Probabilistic Estimation of 3D Human Shape and Pose with a Semantic Local Parametric Model

Akash Sengupta
as2562@cam.ac.uk
Ignas Budvytis
ib255@cam.ac.uk
Roberto Cipolla
rc10001@cam.ac.uk

Department of Engineering
University of Cambridge
Cambridge, UK

Abstract

This paper addresses the problem of 3D human body shape and pose estimation from RGB images. Some recent approaches to this task predict probability distributions over human body model parameters conditioned on the input images. This is motivated by the ill-posed nature of the problem wherein multiple 3D reconstructions may match the image evidence, particularly when some parts of the body are locally occluded. However, body shape parameters in widely-used body models (*e.g.* SMPL) control global deformations over the whole body surface. Distributions over these *global* shape parameters are unable to meaningfully capture uncertainty in shape estimates associated with *locally*-occluded body parts. In contrast, we present a method that (i) predicts distributions over local body shape in the form of semantic body *measurements* and (ii) uses a linear mapping to transform a local distribution over body measurements to a global distribution over SMPL shape parameters. We show that our method outperforms the current state-of-the-art in terms of identity-dependent body shape estimation accuracy on the SSP-3D dataset, and a private dataset of tape-measured humans, by probabilistically-combining local body measurement distributions predicted from multiple images of a subject.

1 Introduction

3D human shape and pose estimation from RGB images is a challenging computer vision problem, with direct applications in virtual retail, virtual reality and computer animation. Several deep-learning-based approaches to this task yield impressive human pose estimates [6, 9, 10, 15, 20, 21, 25, 45]. However, body shape estimates tend to be inaccurate or inconsistent for subjects in-the-wild. Recently, [35, 36] attempt to predict accurate and consistent body shapes from multiple images of a subject, without assuming a fixed body pose or background and lighting conditions. This involves (i) predicting independent Gaussian distributions (i.e. with diagonal covariance matrices) over SMPL [24] shape parameter vectors conditioned on the input images and (ii) probabilistically combining the shape distributions

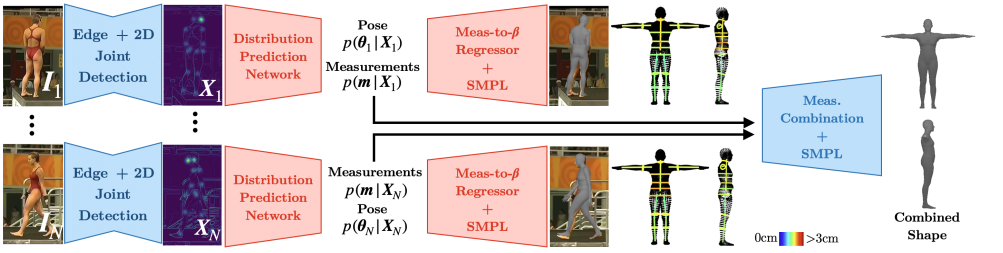

Figure 1: Our three-stage approach for body measurement and pose estimation from a set of images. Each image is converted into an edge and joint heatmap proxy representation, which is passed through a distribution prediction network that yields distributions over body measurements and pose conditioned on the input images. Individual measurement distributions are probabilistically combined into a final measurement estimate, which can be mapped to SMPL shape coefficients using the proposed measurements-to-βs linear regressor.

predicted from each image, to yield a final consistent shape estimate. However, independent Gaussian distributions over SMPL shape parameters are unable to quantify uncertainty in *local* body parts, since SMPL shape parameters (i.e. coefficients of a PCA shape space) control shape deformation over the *global* body surface. Given multiple images of a subject, meaningful probabilistic shape combination benefits from local shape uncertainty estimation, where part-specific uncertainty arises from variation in camera viewpoints and body poses within the images, as well as occlusion (see Figures 3 and 4).

To this end, we extend [36] by predicting distributions over local semantic body shape *measurements* (*e.g.* chest width, arm length, calf circumference, etc), conditioned on an input image. This necessitates learning a mapping from semantic body measurements to SMPL shape coefficients (βs), which enables local, human-interpretable control of SMPL body shapes. Independent Gaussian distributions defined over measurements translate to localised uncertainty over SMPL T-pose vertices (as shown in Figure 2), in contrast with independent Gaussian distributions over SMPL βs. Furthermore, we define the mapping from measurements to SMPL βs to be a linear regression. Thus, a Gaussian distribution over measurements can be analytically transformed into a distribution over SMPL βs, and subsequently 3D vertex locations, using simple linear transformations (see Equation 3).

Having learned a linear mapping from measurements to SMPL βs, our pipeline for 3D multi-image body shape and pose estimation consists of 3 stages (see Figure 1). First, we compute proxy representations using an off-the-shelf 2D keypoint detector [12, 42] and Canny edge detection [6]. This decreases the domain gap between synthetic training data and real test data [35]. Second, a deep neural network predicts means and variances of Gaussian distributions over SMPL pose parameters and body measurements, conditioned on the input proxy representations. Third, body measurements from each image are probabilistically combined [36] to give a final measurements estimate, which is converted into a full body shape estimate using our measurements-to-βs regressor and the SMPL function [24]. Probabilistic combination intuitively amounts to uncertainty-weighted averaging (Equation 5) - since our measurement distributions are able to better capture *local* shape uncertainty than independent Gaussians over SMPL βs [36], we obtain improved body shape estimation accuracy. This is quantitatively corroborated by shape metrics on the SSP-3D dataset [34], as well two private datasets of tape-measured humans, in an A-pose and in varying poses.

