Towards Accurate 3D Human Body Reconstruction from Silhouettes

Brandon M. Smith, Visesh Chari , Amit Agrawal, James Rehg, Ram Sever
Amazon.com Inc.
{smithugh,viseshc,aaagrawa,jamerehg,severram}@amazon.com

Abstract

*We propose a novel computer vision system for reconstructing 3D body shapes from 2D images with the goal of producing highly accurate anthropomorphic measurements from a pair of images. We adopt a supervised learning approach that maps silhouette images to 3D body shapes via a convolutional neural network (CNN). We propose three key improvements over previous approaches: (1) Large-scale realistic synthetic data generation, including more realistic variations in segmentation noise and camera viewpoints. (2) A multi-task learning (MTL) approach to predicting multiple outputs such as shape, 3D joint locations, pose angles, and body volume. (3) A new network architecture that additionally takes known body measurements (*e.g.*, height) and per-pixel segmentation confidence as input. Ablation studies show the improvement in accuracy due to the various components of our system. Results demonstrate that our system produces state-of-the-art results on body circumference errors. We also analyze the repeatability of our system in the presence of realistic camera, background, and pose variations. Our system achieves a vertex standard deviation of $\sim 3mm$ on the CAESAR [36] dataset.*

1. Introduction

Inferring properties of human bodies from images is a fundamental, ill-posed problem in computer vision. Historically, the primary focus has been on estimating keypoints, which define the joints of the human figure in 2D and 3D, including the task of tracking keypoints over time. There is now growing interest in moving beyond the stick-figure view of the human body toward recovering richer representations of shape. For example, recent approaches estimate body segments [18], dense correspondences [20], or 3D volumetric descriptions of bodies from images via, *e.g.*, voxels [17, 41], Gaussian density functions [35], and deformable pre-defined meshes [15]. Interest in the latter representation has coincided with the availability of standardized body shape models, such as SCAPE [3] and SMPL [29], and 3D body datasets, such as CAESAR [36], from which these models were constructed.

Recent deep learning approaches have demonstrated compelling results for estimating 3D body shape information from RGB images "in the wild," meaning from real-world images containing unconstrained poses [25, 26, 32, 33, 42, 41]. In contrast, our goal is recovering *accurate* and *repeatable* anthropomorphic measurements, such as the body dimensions illustrated in Table 1. In order to ensure that detailed shape reconstruction is feasible, we consider images of people in an 'A' pose with legs separated slightly and arms to the side away from the torso. Even with a canonical pose, there are substantial challenges, including appearance variations, poor illumination, camera perspective, and sensor noise in real-world images that make accurate estimation difficult.

Similar to [14, 15, 16], we utilize body segmentation masks as our primary input source and present a novel deep learning architecture called *BfSNet* (Body from Silhouette Network) for recovering 3D parametric model fits that lead to more accurate anthropomorphic measurements compared to state-of-the-art approaches, such as Dibra *et al*. [14] and Kanazawa *et al*. [25]. In real-world applications, the repeatability (test–retest reliability) of the system is also important, but is not sufficiently analyzed in previous works. We analyze our system's repeatability for the *same* body shape across realistic variations in camera position and orientation, backgrounds, and segmentation errors. Under these challenging conditions, our system achieves a mean vertex standard deviation of 3.09mm.

Contributions The key contributions of our approach are:

1. Large-scale realistic synthetic data generation to augment the limited 3D data available for training;
2. Multi-task learning [31] for predict additional outputs to further constrain the estimation process; and
3. Using additional network inputs such as height and segmentation confidence maps to improve robustness to missing body parts and segmentation errors.

2. Related Work

In this section we review related work and provide context for our contributions.

Anthropomorphic Measurements: Prior to Boisvert *et al*. [5] in 2013, few papers evaluated 3D reconstruction accuracy using a variety of anthropomorphic measurements. There were a few exceptions, *e.g.*, [6, 19, 38], but even

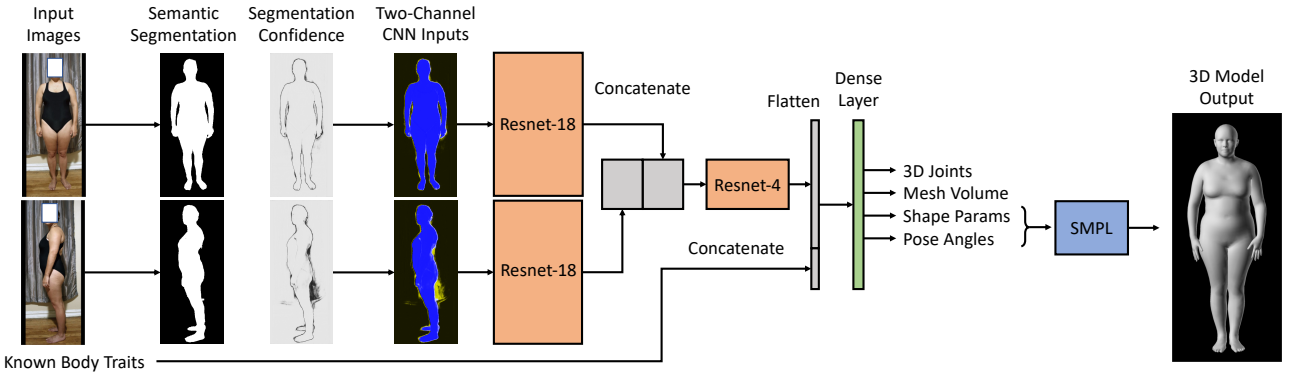


Figure 1. Our proposed system, called *BfSNet*, for 3D body shape reconstruction from noisy silhouettes. We train gender-specific CNNs, which are constructed to use multi-modal inputs (binary silhouettes, segmentation confidence masks, known body attributes such as height), and predict multi-task outputs (3D joint locations, mesh volume, shape parameters, and pose angles). The combination of multi-modal inputs, multi-task outputs, and a deeper network with modern features (*e.g.*, batch normalization, residual blocks) results in more accurate and repeatable 3D reconstructions, which we demonstrate in Section 4. Please see Figure 5 for the source of the input images, and the supplementary material for architecture details.

in those cases, measurements were limited to two or three types (*e.g.*, height, chest and waist circumference) and anthropomorphic measurements were not the focus. In 2016, Dibra *et al.* [14] demonstrated, for the first time, the effectiveness of using CNNs to map silhouette images directly to 3D body shapes, and subsequently introduced Heat Kernel Signature (HKS) descriptors [39] as an intermediate shape representation [15]. Our approach builds on these ideas and we show several improvements using synthetic data generation and multi-task learning. In Section 4, we quantitatively compare with Dibra *et al.* [14] and others [5, 10, 15, 43]. On an existing benchmark, which assumes noise- and error-free silhouettes and a fixed calibrated camera, we demonstrate favorable accuracy. On a more realistic test dataset with a range of camera heights (0 to 2 meters) and tilts (-30 to +30 degrees), partial occlusions, segmentation boundary noise, and large segmentation errors, we show that BfSNet produces a mean measurement error of 7.7mm compared to 11.7mm for Dibra *et al.* [14].

