Shape of You: Precise 3D shape estimations for diverse body types

Rohan Sarkar[1,2], Achal Dave[2], Gerard Medioni[2], and Benjamin Biggs[2]

[1]Purdue University, [2]Amazon

Abstract

This paper presents Shape of You (SoY), an approach to improve the accuracy of 3D body shape estimation for vision-based clothing recommendation systems. While existing methods have successfully estimated 3D poses, there remains a lack of work in precise shape *estimation, particularly for diverse human bodies. To address this gap, we propose two loss functions that can be readily integrated into parametric 3D human reconstruction pipelines. Additionally, we propose a test-time optimization routine that further improves quality. Our method improves over the recent SHAPY [7] method by 17.7% on the challenging SSP-3D dataset [16]. We consider our work to be a step towards a more accurate 3D shape estimation system that works reliably on diverse body types and holds promise for practical applications in the fashion industry.*

1. Introduction

Clothing retailers have been designing interactive experiences that make clothing recommendations based on customer selfie photos, which contain valuable visual cues pertaining to body shape. Approaches promise to transform digital shopping experiences by empowering customers to shop more confidently without physical access to the inventory and by reducing the rate of clothing returns. However, estimating shape characteristics from 2D images of diverse human bodies remains a challenging problem.

Techniques for image-based 3D human reconstruction present compelling opportunities in this space. These approaches disentangle the space of human bodies into *shape* deformations that control relative body proportions, and *pose* deformations that control the position and orientation of limbs. Precisely estimating a customer's *shape* enables valuable fashion applications, such as being able to predict suitable clothing sizes or identify flattering garments. Pose deformations are typically nuisance factors, but they must still be modeled carefully to avoid noisy shape estimates.

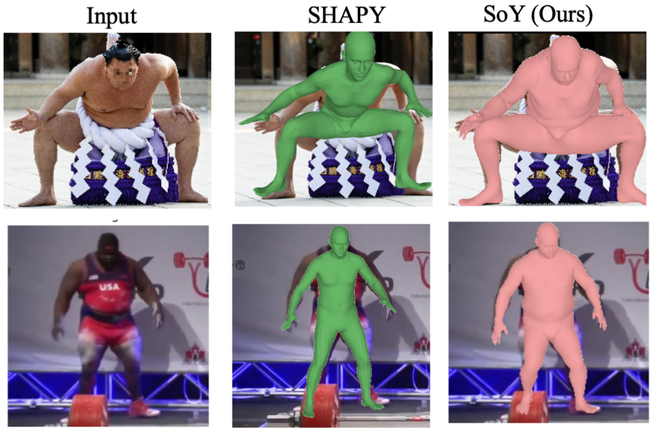

Figure 1. Our SoY method generates improved body shape estimates compared to prior work [7].

Despite significant recent progress for image-based 3D human reconstruction, most improvements have been made in robust 3D pose estimation. Precisely estimating *shape* characteristics, particularly for plus-size body types, remains a challenging problem, in part due to the scarcity of training data. Figure 1 compares our proposed approach to that of Choutas et al. [7]. We achieve improved shape estimates despite training without annotated body measurements or semantic textual attributes that are used by their method. Next, we summarize our paper's key contributions:

1. We propose SoY, a novel 3D human reconstruction method that incorporates specific loss functions to promote detailed 3D shape recovery, particularly for diverse humans shapes. Our method includes (a) a 2D loss based on dense correspondences [8] between the 3D mesh and foreground person – an improved shape-specific signal over 2D keypoints [4, 10] or 2D silhouettes [20] – and (b) a 3D loss that promotes shape-specific vertex alignment by subtracting pose effects.

2. We propose a refinement step which further improves the quality of shape recovery. This procedure is particularly useful when testing on humans with body shapes that are under-represented in the training distribution.

3. We show that SoY with refinement achieves an improvement of 17.7% over Choutas et al. [7] on SSP-3D without *any body measurements or semantic attributes for training*. Additionally, we show our method performs on-par when the refinement step is omitted.

2. Related Work

The problem of recovering 3D shape and pose for articulated subjects has received significant attention in the literature. One common approach is to fit 3D morphable body models (3DMM) to the subjects of interest. Early methods relied on iterative optimization algorithms [4] to align 3DMMs with 2D observations, while recent works [9, 11] estimate 3DMM parameters with direct learning-based regression. Our work builds on the hybrid approaches [10] which integrate an iterative optimization loop for training the deep network that performs the regression. All of these methods fall short in accurately estimating shape for plus-size individuals. In this paper, we propose loss functions and a refinement procedure to overcome this limitation.