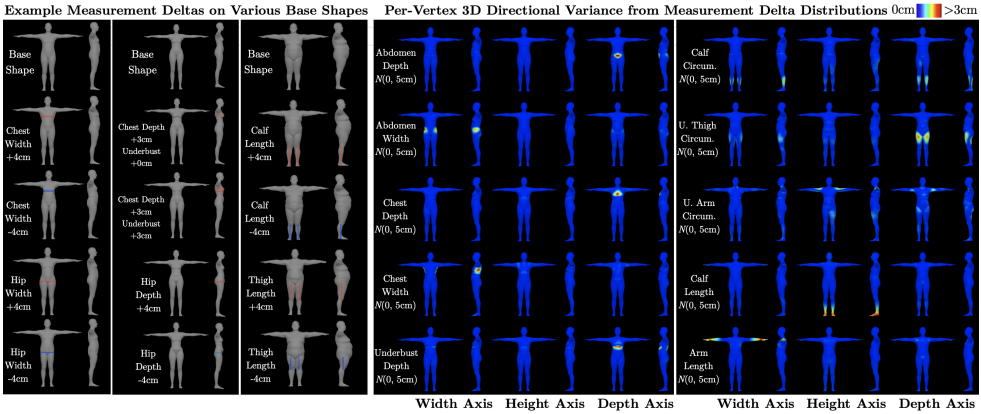

Figure 2: Left: Effect of various input measurement offsets applied to 3 different base body shapes. Measurement offsets are mapped to SMPL β offsets using the proposed measurements-to-βs linear regressor. Right: Transformation of Gaussian distributions over various measurement offsets to Gaussians over 3D vertex locations. Visualisation of 3D vertex variances along the width (x), height (y) and depth (z) axes aligned with the front-facing human body. Note that semantic measurement offsets result in local shape deformations, and distributions over these offsets result in localised vertex variance, representing uncertainty in each vertex's 3D location in the T-pose.

2 Related Work

This section reviews recent approaches to 3D human shape and pose estimation from images.

Monocular shape and pose estimators may be classified as optimisation-based or learning-based. Optimisation-based approaches fit a parametric 3D body model [14, 24, 27, 29] to 2D observations, such as 2D keypoints [9, 22, 29], silhouettes [22] or part segmentations [43] by minimising a suitable error function. They do not require expensive 3D-labelled training, but are sensitive to poor initialisations and inaccurate 2D observations. Learning-based approaches can be further split into model-free or model-based. Model-free methods train deep networks to directly predict human body meshes [6, 21, 25, 44], voxel occupancies [40] or implicit surface representations [32, 33] given an input image. Model-based methods [5, 9, 10, 15, 20, 26, 28, 38, 45] regress 3D body model parameters [24, 27, 29], which give a low-dimensional representation of a 3D human body. Learning-based methods yield impressive 3D pose estimates in-the-wild, but shape predictions are often inaccurate, due to the lack of shape diversity in training datasets. Some recent works improve shape estimates using synthetic training data [34, 35, 36, 37], which we adopt in our method.

Multi-image shape and pose estimators leverage the extra shape information present in videos [1, 2, 16, 19, 30, 39], as well as multi-view images [23, 37] of a subject in a fixed pose captured from multiple camera angles. In contrast, [36] propose to estimate body shape from a set of *unconstrained* images of a subject, by probabilistically-combining distributions over SMPL [24] shape parameters. We extend this work by predicting distributions over local body measurements instead of global shape parameters, and demonstrate that this improves shape estimation accuracy from sets of unconstrained images.