Learning from Synthetic Data: CNNs perform best when they can learn from large, diverse datasets. Unfortunately, databases suitable for learning a mapping between 3D bodies and 2D images are relatively small, and scanning equipment is expensive. For example, SizeUSA [1] and CAESAR [36] were introduced more than 15 years ago and include only a few thousand scans, but remain popular 'large' databases.

Learning from synthetic data is a popular strategy for overcoming the lack of 3D data. For example, the SURREAL (Synthetic hUmans for REAL tasks) database [42] consists of 1k body textures, 70k background images, and 4k body shapes created using the SMPL body model [29]. It combines different textures, body shapes, backgrounds, lighting, and virtual camera locations to generate a corpus of 6 million synthetic images. Several approaches such as [26, 40, 41] use SURREAL for learning to estimate body shape, pose, camera location and orientation from images.

Direct Recovery from Images in Unconstrained Poses: Several previous approaches have used a render-and-compare strategy to leverage existing, large-scale 2D image datasets by adding additional loss terms during training. Bogo *et al.* [4] showed that even 2D keypoints provide enough constraints in many cases to fit 3D models to images. Given a collection of high-quality fits, it's possible to train discriminative models for 2D and 3D landmark estimation and body part segmentation [27]. Kanazawa *et al.* [25] extended this idea by training an end-to-end system that maps image pixels directly to model parameters. Kundu *et al.* [26], Pavlakos *et al.* [33], and Omran *et al.* [32] proposed CNN-based approaches with differentiable render-and-compare training losses, allowing 3D shape and pose to be learned from extensive 2D datasets. Popa *et al.* [34] proposed a deep multi-task architecture for estimating both pixel-level body part labels and 2D and 3D keypoints. Zanfir *et al.* [45] extended this work, focusing on visually plausible 3D reconstructions of multiple people in a scene. The training data used by these approaches depict people in unconstrained poses, and are therefore better suited for tasks like pose estimation, body part segmentation, and depth estimation,rather than accurate anthropomorphic reconstruction, which is our goal.

3D Body Reconstruction from Videos: Alldieck *et al.* [2] proposed a method for reconstructing 3D bodies from videos (*e.g.*, 120 frames) within an iterative optimization framework, similar in spirit to classical shape-from-silhouette approaches. In contrast, our system relies on an efficient deep learning approach, uses only two silhouette images per example, and computes results in seconds rather than minutes or hours. For fair comparison, we include only one- or two-view approaches in our experimental re-

sults section.

Segmentation Robustness: In real-world scenarios, segmentation can be noisy, which can affect the accuracy of 3D modeling from silhouettes. Typically, the network is trained with noise augmentations (*e.g.*, silhouette boundary noise, occlusions) [14, 16] so that it learns features and a mapping that are more robust to noise. Previous methods relied on foreground/background segmentation labels, with no consideration of pixel-level confidence. We train a CNN with an additional segmentation confidence input and show that it improves the estimates.

Body Models: Several deformable parametric 3D body models have been proposed over the years. One of the most popular has been the SCAPE (Shape Completion and Animation of People) method of building a 3D body model [3]. The SCAPE model was data-driven, and learned pose and shape (a.k.a. phenotype [11]) variation separately. SCAPE has been used for reconstructing 3D shapes from images (mostly silhouettes) [6, 7, 9, 19, 38]. More recently, Loper *et al*. [29] proposed the SMPL model, which is now a dominant representation among state-of-the-art methods, *e.g.*, [25, 26, 27, 32, 33]. We used SMPL for our body modeling

3. Proposed Method

In this section we begin with a high-level overview of our approach, and then describe the technical details.

3.1. Overview

Figure 1 illustrates our overall system. We train a deep neural network that takes as inputs binary segmentation, along with segmentation confidence and a person's known traits (*e.g.*, gender, height, weight), and outputs SMPL shape and pose parameters. The first two blocks are Resnet-18 [21] networks, and the third block after concatenation is a Resnet-4 network, followed by a densely-connected layer. Please see the supplementary material for details.

3.2. Background: SMPL Model

The SMPL (Skinned Multi-Person Linear) model [29] is a realistic data-driven model of 3D human shape and pose. SMPL was trained using the CAESAR (Civilian American and European Surface Anthropometry Resource) dataset [36], which is composed of approximately 2k scans per gender. SMPL decomposes innate 3D body shape and pose. Innate 3D body shape variation is modeled linearly, and all body meshes share the same pre-defined topology on $V = 6890$ vertexes. Specifically, the vertices $\mathrm{v} \in \mathbb{R}^{3V}$ are parameterized by β via a simple linear equation:

$$\mathrm{v} = \mathbf{M}\beta + \mu, \tag{1}$$

where $\mathbf{M}$ is computed via principle component analysis (PCA) and μ is the mean shape. SMPL provides either a gender-neutral model or a gender-specific model. We adopt a gender-specific model while assuming known gender. Pose is parameterized by local 3D rotation angles θ on 24 skeleton joints. The final articulated mesh is a function of shape and pose, and is achieved by *blend shapes*, which are learned from data and correct for the limitations of standard linear blend skinning.