Another emerging trend is to reconstruct 3D articulated subjects without an explicit 3D template prior [1, 6, 14, 15]. While precise shapes can be estimated, they rely heavily on paired 3D training data which is limited for diverse body shapes and tend to entangle clothing in the reconstructed output. Furthermore, it is unclear how best to integrate these neural network-encoded shape representations into downstream fashion applications.

Recent animal reconstruction literature has made progress for complex shape categories [2, 3, 13, 20]. Our paper takes particular inspiration from BANMo [19] who fit to dense correspondences tracked within a video sequence.

However, most relevant are techniques that reconstruct diverse human shapes. Some methods [16, 17] augment existing 3D training datasets with synthetically-generated diverse body shapes. However, this is achieved using in-the-wild datasets [12, 18] of 3D joint angles which is more than we require here. SHAPY [7] proposes a method to recover 3D shapes for diverse fashion models. However, they train using annotated measurements and semantic textual attributes – a requirement we overcome in this work.

3. Preliminaries

Before introducing our method, we cover the necessary background starting with SMPL.

SMPL: SMPL is a 3D model of the human body parameterized by (23×3) axis-angle rotations $\theta \in \mathbb{R}^P$ of the limbs, shape coefficients $\beta \in \mathbb{R}^B$ that control body proportions, and a global rotation parameter $\gamma \in \mathbb{R}^3$. SMPL is supplied with a linear blend skinning function $S : (\theta, \beta, \gamma) \mapsto V$, which generates a set of vertex positions $V \in \mathbb{R}^{6890 \times 3}$ of a 3D mesh.

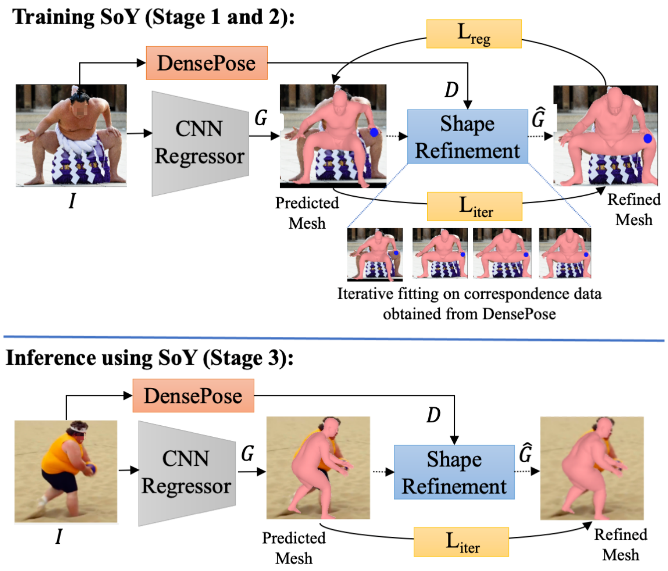

Figure 2. System Overview. **Stage 1**: A CNN Regressor is trained to estimate the SMPL parameters G using the L_{reg} loss (ref. Eq. (3)). **Stage 2**: The Shape Refinement module is initialized with G and runs an iterative optimization process with loss L_{iter} (ref. Eq. (4)) based on dense 3D correspondences D to produce updated parameters $\hat{G}$, used as pseudo-ground truth in the next epoch. **Stage 3**: At inference time, the trained regressor's prediction G is refined using L_{iter} to produce the final fit.

Predicting SMPL parameters from a single monocular image: For a person in input image I, the 3D reconstruction task is to estimate their SMPL parameters (θ, β, γ). Modern algorithms [9, 10] train a deep network $G(I)$ that directly estimates these parameters as well as the translation $t \in \mathbb{R}^3$ of the perspective camera with fixed parameters. Methods are trained on a mix of datasets with various levels of annotation. Vertex losses can be applied when dense 3D scans are available. Otherwise, a fixed linear regressor $J : V \mapsto X$ translates SMPL mesh vertices to 3D joints enabling comparison with sparse 3D annotations. Furthermore, 2D joint losses can be formed using a function $\pi_t(X)$ that projects 3D joints to the image plane.