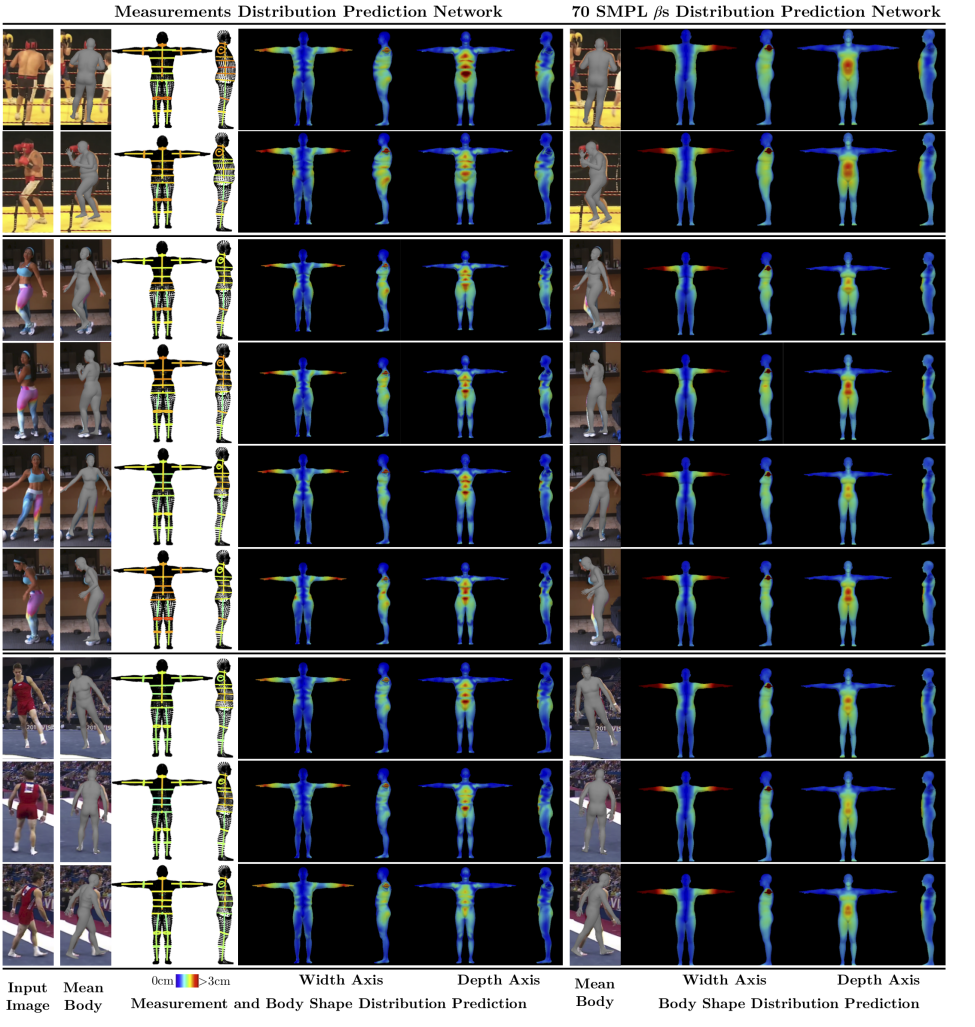

Figure 3: Comparing independent Gaussian measurement distributions and SMPL β distributions on images from SSP-3D [34]. Measurement distributions predictions (left) exhibit meaningful *local* shape uncertainty arising from varying camera angles, challenging poses and self-occlusions. For example, comparing rows 1 vs 2 shows that front/back-facing images result in larger predicted uncertainty (i.e. variance) for *depth* measurements (columns 4, 7, 8), while side-facing images result in greater uncertainty for *width* measurements (columns 3, 5, 6). This is reasonable as body depth is ambiguous from a front-on viewpoint while body width is ambiguous from the side. Moreover, in row 3 the subject's hips are obscured by their pose but the upper torso is visible, while the opposite is true in row 4 where hair occludes the torso. Accordingly, row 3 shows larger hip width uncertainty while row 4 shows larger torso width uncertainty (columns 3 and 6). In contrast, independent Gaussian SMPL β distributions (right), as proposed by [36], cannot model local shape uncertainty arising from ambiguous inputs, since βs control global deformations over the whole body surface. Global shape uncertainty is less useful for downstream probabilistic combination as it does not specify which local body parts are uncertain.

3 Method

This section provides a brief overview of SMPL [24], introduces our measurements-to-βs linear regressor and presents our three-stage pipeline for probabilistic human pose and body measurement estimation from multiple images of a subject.

SMPL [24] is a parametric 3D human body model. It provides a differentiable function that maps pose parameters $\boldsymbol{\theta}$, shape parameters $\boldsymbol{\beta}$ and global body rotation $\boldsymbol{\gamma}$ to a 3D vertex mesh $\boldsymbol{V} \in \mathbb{R}^{6890\times 3}$. $\boldsymbol{\theta}$ represents 3D joint rotations, relative to each joint's parent in the kinematic tree, in axis-angle form (i.e. $\boldsymbol{\theta} \in \mathbb{R}^{69}$ for 23 SMPL joints). Similarly, $\boldsymbol{\gamma} \in \mathbb{R}^3$ represents root joint rotation (i.e. global body orientation) in axis-angle form. The shape parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$ consists of coefficients quantifying the contribution of PCA shape-space basis vectors, $\{\boldsymbol{S}_i\}_{i=1}^{|\boldsymbol{\beta}|}$ where $\boldsymbol{S}_i \in \mathbb{R}^{6890\times 3}$, to the identity-dependent body shape. Specifically, shape-space basis vectors represent deformations from a template mesh $\boldsymbol{T} \in \mathbb{R}^{6890\times 3}$ over the full body surface. The identity-dependent (i.e. T-pose) 3D vertex mesh is then given by

$$\tilde{\boldsymbol{V}} = \sum_{i=1}^{|\boldsymbol{\beta}|} \beta_i \boldsymbol{S}_i + \boldsymbol{T} = \text{vec}^{-1}(\mathcal{S}\boldsymbol{\beta} + \boldsymbol{t}) \tag{1}$$

where $\boldsymbol{t} = \text{vec}(\boldsymbol{T}) \in \mathbb{R}^{20670}$ and $\mathcal{S} = [\text{vec}(\boldsymbol{S}_1), ..., \text{vec}(\boldsymbol{S}_{|\boldsymbol{\beta}|})] \in \mathbb{R}^{20670\times|\boldsymbol{\beta}|}$ represent shape-space bases and template vertices flattened with the vec() operation. $\text{vec}^{-1}()$ denotes the inverse, converting a vector back into a matrix containing 3D vertices.