3.3. Training Loss

We train our CNN with the following loss terms:

1. $\mathcal{L}_{\mathsf{vertex}}$: Mean vertex-to-vertex error in a fixed pose,
2. $\mathcal{L}_{\mathsf{vol}}$: Mesh volume error,
3. $\mathcal{L}_{\mathsf{joints}}$: Error on articulated 3D skeleton joint locations,
4. $\mathcal{L}_{\mathsf{pose}}$: Error on articulated 3D joint angles.

$\mathcal{L}_{\mathsf{vertex}}$ is computed as:

$$\mathcal{L}_{\mathsf{vertex}} = \frac{1}{V}\sum_{j}^{V} w_j \|\mathrm{v}_j^{\mathsf{pred}} - \mathrm{v}_j^{\mathsf{true}}\|^2, \tag{2}$$

where $\mathrm{v}_j^{\mathsf{pred}}$ and $\mathrm{v}_j^{\mathsf{true}}$ are the j-th predicted and true 3D vertex locations, respectively. $\mathcal{L}_{\mathsf{vertex}}$ is weighted in a way that compensates for the non-uniform distribution of vertices in the SMPL model. Specifically, the weight w_j for vertex j is proportional to the average area of the mesh triangles connected to vertex j. Intuitively, this prevents body regions like the hands and face, which have many tightly-space vertices, from dominating $\mathcal{L}_{\mathsf{vertex}}$. For pose error, the model predicts global rotation matrix entries, and then the rotations are converted back to local axis angles during inference. Including volume estimation improves body shape accuracy, as we show in Table 2. The final training loss is:

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N} \mathcal{L}_{\mathsf{vertex}} + \alpha_{\mathsf{vol}}\mathcal{L}_{\mathsf{vol}} + \alpha_{\mathsf{joints}}\mathcal{L}_{\mathsf{joints}} + \alpha_{\mathsf{pose}}\mathcal{L}_{\mathsf{pose}}, \tag{3}$$

where N is the training batch size, and each term penalizes L2 error. The units of $\mathcal{L}_{\mathsf{vertex}}$ and $\mathcal{L}_{\mathsf{joints}}$ are squared millimeters, $\mathcal{L}_{\mathsf{vol}}$ is squared liters, and $\mathcal{L}_{\mathsf{pose}}$ is the mean squared error of pose rotation matrix entries. We set $\alpha_{\mathsf{vol}} = 0.6$, $\alpha_{\mathsf{joints}} = 0.01$, and $\alpha_{\mathsf{pose}} = 1.0$ empirically in order to minimize body measurement errors. With these weights, $\mathcal{L}_{\mathsf{vertex}}$ is large compared to the other terms, which reflects the fact that we care most about shape accuracy.

Improving Training Time: In the SMPL model, vertices v are a function of pose-dependent blend shapes. However, we observe that mesh corrections due to the pose blend shapes are negligibly different for nearby poses, *e.g.*, poses around the 'A' pose. Therefore, we construct a linear transformation between β and v that takes into account the blend shape for the average 'A' pose. After substituting Eq. 1 into

Eq. 2 and simplifying, we obtain

$$\mathcal{L}_{\text{vertex}} = \Delta\beta^{\mathrm{T}}(\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{M})\Delta\beta, \text{ where} \quad (4)$$

$$\Delta\beta = \beta^{\text{pred}} - \beta^{\text{true}}, \quad (5)$$

$$\mathbf{W} = \tfrac{1}{V}\text{diag}([w_1, w_2, \cdots, w_V]), \quad (6)$$

where the matrix $(\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{M})$ is pre-computed for efficiency. We set $B = 300$ (the maximum available in the SMPL model). The first few dimensions of β control most of the shape variation in the model, but we let the network learn the importance of each dimension of β on its own. This approach is in contrast to previous methods that set B to 10-30, *e.g.*, [14, 15, 42].

We investigated training with only a vertex-to-vertex loss on articulated body shape, but found it to be dramatically slower, and produce less accurate anthropomorphic measurements than training with a vertex-to-vertex loss on pose-normalized body shapes. On the other hand, directly predicting articulated 3D joint locations and mesh volume empirically improve shape accuracy, as shown in Table 2. It is also possible to prioritize the accuracy of certain measurements, *e.g.*, waist circumference, via additional loss terms. However, because SMPL is a global shape model, this comes at the expense of the accuracy of other measurements.

3.4. Segmentation and Silhouette Pre-Processing

We compute silhouettes via semantic image segmentation. Specifically, starting from DeepLabv3+ [8], we fine-tuned the model on a collection of CAESAR scans [36] rendered in front of random background images; Section 4.3 describes this dataset. Semantic image segmentation generalizes better to real-world scenarios like cluttered backgrounds and camera motion than simple background subtraction [22], which is a popular strategy in prior work. During training, we modified the loss as described in Section 3.5 so that, in addition to foreground and background labels, the network also outputs per-pixel confidence values.

The silhouette and confidence images are normalized before feeding them to the CNN. We first crop the silhouette by computing the tightest bounding rectangle around it, and then resize it according to s, which is a function of the subject's known height, h: $0.8 \cdot \frac{h}{\mu_h}\text{image}_h$, where image_h is the input image height and μ_h is the average height of a person. The resized silhouette is then centered in the image. This has the effect of placing subjects at approximately the same distance from the virtual camera. We tested other strategies, such as scaling the silhouette to a uniform pixel height, but found that the above normalization results in the most accurate shape predictions.

3.5. Confidence Estimation

DeVries and Taylor [12, 13] proposed a simple modification to the final loss that allows a network to additionally output a confidence, which we adopt. The segmentation network outputs a softmax class prediction probability p_i for each pixel p and each label type i (foreground $i = 1$, background $i = 0$). Each p_i is adjusted by interpolating between the original prediction and the target probability distribution y, where the degree of interpolation is indicated by the network's confidence c_i:

$$p'_i = c_i \cdot p_i + (1 - c_i)y_i. \quad (7)$$

The task loss ℓ_t is computed as usual using the updated prediction probabilities p'_i. In order to prevent the network from always choosing $c_i = 0$ an extra confidence loss term ℓ_c is added to the final loss:

$$\ell = \ell_t + \lambda\ell_c, \qquad \text{where } \ell_c = -log(c_i), \quad (8)$$

where λ is a hyperparameter that balances the two terms.