4. Proposed method

We begin with a function that implements $G(I)$ as described above. As shown in Fig. 2, we start with the training protocol of SPIN [10] which contains the following steps:

- **Stage 1**. Train a feed-forward neural network [9] $G(I)$ to regress SMPL parameters G using a loss L_{reg}.
- **Stage 2**. During training, refine the network's predictions G with an energy-minimization framework (based on [4]) with loss L_{iter}. Use the refined predictions $\hat{G}$ as pseudo ground-truth for the feed-forward network in the next epoch.

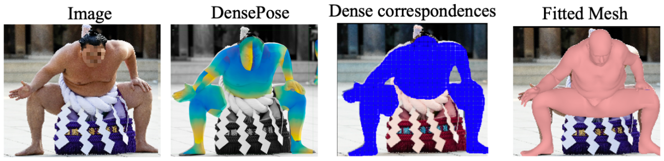

Figure 3. This figure shows the dense image-to-mesh correspondences derived from DensePose. The mesh fitted to these dense correspondences using L_{iter} (ref. Eq. (4)) is shown on the right.

However, as shown later in Table 1, SPIN produces poor shape estimates for plus-sized bodies. To ameliorate this, our SoY method extends SPIN introducing two additional loss functions L_{dp} and L_{tpose} which we add to L_{reg} and L_{iter}. Further, we propose an additional test-time stage:

- **Stage 3**. At test-time, run another pass of energy minimization with loss L_{iter} to refine the test-time predictions.

Note that the energy minimization *does not use any ground-truth data*; the refinement process is based on correspondence data that is predicted from the image. The next section formally introduces our proposed losses.

4.1. Shape-specific losses

In this section, we propose two loss functions that promote strong shape alignment during training.

Dense correspondences: Our first loss function promotes alignment between a set of per-pixel mesh-to-image correspondences. As shown in Fig. 3, we use the popular DensePose [8] method to generate a map $D_G \in \mathbb{R}^{HW \times 3}$ which encodes a 3D point on the posed mesh surface for every pixel x in an image I of size $H \times W$.

$$L_{\text{dp}}(G; D) = \sum_{x}^{H \times W} ||\pi_t(D_G(x)) - x||^2 \tag{1}$$

Pose-invariant vertex alignment: Our second loss function promotes alignment between predicted and refined mesh vertices. Note that we experimented with increasing the weight for SPIN's existing loss between shape coefficients $\beta, \hat{\beta}$ but found this leads to poor results. Instead, we design a loss that penalizes variations between mesh vertices v when *generated in T-Pose* – that is, with any pose effects removed. Here, $\theta_{\text{T}} = \gamma_{\text{T}} = 0$ are the pose and global rotation parameters for the model in T-Pose.

$$L_{\text{tpose}}(G; \hat{G}) = \sum_{v \in V} ||S_v(\theta_{\text{T}}, \beta, \gamma_{\text{T}}) - S_v(\theta_{\text{T}}, \hat{\beta}, \gamma_{\text{T}})||^2 \tag{2}$$

Total losses: The feed-forward regression process is therefore supervised by the following losses where each has a constant weight. $\hat{G}$ are refined SMPL parameters and $\hat{Y}$ are 2D joints predicted with OpenPose [5]:

$$\begin{aligned} L_{\text{reg}}(G;\hat{G}, \hat{Y}) = L_{\text{mesh}}(G; \hat{G}) + L_{\text{3D}}(G; \hat{G}) \\ + L_{\text{2D}}(\pi_t(X); \hat{Y}) + L_{\text{tpose}}(G; \hat{G}) \end{aligned} \tag{3}$$

The iterative optimization process is supervised by the following losses, where $\mu_\beta, \Sigma_\beta, \theta_\mu, \Sigma_\theta$ are mean vectors and covariance matrices for shape and pose parameters respectively. These are provided as part of the SMPL model.

$$L_{\text{iter}}(G; D, \mu, \Sigma) = L_{\text{dp}}(G; D) + L_{\text{prior}}(G; \mu, \Sigma) \tag{4}$$

The losses $L_{\text{mesh}}, L_{\text{3D}}, L_{\text{2D}}$ and L_{prior} were introduced in SPIN and we refer the reader to [10] for details.

4.2. Implementation details

We initialize the feed-forward regressor with weights provided by SPIN [10] and set the loss weights as follows.
Stage 1: $\lambda_{mesh} = 0.1, \lambda_{3D} = 1.0, \lambda_{2D} = 1.0, \lambda_{tpose} = 0.1$. We use the ADAM optimizer with LR $= 5e^{-5}$.
Stage 2: We fix γ, t from Stage 1, and optimize for β, θ with L_{iter} for 250 iterations per epoch. We set $\lambda_{dp} = 99.9, \lambda_{prior,\theta} = 1.0$ and $\lambda_{prior,\beta} = 5.0$.
Stage 3. For the SSP3D dataset [16], the optimal weights for refinement are found to be $\lambda_{prior,\theta} = \lambda_{prior,\beta} = 25.0$. Refinement takes approximately 8s per image.