Measurements-to-βs Linear Regressor. We learn a simple linear regressor from 23 body measurements to SMPL shape coefficients. Please refer to the supplementary material for a list of measurements used and details regarding the definition of measurements over a SMPL T-pose body, which is abstracted here as an operation $\boldsymbol{m} = \text{measure}(\boldsymbol{\beta})$ that outputs body measurements $\boldsymbol{m} \in \mathbb{R}^{23}$ given shape coefficients $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$. We aim to obtain a mapping from measurement *offsets* $\Delta\boldsymbol{m}$ to shape coefficient *offsets* $\Delta\boldsymbol{\beta}$, such that

$$\Delta\boldsymbol{\beta}^T = \Delta\boldsymbol{m}^T\boldsymbol{W} \tag{2}$$

where $\boldsymbol{W} \in \mathbb{R}^{23\times|\boldsymbol{\beta}|}$ is the weight matrix of the linear regressor. Then, each specific measurement of a given base body shape $\boldsymbol{\beta}$ (with measurements $\boldsymbol{m}$) can be offset by (i) setting the corresponding element of $\Delta\boldsymbol{m}$ to the desired value, (ii) obtaining $\Delta\boldsymbol{\beta}$ using Equation 2, and (iii) adding the shape offset to the base shape to yield a new body $\boldsymbol{\beta} + \Delta\boldsymbol{\beta}$ with measurements $\boldsymbol{m} + \Delta\boldsymbol{m}$. Several measurement offsets on varying base body shapes are visualised in Figure 2 (left). Note that the mean SMPL body is given by $\bar{\boldsymbol{\beta}} = \boldsymbol{0}$ with measurements $\bar{\boldsymbol{m}} = \text{measure}(\bar{\boldsymbol{\beta}})$. Thus, if the base body shape is assumed to be the mean SMPL body $\bar{\boldsymbol{\beta}} = \boldsymbol{0}$, coefficient offsets $\Delta\boldsymbol{\beta}$ are equivalent to the new shape coefficients themselves.

To learn the weight matrix $\boldsymbol{W}$, we first randomly sample a range of SMPL shape coefficients and stack them into a matrix $\boldsymbol{B} \in \mathbb{R}^{L\times|\boldsymbol{\beta}|}$, with $L = 10^6$ samples. Corresponding measurements are obtained as $\boldsymbol{M} = \text{measure}(\boldsymbol{B}) \in \mathbb{R}^{L\times 23}$. SMPL mean shape and measurements are subtracted to give $\Delta\boldsymbol{B} = \boldsymbol{B} - \bar{\boldsymbol{\beta}}$ and $\Delta\boldsymbol{M} = \boldsymbol{M} - \bar{\boldsymbol{m}}$. Then, $\boldsymbol{W}$, such that $\Delta\boldsymbol{B} = \Delta\boldsymbol{M}\boldsymbol{W}$, is estimated in a least squares sense using the pseudo-inverse $\boldsymbol{W} = (\Delta\boldsymbol{M}^T\Delta\boldsymbol{M})^{-1}\Delta\boldsymbol{M}^T\Delta\boldsymbol{B}$.

An independent Gaussian distribution over measurement offsets, $\mathcal{N}(\boldsymbol{\mu}_{\Delta m}, \text{diag}(\boldsymbol{\sigma}^2_{\Delta m}))$, can be transformed to a Gaussian distribution over shape coefficients, $\mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, and then over T-pose vertices $\mathcal{N}(\boldsymbol{\mu}_{\tilde{V}}, \boldsymbol{\Sigma}_{\tilde{V}})$ using linear transformations of Gaussians (note that the SMPL mean shape $\bar{\boldsymbol{\beta}} = \boldsymbol{0}$ is assumed as the base body shape):

$$\boldsymbol{\mu}_\beta = \boldsymbol{W}^T\boldsymbol{\mu}_{\Delta m},\ \boldsymbol{\Sigma}_\beta = \boldsymbol{W}^T\text{diag}(\boldsymbol{\sigma}^2_{\Delta m})\boldsymbol{W},\quad \boldsymbol{\mu}_{\tilde{V}} = \mathcal{S}\boldsymbol{\mu}_\beta + \boldsymbol{t},\ \boldsymbol{\Sigma}_{\tilde{V}} = \mathcal{S}\boldsymbol{\Sigma}_\beta\mathcal{S}^T \tag{3}$$

which follows from Equations 1 and 2. The diagonal terms of the covariance matrix $\mathbf{\Sigma}_{\tilde{V}}$ quantify the variance (i.e. uncertainty) in the 3D locations of T-pose vertices in the x, y and z directions (i.e. width, height and depth axis). Figure 2 (right) visualises these directional variances given different input measurement offset distributions.