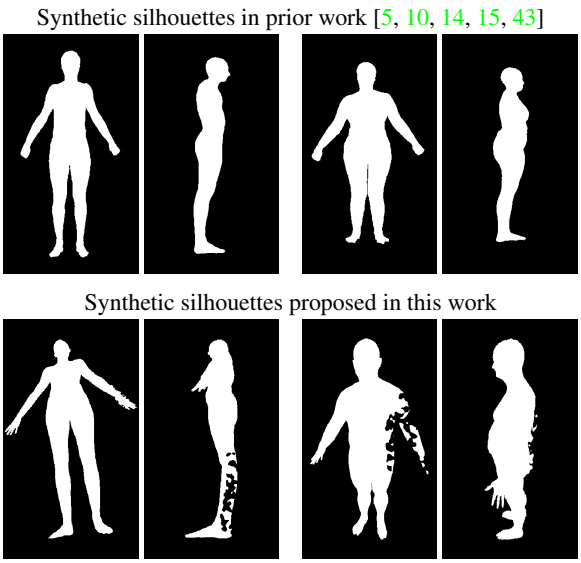


Figure 2. Synthetic data comparison. We introduce a synthetic training dataset that includes significantly more camera variation and segmentation errors (bottom row) compared to prior work (top row). These are important considerations for improving accuracy and repeatability in the real world.

3.6. Generating Synthetic Training Data

In this section we describe our process for generating millions of realistic synthetic training instances with a wide range of body shapes, virtual camera heights and tilts, natural body poses, and realistic segmentation artifacts. Each training instance is associated with ground truth SMPL shape and pose parameters, and the silhouette images for front and side views. Figure 2 shows a qualitative comparison between the kinds of silhouettes used for training and evaluation in prior work [5, 10, 14, 15, 43] and our synthetic silhouettes.

	SMPL model [29]		SCAPE model [3]				
Measurements	**BfSNet** (Our system)	Dibra '16 [14] Our implem.	Dibra '16 [14] From [14]	Dibra '17 [15]	Boisvert *et al.* [5]	Chen *et al.* [10] From [15]	Xi *et al.* [43]
A. Head circumference	5.1 ± 6.4	3.0 ± 3.8	$\mathbf{2} \pm 3$	3.2 ± 2.6	10 ± 12	23 ± 27	50 ± 60
B. Neck circumference	3.0 ± 3.9	3.0 ± 3.9	2 ± 1	$\mathbf{1.9} \pm 1.5$	11 ± 13	27 ± 34	59 ± 72
C. Shoulder to crotch	$\mathbf{1.5} \pm 2.2$	2.9 ± 3.8	3 ± 5	4.2 ± 3.4	4 ± 5	52 ± 65	119 ± 150
D. Chest circumference	4.7 ± 7.7	7.2 ± 9.2	$\mathbf{2} \pm 1$	5.6 ± 4.7	10 ± 12	18 ± 22	36 ± 45
E. Waist circumference	$\mathbf{4.8} \pm 7.5$	8.1 ± 10.2	7 ± 5	7.1 ± 5.8	22 ± 23	37 ± 39	55 ± 62
F. Pelvis circumference	$\mathbf{3.0} \pm 5.1$	6.0 ± 7.7	4 ± 4	6.9 ± 5.6	11 ± 12	15 ± 19	23 ± 28
G. Wrist circumference	2.5 ± 3.3	2.0 ± 2.7	2 ± 2	$\mathbf{1.6} \pm 1.3$	9 ± 12	24 ± 30	56 ± 70
H. Bicep circumference	2.7 ± 3.8	3.3 ± 4.2	$\mathbf{2} \pm 1$	2.6 ± 2.1	17 ± 22	59 ± 76	146 ± 177
I. Forearm circumference	1.9 ± 2.5	2.3 ± 2.9	$\mathbf{1} \pm 1$	2.2 ± 1.9	16 ± 20	76 ± 100	182 ± 230
J. Arm length	$\mathbf{1.7} \pm 2.4$	2.7 ± 3.5	3 ± 2	2.3 ± 1.9	15 ± 21	53 ± 73	109 ± 141
K. Inside leg length	$\mathbf{1.5} \pm 2.7$	2.8 ± 3.5	9 ± 6	4.3 ± 3.8	6 ± 7	9 ± 12	19 ± 24
L. Thigh circumference	$\mathbf{2.4} \pm 4.0$	4.9 ± 6.2	6 ± 4	5.1 ± 4.3	9 ± 12	19 ± 25	35 ± 44
M. Calf circumference	$\mathbf{2.3} \pm 3.6$	3.3 ± 4.3	3 ± 1	2.7 ± 1.9	6 ± 7	16 ± 21	33 ± 42
N. Ankle circumference	2.1 ± 2.8	2.0 ± 2.6	2 ± 1	$\mathbf{1.4} \pm 1.1$	14 ± 16	28 ± 35	61 ± 78
O. Overall height	$\mathbf{2.3} \pm 4.6$	4.0 ± 5.0	12 ± 10	7.1 ± 5.5	9 ± 12	21 ± 27	49 ± 62
P. Shoulder breadth	$\mathbf{1.9} \pm 2.5$	2.9 ± 3.6	2 ± 4	2.1 ± 1.8	6 ± 7	12 ± 15	24 ± 31
Mean measurement error	$\mathbf{2.72}$ mm	3.78 mm	4.02 mm	3.77 mm	11 mm	31 mm	66 mm

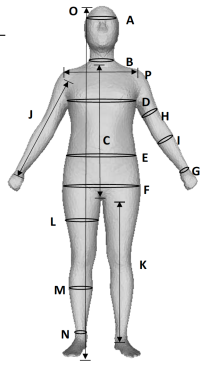


Table 1. Quantitative comparison with prior work. Column 3 is our implementation of Dibra *et al*. in 2016 [14]. Column 4 is reproduced from Dibra *et al*. 2016 [14]. Columns 5-8, and the body measurements definitions shown on the right, are reproduced from Dibra *et al*. 2017 [15]. BfSNet produces favorable or comparable accuracy (within 3mm) on all measurements compared to the state of the art.