5. Experiments

This section describes a quantitative and qualitative evaluation against competitive baselines.

5.1. Baselines, datasets and evaluation protocol

We evaluate our approach against three state-of-the-art techniques: HMR [9], SPIN [10] and SHAPY [7]. We do not compare to methods [16, 17] that require in-the-wild datasets [12, 18] of 3D joint angles for training.

We pretrain on the SPIN [10] training datasets and finetune on the Model Agency dataset [7] with generated OpenPose [5] joints and DensePose maps [8]. We do not use body measurement data or linguistic body shape annotations.

Our quantitative evaluation is based on the SSP-3D dataset [16] which contains 311 images of sportspeople with diverse shapes in tight-fitting clothes. Following SHAPY [7], we report per-vertex error in millimeters in T-pose after scale correction (PVE-T-SC) and mean intersection-over-union (mIoU) between predicted and ground-truth 2D silhouettes. We also provide qualitative results using a set of diverse bodies we sourced online.

5.2. Results

Baselines. Table 1 compares SoY against the baselines. Our approach outperforms all three baselines and does not

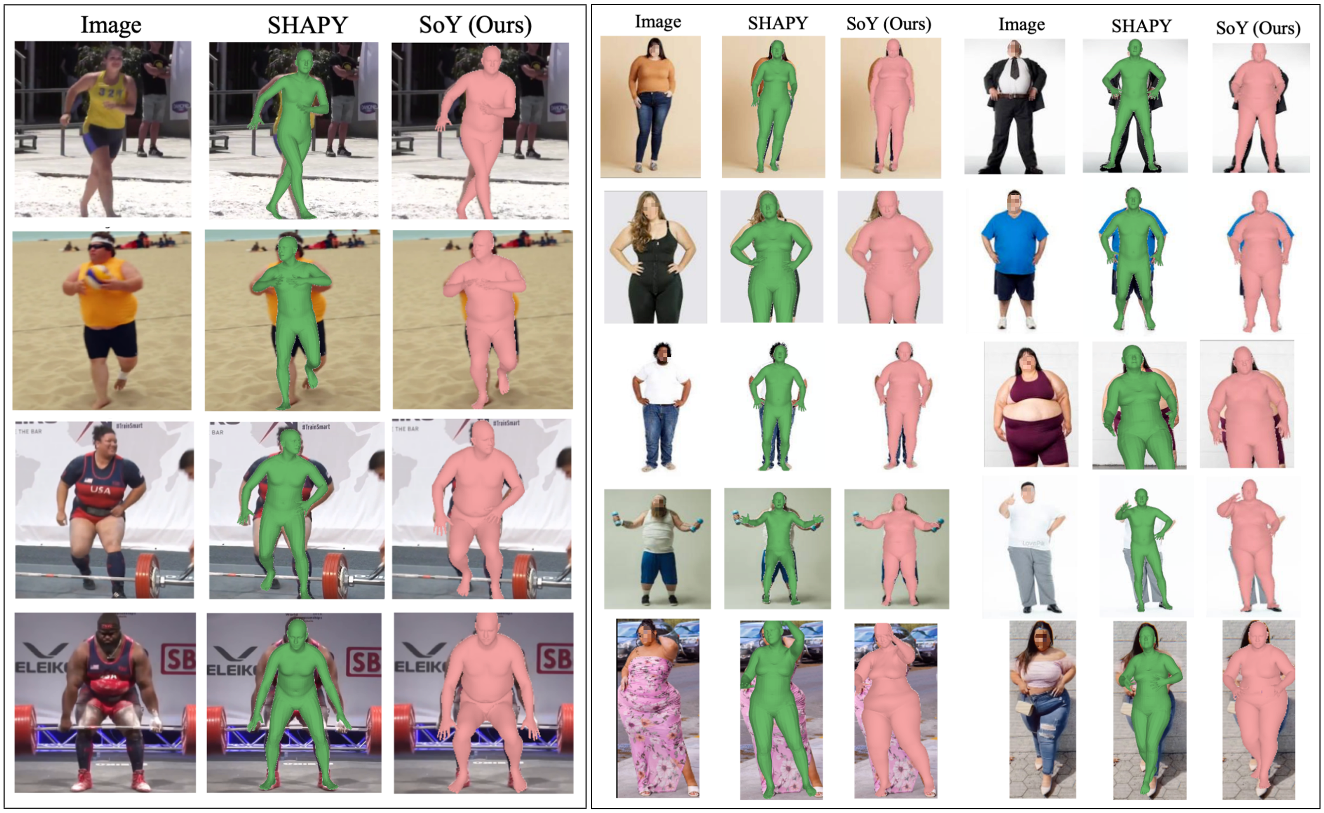

Figure 4. Comparison of results generated using SHAPY and SoY on (left) diverse body types of sports athletes from SSP3D and (right) high BMI body types in a variety of different clothing.

Method	PVE-T-SC (↓)	mIoU (↑)
HMR	22.9	0.69
SPIN	22.2	0.70
SHAPY	19.2	0.71
SoY (Ours)	**15.8**	**0.76**

Table 1. Comparison of scaled mean vertex-to-vertex error in T-pose (PVE-T-SC) and mean intersection-over-union (mIoU) for our SoY against competitive baselines.

L_{tpose}	Stage 3	PVE-T-SC (↓)	mIoU (↑)
✓	✓	**15.8**	**0.76**
✓	✗	19.1	0.65
✗	✗	19.6	0.64

Table 2. Ablation Study on SSP-3D test set. We compare our full method (R1) with a version of our method trained without the Stage 3 refinement step (R2) and without the L_{tpose} loss (R3).

require the semantic linguistic labels or body measurement labels employed by SHAPY.

Qualitative. Fig. 4 shows examples of our method running on a set of images of diverse body shapes from SSP3D, and on plus-size individuals in different clothing we downloaded online. It can be seen that our model improves the accuracy of shape estimation on a diverse set of bodies as compared to prior work [7]. The improvements are most noticeable for plus-sized individuals whose body characteristics are most under-represented in the training datasets. As shown in the bottom row, our refinement process also improves the quality of the estimated pose.

Ablation Study. Table 2 demonstrates that the Stage 3 refinement step leads to a 17.3% improvement on PVE-T-SC and our proposed T-pose loss improves the performance of our feed-forward network by 2.6%. Note that our method performs on par with SHAPY with the test-time refinement step removed.