Proxy representation computation. Given N RGB images $\{\boldsymbol{I}_n\}_{n=1}^N$ of a subject, we first compute edge-images and 2D joint heatmaps (see Figure 1), using Canny edge detection [5] and Detectron2 [42]. The edge-image and joint heatmaps of $\boldsymbol{I}_n \in \mathbb{R}^{H\times W\times 3}$ are stacked to form a proxy representation $\boldsymbol{X}_n \in \mathbb{R}^{H\times W\times (J+1)}$ (for J joints). We use this proxy representation as our input, instead of the RGB image, to decrease the domain gap between synthetic training images [34, 35, 36] and real test images.

Body measurements and pose distribution prediction. Next, we follow [36] and pass each $\boldsymbol{X}_n$ into a distribution prediction neural network (as shown in Figure 1). However, instead of predicting a distribution over SMPL shape coefficients, our network outputs the means and variances of independent Gaussian distributions over *measurement offsets* $\boldsymbol{\Delta m}$ (from the mean SMPL body measurements $\bar{\boldsymbol{m}}$), as well as pose parameters $\boldsymbol{\theta}$, conditioned on the inputs:

$$p(\boldsymbol{\theta}_n|\boldsymbol{X}_n) = \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{X}_n), \mathbf{\Sigma}_\theta(\boldsymbol{X}_n)), \;\; p(\boldsymbol{\Delta m}|\boldsymbol{X}_n) = \mathcal{N}(\boldsymbol{\mu}_{\Delta m}(\boldsymbol{X}_n), \mathbf{\Sigma}_{\Delta m}(\boldsymbol{X}_n)) \tag{4}$$

where $\mathbf{\Sigma}_\theta(\boldsymbol{X}_n) = \text{diag}(\boldsymbol{\sigma}^2_\theta(\boldsymbol{X}_n))$ and $\mathbf{\Sigma}_{\Delta m}(\boldsymbol{X}_n) = \text{diag}(\boldsymbol{\sigma}^2_{\Delta m}(\boldsymbol{X}_n))$. Specifically, $\boldsymbol{\sigma}^2_\theta(\boldsymbol{X}_n)$ and $\boldsymbol{\sigma}^2_{\Delta m}(\boldsymbol{X}_n)$ estimate the heteroscedastic aleatoric uncertainty [8, 17] in pose and measurement predictions, arising from ambiguities in the inputs due to varying camera views and poses (resulting in self-occlusion), or occluding objects. Furthermore, our network also outputs per-image deterministic estimates of weak-perspective camera parameters $\{\boldsymbol{c}_n\}_{n=1}^N$, representing scale and xy translation, and global body rotations $\{\boldsymbol{\gamma}_n\}_{n=1}^N$.

As an aside, our measurements-to-βs regressor, in theory, can be subsumed into the SMPL β distribution network of [36]. However, in practice, *independent* Gaussian distributions (with diagonal covariances) over SMPL βs cannot model local shape uncertainty. We would need to predict Gaussian distributions with full $|\boldsymbol{\beta}| \times |\boldsymbol{\beta}|$ positive semi-definite covariance matrices (see Equation 3). This is difficult compared to (i) learning the measurements-to-βs regressor separately, and (ii) predicting per-measurement variances $\boldsymbol{\sigma}^2_{\Delta m}(\boldsymbol{X}_n) \in \mathbb{R}^{23}$.

Multi-image measurement combination. Finally, we implement a similar probabilistic combination operation to [36], that combines the shape distributions from the individual images into a final, consistent body shape. However, instead of combining predicted distributions over SMPL shape coefficients, our combination is done in the body measurement space using the predicted measurement distributions:

$$\begin{aligned} p(\boldsymbol{\Delta m}|\{\boldsymbol{X}_n\}_{n=1}^N) &\propto \prod_{n=1}^N p(\boldsymbol{\Delta m}|\boldsymbol{X}_n) \propto \mathcal{N}(\boldsymbol{\Delta m}; \boldsymbol{\mu}_{\text{comb}}, \mathbf{\Sigma}_{\text{comb}}) \\ \mathbf{\Sigma}_{\text{comb}} &= \left(\sum_{n=1}^N \mathbf{\Sigma}^{-1}_{\Delta m}(\boldsymbol{X}_n)\right)^{-1}, \;\; \boldsymbol{\mu}_{\text{comb}} = \mathbf{\Sigma}_{\text{comb}} \left(\sum_{n=1}^N \mathbf{\Sigma}^{-1}_{\Delta m}(\boldsymbol{X}_n)\boldsymbol{\mu}_{\Delta m}(\boldsymbol{X}_n)\right). \end{aligned} \tag{5}$$

We observe that combining measurement distributions instead of shape coefficient distributions results in improved shape estimation accuracy (see Section 5), since distributions over measurements are able to predict local shape uncertainty due to varying camera views, poses and occlusions, unlike independent Gaussian distributions over global shape coefficients (see Figures 3 and 4). Please refer to [36] for more details on probabilistic shape combination.

At any stage of the inference pipeline, predicted measurement distributions or final combined measurement estimates can be easily converted into SMPL shape coefficient distributions/estimates using Equations 2 and 3.