Shape and Pose Sampling We use two strategies for sampling SMPL shape and pose. First, we construct multivariate Gaussian distributions for shape and pose parameters (joint axis angles), and randomly sample from the two distributions with shape $\sigma = 1$ and, to add more pose variation, pose $\sigma = 2$. Second, for a subset of CAESAR scans, we randomly sample the local shape and pose neighborhood centered on each scan with $\sigma = 0.1$, to create examples with slight perturbations around CAESAR instances. We also reject shapes that include self-intersection. In general, we use both strategies to enrich our training dataset, but for fair comparison with prior work in Section 4.1, we set $\sigma = 1$ for both shape and pose, and use only the global multivariate distribution.

Virtual Camera Sampling The camera position, orientation, and focal length were fixed for each viewpoint in prior work [5, 14, 15]. Indeed, for fair comparison with these works in Section 4.1 we trained and tested with images from two static virtual camera viewpoints. However, in order to match the amount of variation that we expect in the real world (*e.g.*, images from selfies) we generated a second dataset with body shapes rendered from a range of viewpoints, as describe below.

For each instance, the virtual camera is placed randomly at a distance between 1 to 2 meters at a height of 0 (floor plane) to 2 meters. Focal length is adjusted to ensure full body visibility. We further add realism by allowing the pose to change between the front and side view.

Adding Segmentation Noise to Improve Robustness We added two types of segmentation noise for robustness improvement: (1) segmentation boundary noise, and (2) large segmentation errors. Segmentation boundary noise is added by (A) dilating the silhouette a few pixels, (B) eroding the silhouette a few pixels, and randomly choosing between (A) and (B) for each pixel. Large segmentation errors are created by selecting regions of the silhouette at random and filling them with random splotches, similar to the examples in Figure 2. Synthetic confidence is set low in noise regions, and high everywhere else.

3.7. Implementation Details

The training and evaluation code was implemented in Python using the Keras framework with TensorFlow as the backend. We use 640×360 as the CNN input resolution, batch normalization [23], and ReLU activation layers [30]. We found that higher resolutions did not significantly improve the accuracy of the results. Images were rendered ahead of time using OpenGL.

4. Experimental Results and Discussion

In this section, we evaluate the accuracy and repeatability of our system. We first compare with recent methods on an established benchmark that measures the accuracy of 16 anthropomorphic measurements. Second, we perform an ablation study on a much larger and more challenging and realistic synthetic dataset to highlight the impact of different components of our system. Third, we compare with a state-of-the-art CNN-based end-to-end approach [25] that maps RGB pixels directly to 3D models. Fourth, we investigate the impact of segmentation errors on the accuracy and repeatability (test–retest reliability) of three models, each trained with (1) no large segmentation errors, (2) with large segmentation errors, and (3) with large segmentation errors and segmentation confidence masks as additional input channels to the network. Finally, we show qualitative results on images downloaded from the web.

4.1. Quantitative Results

We first present quantitative results and compare with recent methods [5, 10, 14, 15, 43]. Unfortunately, the specific

Measurements	BfSNet *2V-Late-HW-Conf-Vol-Pose*	*2V-Late-HW-Vol-Pose*	*2V-Late-HW*	*2V-Late*	*2V-Early*	*1V*	Dibra '16 [14] Our implem.
A. Head circumference	6.7 ± 8.4	8.0 ± 10.1	8.1 ± 10.4	8.8 ± 11.2	9.3 ± 11.7	8.9 ± 11.2	9.3 ± 11.7
B. Neck circumference	8.0 ± 10.1	8.8 ± 11.0	9.0 ± 11.5	9.3 ± 11.9	9.8 ± 12.4	9.0 ± 11.6	10.0 ± 12.8
C. Shoulder to crotch	5.1 ± 6.5	5.6 ± 7.1	5.9 ± 7.6	5.7 ± 7.4	6.4 ± 8.1	6.7 ± 8.5	6.6 ± 8.6
D. Chest circumference	12.5 ± 15.9	14.4 ± 18.5	16.0 ± 20.8	16.7 ± 21.8	18.8 ± 24.5	25.0 ± 31.8	22.8 ± 29.2
E. Waist circumference	15.8 ± 20.0	17.4 ± 22.3	17.8 ± 22.5	19.3 ± 24.5	20.1 ± 25.5	22.1 ± 28.5	24.0 ± 30.5
F. Pelvis circumference	9.3 ± 11.8	11.3 ± 14.3	12.0 ± 16.6	13.8 ± 19.9	15.9 ± 21.9	18.1 ± 24.2	20.0 ± 27.5
G. Wrist circumference	9.3 ± 13.4	9.6 ± 13.7	9.7 ± 13.6	9.8 ± 13.8	9.7 ± 13.7	9.7 ± 13.6	9.9 ± 13.8
H. Bicep circumference	8.1 ± 10.6	9.7 ± 12.2	9.7 ± 12.4	10.9 ± 14.0	10.6 ± 13.6	9.1 ± 12.0	12.0 ± 15.6
I. Forearm circumference	5.7 ± 7.1	6.2 ± 7.8	6.2 ± 7.9	7.2 ± 9.1	7.1 ± 8.9	6.6 ± 8.2	7.9 ± 9.9
J. Arm length	5.1 ± 6.4	5.7 ± 7.1	6.0 ± 7.4	5.5 ± 7.0	6.2 ± 7.7	5.9 ± 7.5	6.4 ± 8.0
K. Inside leg length	6.8 ± 8.6	7.5 ± 9.4	8.0 ± 10.1	7.2 ± 9.3	8.2 ± 10.5	8.4 ± 10.9	8.9 ± 11.5
L. Thigh circumference	8.8 ± 11.0	9.6 ± 12.3	10.3 ± 13.7	11.5 ± 15.7	12.5 ± 16.7	13.3 ± 17.6	15.5 ± 20.4
M. Calf circumference	7.2 ± 9.1	7.9 ± 10.0	8.1 ± 10.4	9.6 ± 12.4	9.9 ± 12.6	9.2 ± 11.8	13.2 ± 16.6
N. Ankle circumference	5.0 ± 6.4	5.6 ± 7.1	5.6 ± 7.2	6.2 ± 8.0	6.5 ± 8.2	5.9 ± 7.6	7.6 ± 9.5
O. Overall height	5.8 ± 7.5	6.4 ± 8.2	6.8 ± 9.1	6.9 ± 9.2	7.3 ± 9.6	7.5 ± 9.9	7.8 ± 10.2
P. Shoulder breadth	4.5 ± 5.7	5.1 ± 6.5	5.2 ± 6.7	5.4 ± 6.8	5.7 ± 7.3	5.6 ± 7.1	6.0 ± 7.6
Mean measurement error	7.7 mm	8.7 mm	9.0 mm	9.6 mm	10.2 mm	10.7 mm	11.7 mm