6. Conclusion

This paper explores an important and understudied problem of estimating the 3D shape of humans whose body characteristics are under-represented in computer vision datasets. We demonstrate two shape-specific loss functions and a test-time iterative refinement technique that improves the quality of shape estimates for this group, as tested on the challenging SSP-3D dataset. We achieve this without using semantic text attributes or body measurements for training.

References

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proc. CVPR*, 2022. 2

[2] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out?: 3D animal reconstruction with expectation maximization in the loop. In *Proc. ECCV*, 2020. 2

[3] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and SMAL: Recovering the shape and motion of animals from video. In *Proc. ACCV*, 2018. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Proc. ECCV*, pages 561–578, 10 2016. 1, 2

[5] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. In *TPAMI*, 2019. 3

[6] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proc. ICCV*, 2021. 2

[7] Vasileios Choutas, Lea Muller, Chun-Hao P. Huang, Siyu Tang, Dimitris Tzionas, and Michael J. Black. Accurate 3d body shape regression using metric and semantic attribute. In *Proc. CVPR*, June 2022. 1, 2, 3, 4

[8] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proc. CVPR*, pages 7297–7306, 2018. 1, 3

[9] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. CVPR*, 2018. 2, 3

[10] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proc. ICCV*, 2019. 1, 2, 3

[11] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proc. CVPR*, 2019. 2

[12] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proc. CVPR*, July 2017. 2, 3

[13] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. Barc: Learning to regress 3d dog shape from images by exploiting breed information. In *Proc. CVPR*, 2022. 2

[14] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. ICCV*, October 2019. 2

[15] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proc. CVPR*, 2020. 2

[16] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *Proc. BMVC*, September 2020. 1, 2, 3

[17] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D Human Shape and Pose Estimation from Multiple Unconstrained Images in the Wild. In *Proc. CVPR*, June 2021. 2, 3

[18] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proc. ECCV*, September 2018. 2, 3

[19] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proc. CVPR*, pages 2863–2873, June 2022. 2

[20] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *Proc. CVPR*, July 2017. 1, 2