Num. βs Used	Input Meas. Δ	Local Offset Evaluation Output Meas. Δ						Reconstruction Eval.	
		Ch. W.	Ch. D.	St. W.	St. D.	Ca. C.	Ca. L.	Meas. MAE	PVE-T
10	Ch. W. +50	+27.3	+5.0	+10.1	-4.1	+6.1	-1.1	**0.9**	**1.9**
	St. D. +50	-2.7	+7.8	+11.2	+29.9	+6.0	+5.8		
	Ca. L. +50	-3.1	+0.8	+5.3	+5.2	-2.5	+27.8		
70	Ch. W. +50	**+51.1**	+0.5	+0.5	+0.0	+0.3	+0.0	3.9	15.4
	St. D. +50	-0.3	-0.5	+0.0	**+49.8**	-3.6	+0.0		
	Ca. L. +50	+0.0	+0.2	+0.3	+1.9	+0.2	**+50.1**		
90	Ch. W. +50	+51.6	+0.4	+0.0	+0.5	+0.4	+0.0	6.2	23.9
	St. D. +50	+0.0	+0.1	+0.0	+54.3	-0.6	+0.0		
	Ca. L. +50	+0.1	-0.2	-0.4	-1.0	+1.7	+50.3		

Table 1: Local controllability and reconstruction ability of our measurements-to-βs regressor, using different numbers of SMPL βs. Local Offset Evaluation involves passing an input offset of +50mm for chest width, stomach depth and calf length in turn through the linear regressor, and computing the corresponding output measurement offsets, thereby quantifying the local controllability of our approach. Reconstruction Evaluation quantifies how well an SMPL body (represented by T-pose vertices) can be reconstructed from just its corresponding measurements, in terms of measurement error (Meas. MAE) and per-vertex error (PVE-T). These are computed by sampling 100,000 random input SMPL bodies, passing their measurements through the linear regressor and comparing the output SMPL bodies with the inputs. All numbers in mm. Abbreviations: Ch. = Chest, St. = Stomach, Ca. = Calf, W. = Width, D. = Depth, C. = Circumference, L. = Length.

Loss functions. At test-time, our measurement and pose prediction pipeline deals with sets of input images. However, training occurs using a dataset of single-image input-label pairs, denoted by $\{\boldsymbol{X}_k, \{\boldsymbol{\theta}_k, \boldsymbol{\Delta m}_k, \boldsymbol{\gamma}_k\}\}_{k=1}^K$, with K i.i.d training samples. Note that measurement offset labels $\boldsymbol{\Delta m}_k$ represent offsets from the mean SMPL body measurements $\bar{\boldsymbol{m}}$.

We train the distribution prediction network with a negative log-likelihood loss $\mathcal{L}_{\text{NLL}} = -\sum_{k=1}^K \Big(\log p(\boldsymbol{\theta}_k|\boldsymbol{X}_k) + \log p(\boldsymbol{\Delta m}_k|\boldsymbol{X}_k) \Big)$. We also apply the same 2D joints samples loss proposed in [36], as well as a mean-squared-error loss over global body rotation matrices.

4 Implementation Details

Network architecture. Our distribution prediction network consists of a ResNet-18 [11] convolutional encoder, followed by a 3 layer fully-connected network with 512 neurons in the two hidden layers and ELU activations [7], and 190 output neurons. Output variances are forced to be positive using an exponential activation function.

Synthetic training. We adopt the training frameworks presented in [28, 34, 36, 37], which entail on-the-fly generation of synthetic training inputs and corresponding SMPL body shape and pose labels during training. In short, for each training iteration, ground-truth body shapes are randomly sampled from a Gaussian distribution over the SMPL shape space, while ground-truth poses are obtained from the training sets of UP-3D [22], 3DPW [41] and H3.6M [13]. These are rendered into synthetic input proxy representations using the SMPL function, a light-weight renderer [31] and Canny edge detection [5]. Synthetic inputs are augmented using various occlusion and corruption transforms. Our method differs from past work in 2 main ways: (i) our measurements-to-βs regressor allows us to randomly sample measurement offsets to further augment the random SMPL shape samples, and (ii) we use

Method	Synthetic Meas. MAE-SC			SSP-3D Meas. MAE-SC			SSP-3D PVE-T-SC			3DPW	
	SI	NA	PC	SI	NA	PC	SI	NA	PC	MPJPE-SC	MPJPE-PA
10 βs Net	23.3	18.5	18.3	23.4	20.2	20.2	13.6	13.0	12.8	87.3	60.3
70 βs Net	23.2	18.6	18.2	23.7	20.1	20.0	13.6	12.9	12.9	92.9	63.8
Measure Net (Ours)	21.6	17.8	**16.4**	22.8	19.9	**19.5**	13.7	12.8	**12.4**	88.3	61.6
GraphCMR [21]	-	-	-	47.2	47.0	-	19.5	19.3	-	102.0	70.2
SPIN [20]	-	-	-	49.8	49.7	-	22.2	21.9	-	89.4	59.2
DaNet [35]	-	-	-	49.9	49.7	-	22.1	22.1	-	**82.4**	**54.8**
STRAPS [30]	-	-	-	24.5	21.0	-	15.9	14.4	-	99.0	66.8
Sengupta *et al.* [36]	-	-	-	24.4	20.6	20.4	15.2	13.6	13.3	90.9	61.0
VIBE* [19]	-	-	-	-	50.1	-	-	24.1	-	-	**51.9**

Table 2: Single-image (SI) and multi-image (NA/PC) body shape evaluation on synthetic ablation data and SSP-3D [34], and single-image pose evaluation on 3DPW [41]. Multi-image shape evaluation compares naive-averaging (NA) of shapes predicted from individual inputs against probabilistic shape combination (PC) (i.e. uncertainty-weighted averaging). The top half compares SMPL shape β distribution predictors against our measurement distribution predictor (both trained on the same synthetic data). The bottom half presents metrics from competing approaches. Note that our 10 βs Net is equivalent to [36], except we use improved edge-based training inputs [35]. All numbers in mm. *VIBE [19] uses video inputs.

body measurement labels to train our network using $\mathcal{L}_{\text{NLL}}$, and thus need to compute ground-truth measurements from the sampled ground-truth shape coefficients.