Table 2. Ablation study. *1V:* Single view input. *2V-Early:* Two input views with early fusion. *2V-Late:* Two input views with late fusion. *2V-Late-HW:* Two input views with late fusion and known height and weight as input. *2V-Late-HW-Vol-Pose:* Two input views with late fusion, known height as input, and additional multi-task outputs. *2V-Late-HW-Conf-Vol-Pose:* (full BfSNet system) using confidence masks as additional input. For reference, accuracy of our implementation of Dibra *et al.*'s 2016 system [14] is also shown.

dataset used by these methods for training and testing is not available, and so we recreated it according to the specifications in [15]. First, we randomly sampled 500k meshes from the CAESAR [36] pose and shape multivariate distributions. Dibra *et al.* [15] used the SCAPE model for this purpose, and they limited their shape space to the first 20 PCA bases. We instead use the SMPL model fit to CAESAR to sample its distribution, but we also limit the number of shape bases to 20. For a fair comparison, we reimplemented the CNN-based approached proposed by Dibra *et al.* [14] and trained it to predict SMPL model parameters instead. Second, we rendered each mesh instance to front and side silhouettes using stationary, calibrated virtual cameras. Like Dibra *et al.* [14, 15], we set 1k instances aside for testing, and the remaining instances for training (249.5k per gender). Evaluating on the CAESAR dataset is not ideal. However, a significant challenge in this research area is a lack of good, publicly available benchmarks, *e.g.*, 3D body scans with accompanying RGB images.

Table 1 shows that BfSNet produces circumference and length estimates with errors of 5mm or less on average, with favorable or comparable accuracy with respect to recent work. Note that this benchmark uses perfect silhouettes with no camera height or tilt variation and no self-occlusion in the front view (*e.g.*, hands are never in front of hips). This lack of realism motivates our more challenging dataset.

4.2. Ablation Study on a More Challenging Dataset

We trained and evaluated different versions of our system on the challenging synthetic dataset described in Section 3.6. Specifically, we generated 1.5 million {front, right}-view pairs for training, 2k for validation, and 10k for testing per gender. Table 2 shows the accuracy of results from ablated versions of our pipeline, which highlights the contribution of each component to the overall accuracy. We also trained and tested our implementation of Dibra *et al.* [14] on the same dataset for fair comparison. Our proposed system produces significantly more accurate results (7.4mm vs. 11.7mm measurement error) compared to Dibra *et al.* [14].

In Table 2, *1V* is a Resnet-50 network [21] that uses a single front input view, and predicts only shape parameters. *2V-Early* is the same architecture as *1V*, except the side-view image is added as a second channel two the input image. Importantly, the two views are preprocessed to bring them into approximate correspondence before feeding them to the CNN, *i.e.*, the silhouettes are height-normalized and centered in the image. *2V-Late* is the network architecture illustrated in Figure 1, which extracts features from each viewpoint separately via a sequence of 5 Resnet blocks; the features are then concatenated and fed to an additional Resnet block and two dense layers. *2V-Late-HW* adds known body traits (height, weight) as inputs. *2V-Late-HW-Vol-Pose* adds multi-task outputs (mesh volume, 3D joint locations, and pose angles). *2V-Late-HW-Conf-Vol-Pose* (full BfSNet system) adds confidence masks as additional input channels to each input image.

Figure 3 shows the impact of segmentation noise and confidence on average vertex-to-vertex accuracy. We observe that the model trained without noise does not generalize well to noisy inputs, training with noise offers a significant improvement, and training with noise and segmentation confidence produces the most accurate reconstruction on noisy segmentation masks.

4.3. Comparison with a Direct-from-RGB Approach

Kanazawa *et al.* [25] proposed an impressive method for estimating 3D body models directly from RGB images, which represents the state of the art among similar

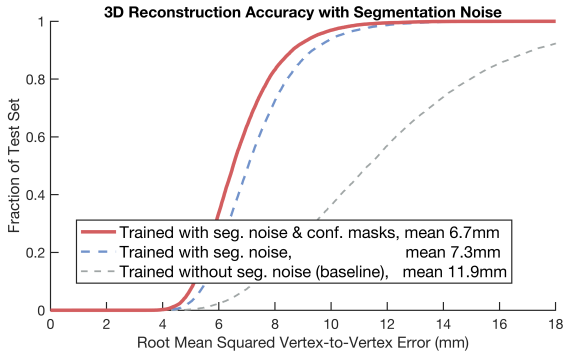


Figure 3. Cumulative error distribution of average vertex-to-vertex reconstruction error for different versions of the pipeline tested on 10k examples with corrupted segmentation masks. Corruptions are similar to the example shown in Figure 2.

approaches [25, 26, 32, 33]. For fair comparisons with [25], we generated a test dataset of color renderings with known body shapes. Specifically, we selected 1000 participants from the CAESAR database [36] and rendered front and side views of each scan in front of random backgrounds from the LSUN dataset [44], which is the same dataset used for background images in the SURREAL dataset [42]. Please see our supplementary material for example images, segmentation results, and estimated body models.

Kanazawa *et al*. and similar methods are trained in a weakly supervised manner from images with the ambitious goal of handling unconstrained poses scenarios. For this reason, their system is well-suited for semantic segmentation and pose estimation, but not on estimating accurate anthropomorphic measurements. Figure 4 highlights the fact that *the estimates from Kanazawa* et al. *tend to have average body shape, regardless of the input image.* In contrast, BfSNet produces significantly more accurate measurements, as shown in Table 3.

4.4. Repeatability Analysis

We now analyze the *repeatability* of our system in order to understand how segmentation errors give rise to variations in body shape estimates. Figure 6 shows repeatability analysis on 1000 test examples from the ablation study, using the *2V-Late-HW-Vol-Pose* (no confidence input) and *2V-Late-HW-Conf-Vol-Pose* (full system with confidence input) models. The average vertex standard deviation with and without confidence is 1.96mm and 2.38mm, respectively, highlighting the robustness of our system.

