experiments of reconstructing shape using our new shape parameterization in mm.

it more robust to noise.

Shape estimation from images. We also experiment using different parameterizations for estimating human body shapes from RGB images. Tab. 5 provides a comparison of the results obtained using the direct shape parameterization ($\boldsymbol{\beta}$) (Li et al. 2021) with our novel parameterization utilizing $n = 1$ and $n = 2$. We use image augmentation in the training. Since it is hard to find a ground truth $\boldsymbol{\beta}$ for augmented images, we use the 2D coordinates of vertices as supervision. We find that using our new parameterization yields better results, but a larger n does not improve performance. The reasons are (1) the parameterization with $n = 1$ already achieves a small shape reconstruction error (2) using larger n complicates the regression task for the CNN backbone, resulting in a reduction in the accuracy of predicting part slicing widths.

The effectiveness of data augmentation. We also make ablation studies with different training data quantitatively. The results are shown in Tab. 6. When the data augmentation module is not used, the performance of our model drops on both HBW and SSP-3D dataset. This shows the effectiveness of our data augmentation module.

Method	PVE-T-SC	$P2P_{20K}$
β	12.3	26.0
$n = 1$	**11.4**	**25.1**
$n = 2$	11.6	26.2

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

Method	PVE-T-SC	$P2P_{20K}$
ShapeBoost (Ours)	**11.4**	**25.1**
w/o Augment	12.1	26.5
w/o Augment, w/o Decompose	12.4	27.0

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

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

Agarwal, A.; and Triggs, B. 2005. Recovering 3D human pose from monocular images. *TPAMI*, 28(1): 44–58.

Bertiche, H.; Madadi, M.; and Escalera, S. 2020. CLOTH3D: clothed 3d humans. In *ECCV*, 344–359. Springer.

Black, M. J.; Patel, P.; Tesch, J.; and Yang, J. 2023. BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8726–8737.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*.

Choi, H.; Moon, G.; and Lee, K. M. 2020. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, 769–787. Springer.

Choutas, V.; Müller, L.; Huang, C.-H. P.; Tang, S.; Tzionas, D.; and Black, M. J. 2022. Accurate 3D body shape regression using metric and semantic attributes. In *CVPR*, 2718–2728.

Corona, E.; Pons-Moll, G.; Alenyà, G.; and Moreno-Noguer, F. 2022. Learned Vertex Descent: A New Direction for 3D Human Model Fitting. In *ECCV*.

Dwivedi, S. K.; Athanasiou, N.; Kocabas, M.; and Black, M. J. 2021. Learning to regress bodies from images using differentiable semantic rendering. In *ICCV*, 11250–11259.

Fang, H.-S.; Sun, J.; Wang, R.; Gou, M.; Li, Y.-L.; and Lu, C. 2019. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *ICCV*, 682–691.

Guan, P.; Weiss, A.; Balan, A. O.; and Black, M. J. 2009. Estimating human shape and pose from a single image. In *ICCV*, 1381–1388. IEEE.

Hoffmann, D. T.; Tzionas, D.; Black, M. J.; and Tang, S. 2019. Learning to train with synthetic humans. In *Pattern Recognition*, 609–623. Springer.

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*.

Joo, H.; Neverova, N.; and Vedaldi, A. 2021. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *3DV*.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *CVPR*.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. VIBE: Video inference for human body pose and shape estimation. In *CVPR*.

Kocabas, M.; Huang, C.-H. P.; Hilliges, O.; and Black, M. J. 2021. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 11127–11137.

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*.

Kolotouros, N.; Pavlakos, G.; and Daniilidis, K. 2019. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 4501–4510.

Li, J.; Bian, S.; Liu, Q.; Tang, J.; Wang, F.; and Lu, C. 2023a. NIKI: Neural Inverse Kinematics with Invertible Neural Networks for 3D Human Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12933–12942.

Li, J.; Bian, S.; Xu, C.; Chen, Z.; Yang, L.; and Lu, C. 2023b. HybrIK-X: Hybrid Analytical-Neural Inverse Kinematics for Whole-body Mesh Recovery. *arXiv preprint arXiv:2304.05690*.

Li, J.; Bian, S.; Xu, C.; Liu, G.; Yu, G.; and Lu, C. 2022a. D&D: Learning Human Dynamics from Dynamic Camera. In *ECCV*.

Li, J.; Xu, C.; Chen, Z.; Bian, S.; Yang, L.; and Lu, C. 2021. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 3383–3393.

Li, Z.; Liu, J.; Zhang, Z.; Xu, S.; and Yan, Y. 2022b. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 590–606. Springer.

Liang, J.; and Lin, M. C. 2019. Shape-aware human pose and shape reconstruction using multi-view images. In *ICCV*, 4352–4362.

Lin, K.; Wang, L.; and Liu, Z. 2021a. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 1954–1963.

Lin, K.; Wang, L.; and Liu, Z. 2021b. Mesh graphormer. In *ICCV*, 12939–12948.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *TOG*.

Ma, X.; Su, J.; Wang, C.; Zhu, W.; and Wang, Y. 2023. 3D Human Mesh Estimation from Virtual Markers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 534–543.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 5442–5451.

Moon, G.; and Lee, K. M. 2020. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 752–768. Springer.

Muller, L.; Osman, A. A.; Tang, S.; Huang, C.-H. P.; and Black, M. J. 2021. On self-contact and human pose. In *CVPR*, 9990–9999.

Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; and Schiele, B. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 484–494. IEEE.

Patel, P.; Huang, C.-H. P.; Tesch, J.; Hoffmann, D. T.; Tripathi, S.; and Black, M. J. 2021. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, 13468–13478.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*.

Pavlakos, G.; Zhu, L.; Zhou, X.; and Daniilidis, K. 2018. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 459–468.

Pumarola, A.; Sanchez-Riera, J.; Choi, G.; Sanfeliu, A.; and Moreno-Noguer, F. 2019. 3dpeople: Modeling the geometry of dressed humans. In *ICCV*, 2242–2251.

Ruiz, N.; Bellver, M.; Bolkart, T.; Arora, A.; Lin, M. C.; Romero, J.; and Bala, R. 2022. Human body measurement estimation with adversarial augmentation. In *2022 International Conference on 3D Vision (3DV)*, 219–230. IEEE.

Sarkar, R.; Dave, A.; Medioni, G.; and Biggs, B. 2023. Shape of You: Precise 3D shape estimations for diverse body types. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3519–3523.

Sengupta, A.; Budvytis, I.; and Cipolla, R. 2020. Synthetic Training for Accurate 3D Human Pose and Shape Estimation in the Wild. In *British Machine Vision Conference (BMVC)*.

Sengupta, A.; Budvytis, I.; and Cipolla, R. 2021a. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In *ICCV*, 11219–11229.

Sengupta, A.; Budvytis, I.; and Cipolla, R. 2021b. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *CVPR*, 16094–16104.

Tung, H.-Y.; Tung, H.-W.; Yumer, E.; and Fragkiadaki, K. 2017. Self-supervised learning of motion capture. *NeurIPS*, 30.

Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; and Schmid, C. 2018. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*, 20–36.

Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M. J.; Laptev, I.; and Schmid, C. 2017. Learning from synthetic humans. In *CVPR*, 109–117.

von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*.

Weitz, A.; Colucci, L.; Primas, S.; and Bent, B. 2021. InfiniteForm: A synthetic, minimal bias dataset for fitness applications. *arXiv preprint arXiv:2110.01330*.

Zhang, H.; Tian, Y.; Zhang, Y.; Li, M.; An, L.; Sun, Z.; and Liu, Y. 2022. PyMAF-X: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*.