Training details. We use Adam [18] with a learning rate of 0.0001, batch size of 80 and train for 150 epochs, which takes 2 days on a 2080Ti GPU.

Evaluation datasets. SSP-3D [34] is used to evaluate body shape prediction accuracy in the wild. We report per-vertex Euclidean error in the T-pose after scale correction (PVE-T-SC) [34] and mean absolute measurement error after scale correction (Meas. MAE-SC), both in mm. In addition, we evaluate on two private datasets of tape-measured humans: "A-Pose Subjects" consists of front and side views of 8 subjects in an A-pose and "Varying-Pose Subjects" consists of 27 images of 4 subjects in a range of poses and camera views. We report chest, stomach and hip circumference measurement errors. The test set of 3DPW [41] is used to evaluate body pose accuracy, using mean-per-joint-position-error after scale correction (MPJPE-SC) and after Procrustes analysis (MPJPE-PA). Finally, we utilise a synthetic evaluation dataset for our ablation studies, which consists of 1000 synthetic subjects with randomly sampled body measurements. Each subject is posed using 4 SMPL poses sampled from Human3.6M [13] and global body rotations are set to face forward, backwards, left and right. Synthetic evaluation inputs are generated in the same way as our training inputs.

Please refer to the supplementary material for further details regarding synthetic data generation, training hyperparameters and example test images from the private shape evaluation datasets with tape-measured humans and the synthetic ablation dataset.

5 Experimental Results

This section discusses our ablation studies on the measurements-to-βs linear regressor, compares measurement versus shape coefficient distribution prediction and evaluates the performance of our method on real datasets against state-of-the-art approaches.

Measurements-to-βs linear regressor. Table 1 investigates the proposed measurements-to-βs regressor using $|\boldsymbol{\beta}| = 10$, 70 and 90 SMPL shape coefficients. From the local offset analysis (Table 1, left), it is clear that 10 shape PCA coefficients are not expressive enough

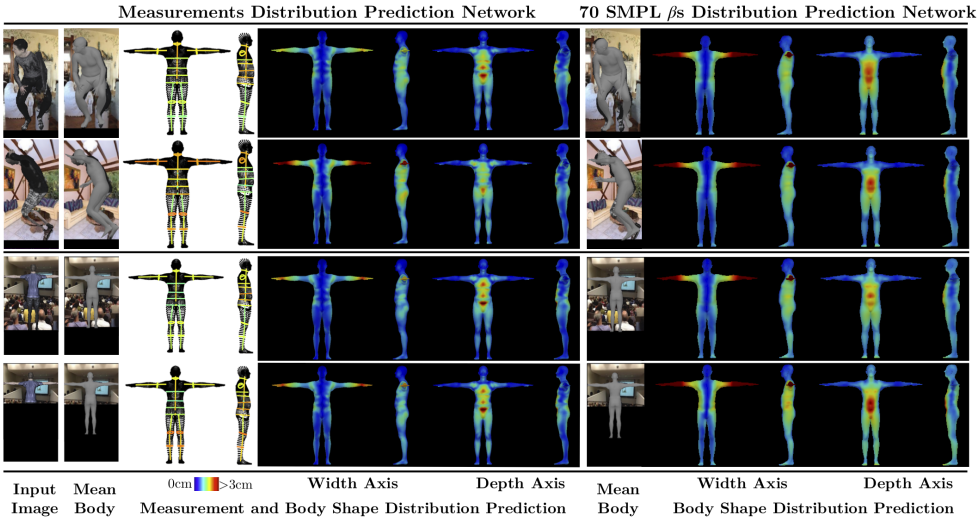

Figure 4: Comparing independent Gaussian measurement distributions and SMPL β distributions on synthetic images. Similar to Figure 3, measurement distributions (left) capture *local* shape uncertainty. For example, a front-on camera viewpoint (row 1) results in larger predicted depth measurement uncertainties (column 4) compared to a side-on viewpoint (row 2). Bent limbs result in higher limb length uncertainties (see arm lengths in row 1 vs row 2). Furthermore, when body parts are locally occluded, measurements *specific to the occluded part* have larger predicted uncertainties (see rows 3-4, columns 3-8). In contrast, independent Gaussian β distributions [36] (right) cannot model local shape uncertainty. For example, a locally occluded input results in an undesired increase of *global* shape uncertainty over the whole body surface (see rows 3-4, columns 10-13), which is less useful for downstream probabilistic combination as it does not specify which body parts are uncertain.

to locally control body shape, since an input offset for one measurement (*e.g.* +50mm chest width in row 1) results in significant output offsets for several other measurements. Increasing the number of PCA shape coefficients used, from 10 to 70, greatly improves the local controllability of the model, but further increasing to 90 does not provide much additional benefit. Qualitative examples of local offsets are given in Figure 2.