We also investigated the repeatability of our system on more realistic CAESAR renderings. We rendered each color scan in front of ten different backgrounds, and virtual camera viewpoints. This produced a dataset of 10k instances from which 9k were used for fine-tuning. On the remaining 1k instances, the average standard deviation across all examples and vertices was 3.09mm, which indicates BfSNet is robustness for camera changes and realistic segmentation noise.

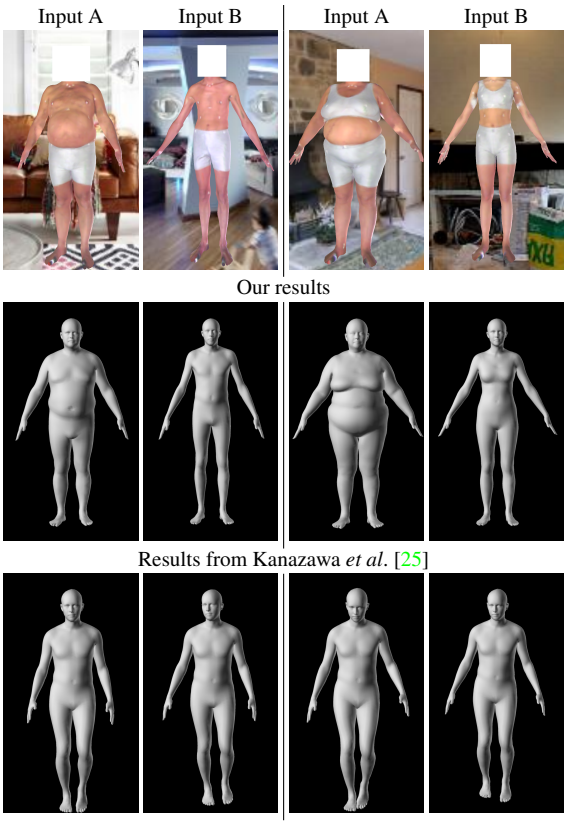


Figure 4. Qualitative comparison with state-of-art approach of [25] that estimates 3D models directly from RGB images (rendered scans from the CAESAR database). Our proposed method estimates body dimensions more accurately.

Measurements	BfSNet	Kanazawa*et al*. [25]
A. Head circumference	14.2 ± 18.6	16.7 ± 26.5
B. Neck circumference	11.4 ± 18.7	35.7 ± 63.3
C. Shoulder to crotch	11.0 ± 13.8	33.8 ± 42.2
D. Chest circumference	16.2 ± 20.6	92.8 ± 116.5
E. Waist circumference	25.0 ± 32.1	118.3 ± 146.7
F. Pelvis circumference	15.2 ± 19.6	68.7 ± 90.0
G. Wrist circumference	5.5 ± 7.0	12.2 ± 15.1
H. Bicep circumference	10.4 ± 13.5	29.3 ± 37.6
I. Forearm circumference	7.9 ± 10.1	20.6 ± 25.8
J. Arm length	6.0 ± 7.7	29.9 ± 41.7
K. Inside leg length	8.0 ± 10.1	44.3 ± 58.9
L. Thigh circumference	11.1 ± 14.2	38.5 ± 49.7
M. Calf circumference	10.4 ± 13.3	25.8 ± 33.2
N. Ankle circumference	6.3 ± 8.1	14.0 ± 17.6
O. Overall height	7.9 ± 10.5	76.2 ± 97.9
P. Shoulder breadth	8.4 ± 10.7	26.5 ± 32.0
Mean measurement error	10.9 mm	42.7 mm

Table 3. Quantitative comparison (mean $\pm$ standard deviation) with [25] on 1000 rendered scans from the CAESAR database. Each test example is similar to the examples in Figure 4. Our system is better suited for accurately estimating anthropomorphic measurements.

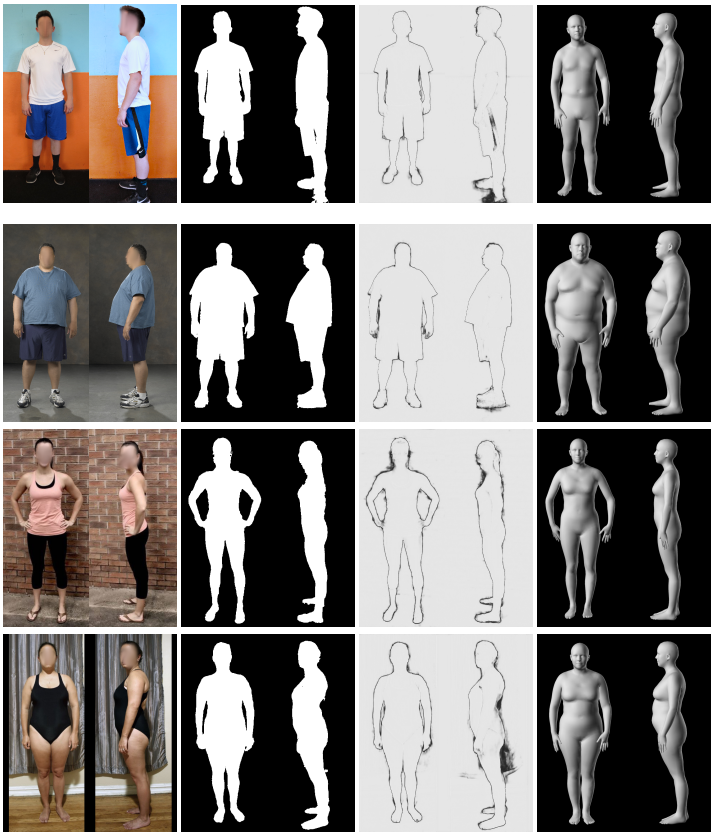
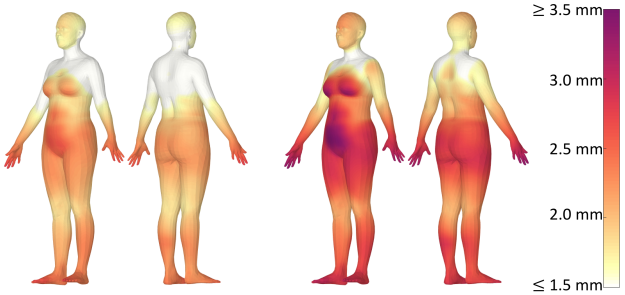


Figure 5. Qualitative results from our full system on images from the web. **First column:** input images. **Second column:** silhouettes computed via semantic segmentation. **Third column:** segmentation confidence (black is low confidence). **Fourth column:** body model result. Our approach is robust to different camera viewpoints, illumination conditions, and natural body pose variation around an 'A' pose. **Attribution:** The input images were downloaded from flickr.com and are free to share via creativecommons.org/licenses/by-nd/2.0/ (Rows 1 and 2) or creativecommons.org/licenses/by/2.0/ (Rows 3 and 4). Row 1 : flic.kr/p/23XQEJv and flic.kr/p/2euFez3. Row 2: flic.kr/p/efiXWW and flic.kr/p/efiWSS. Row 3: flic.kr/p/ S6EeZA. Row 4: flic.kr/p/6tfXfP.



Avg. standard deviation with confidence, 1.96 mm

Avg. standard deviation without confidence, 2.38 mm

Figure 6. Repeatability with and without segmentation confidence. Ten copies of each test instance were created by randomly corrupting the segmentation masks. The heat maps show the average standard deviation for each vertex across test instances. Repeatability is better when confidence masks are add as network inputs.

4.5. Qualitative Results

Figure 5 shows several results on web images. The visual shape of each estimated 3D model closely reflects the visual shape of the person in each pair of images.

5. Conclusions and Future Work

We have presented a novel method for reconstructing 3D body shape from 2D binary silhouettes. In contrast to approaches that target people in unconstrained poses, we focused on the task of estimating accurate and repeatable anthropomorphic measurements. Results demonstrate that our system is more accurate than previous approaches, with good repeatability. Key to our improvements are (1) large-scale synthetic data generation incorporating realistic variations in camera height and tilt, and segmentation errors; (2) a multi-task approach to estimate body shape, 3D joint locations, 3D pose angles, and body volume simultaneously; and (3) a novel architecture that takes multiple kinds of inputs, including segmentation confidence and known body traits. An important future direction is to broaden the range of acceptable poses and camera viewpoints while maintaining the same performance. Toward this goal, it will be important to continue to improve the realism of large synthetic datasets, *e.g.*, by better aligning the distributions of synthetic and real segmentation masks using GANs [37].

References

[1] TC2 Labs LLC, SizeUSA. http://scan2fit.com/sizeusa/about.php. Accessed: 2018-09-30. 2

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3D people models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape completion and animation of people. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 24(3):408–416, July 2005. 1, 3, 5

[4] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[5] J. Boisvert, C. Shu, S. Wuhrer, and P. Xi. Three-dimensional human shape inference from silhouettes: Reconstruction and validation. *Machine Vision and Applications*, 24(1):145–157, 2013. 1, 2, 4, 5

[6] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision (ECCV)*, 2008. 1, 3

[7] A. O. Bălan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 3

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 4

[9] X. Chen, Y. Guo, B. Zhou, and Q. Zhao. Deformable model for estimating clothing and naked human shapes from a single image. *The Visual Computer*, 29(11):1187–1196, 2013. 3

[10] Y. Chen, T.-K. Kim, and R. Cipolla. Inferring 3d shapes and deformations from single views. In *European Conference on Computer Vision*, 2010. 2, 4, 5

[11] Y. Chen, T.-K. Kim, and R. Cipolla. Silhouette-based object phenotype recognition using 3d shape priors. In *International Conference on Computer Vision (ICCV)*, 2011. 3

[12] T. DeVries and G. W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 4

[13] T. DeVries and G. W. Taylor. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502*, 2018. 4

[14] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross. HS-Nets: Estimating human body shape from silhouettes with convolutional neural networks. In *International Conference on 3D Vision (3DV)*, 2016. 1, 2, 3, 4, 5, 6

[15] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 4, 5, 6

[16] E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. Shape from selfies: Human body shape estimation using CCA regression forests. In *European Converence on Computer Vision (ECCV)*, 2016. 1, 3

[17] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton. Volumetric performance capture from minimal camera viewpoints. In *European Conference on Computer Vision*, 2018. 1

[18] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[19] P. Guan, A. Weiss, A. O. Bălan, and M. J. Black. Estimating human shape and pose from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 1, 3

[20] R. A. Güler, N. Neverova, and I. Kokkinos. DensePose: Dense human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 6

[22] T. Horprasert, D. Harwood, and L. S. Davis. A statistical approach for real-time robust background subtraction and shadow detection. In *IEEE International Conference on Computer Vision (ICCV)*, 1999. 4

[23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015. 5

[24] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 5, 6, 7

[26] A. Kundu, Y. Li, and J. M. Rehg. 3D-RCNN: Instance-level 3d object reconstruction via render-and-compare. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 7

[27] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people – closing the loop between 3D and 2D human representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[28] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[29] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 2, 3, 5

[30] V. Nair and G. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning (ICML)*, 2010. 5

[31] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, pages 689–696, 2011. 1

[32] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, 2018. 1, 2, 3, 7

[33] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 7

[34] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[35] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European Conference on Computer Vision*, 2016. 1

[36] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, S. Fleming, T. Brill, D. Hoeferlin, and D. Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) final report. *Tech. Rep. AFRL-HEWP-TR-2002-0169, US Air Force Research Laboratory*, 2002. 1, 2, 3, 4, 6, 7

[37] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[38] L. Sigal, A. O. Bălan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Neural Information Processing Systems (NIPS)*, 2007. 1, 3

[39] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Symposium on Geometry Processing*, 2009. 2

[40] H.-Y. F. Tung and H.-W. T. E. Y. K. Fragkiadaki. Self-supervised learning of motion capture. In *Neural Information Processing Systems (NIPS)*, 2017. 2

[41] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. BodyNet: Volumetric inference of 3d human body shapes. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[42] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 4, 7

[43] P. Xi, W.-S. Lee, and C. Shu. A data-driven approach to human-body cloning using a segmented body database. In *Pacific Conference on Computer Graphics and Applications (PG)*, 2007. 2, 4, 5

[44] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 7

[45] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2