Conversely to the local offset analysis, using a larger number of shape coefficients increases reconstruction error (Table 1, right). While the greater expressiveness of more shape coefficients is beneficial for local offsets, it means that measurements alone do not contain enough information to reconstruct full body T-pose meshes. As a compromise between local controllability and reconstruction error, we use 70 SMPL shape coefficients.

Comparison between β distributions and measurement distributions. Table 2 (top) compares our proposed measurement distribution prediction network against SMPL β distribution predictors using 10 and 70 βs. Probabilistic combination (i.e. uncertainty-weighted averaging) using β distributions only results in marginal improvements over naive-averaging of βs, since the predicted β distributions are unable to quantify local uncertainty. In contrast, probabilistic *measurement* combination yields a significant improvement over naive-averaging of measurements, on both synthetic data and SSP-3D, since measurement distri-

Method	A-Pose Subjects									Varying-Pose Subjects								
	Chest			Stomach			Hip			Chest			Stomach			Hip		
	SI	NA	PC	SI	NA	PC	SI	NA	PC	SI	NA	PC	SI	NA	PC	SI	NA	PC
10 βs Net	63	54	52	61	54	52	54	38	35	83	72	76	47	30	32	42	27	27
70 βs Net	61	44	39	61	47	45	52	32	25	60	56	58	39	32	30	33	27	25
Measure Net (Ours)	52	37	**33**	37	32	**29**	37	29	**23**	63	53	**52**	43	29	**27**	37	23	**17**
SPIN [24]	130	127	-	117	114	-	125	124	-	57	56	-	60	60	-	74	73	-
STRAPS [38]	82	80	-	81	81	-	84	83	-	67	67	-	54	53	-	59	55	-
Sengupta *et al.* [36]	65	53	51	61	54	52	53	49	42	78	68	67	49	41	43	42	33	35

Table 3: Single-image (SI) and multi-image (NA/PC) body shape evaluation on two datasets of tape-measured humans, containing subjects in an A-pose and subjects in varying poses respectively. Multi-image shape evaluation compares naive-averaging (NA) of shapes predicted from individual inputs and probabilistic body shape combination (PC) (i.e. uncertainty-weighted averaging). The top half compares SMPL shape β distribution predictors against our proposed measurement distribution predictor. The bottom half presents metrics from competing state-of-the-art approaches. All numbers are circumference errors in mm. Note that our 10 βs Net is equivalent to [36], except we use improved edge-based training inputs [35].

butions capture local shape uncertainty due to varying poses/camera angles (which cause self-occlusion), as well as occluding objects (see Figures 3 and 4). Table 3 (top) further exhibits the benefits of probabilistic measurement combination on both A-pose humans and humans in varying poses. Note that combining β distributions (rows 1-2) can even result in *worse* measurement errors than naive-averaging for varying-pose subjects (showcasing challenging body poses), while measurement combination always improves errors.

Moreover, Table 2 shows a reduction in pose estimation accuracy for β distribution predictors when the number of βs used is increased from 10 to 70. We hypothesise that it is challenging for the network to learn to estimate distributions over $7\times$ more shape parameters, and pose accuracy suffers as a result. In contrast, predicting distributions over 23 local body measurements allows us to benefit from the increased expressiveness of 70 shape coefficients, without compromising pose.

Comparison with the state-of-the-art. Table 2 (bottom) presents shape and pose metrics from several approaches evaluated on SSP-3D and 3DPW. Our probabilistic measurement combination approach yields the best shape metrics on SSP-3D. In terms of pose metrics, it is competitive with approaches that do not any require 3D-labelled training images [34, 36]. Table 3 (bottom) also shows that measurement combination outperforms all other approaches in terms of measurement errors on both A-pose and varying-pose subjects.

6 Conclusion

In this work, we propose a locally controllable shape model by learning a linear mapping from semantic body measurements to SMPL [24] shape βs. This is motivated by the observation that distributions over SMPL shape βs are unable to meaningfully capture shape uncertainty associated with *locally*-occluded body parts, since the SMPL shape space represents *global* deformations over the whole body surface. Our measurements-to-βs regressor allows us to predict distributions over body measurements conditioned on input images. We demonstrate the value of the proposed procedure when predicting a body shape estimate from a set of images of a subject, where we achieve state-of-the-art identity-dependent body shape estimation accuracy on the SSP-3D [36] dataset and a private dataset of tape-measured humans, using probabilistic measurement combination.

References

[1] Thiemo Alldieck, Marc Kassubeck, Bastian Wandt, Bodo Rosenhahn, and Marcus Magnor. Optical flow-based 3d human motion estimation from monocular video. In Volker Roth and Thomas Vetter, editors, *Proceedings of the German Conference on Pattern Recognition (GCPR)*, pages 347–360, Sep 2017.

[2] Anurag* Arnab, Carl* Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] Benjamin Biggs, Sébastien Erhardt, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotny. 3D multibodies: Fitting sets of plausible 3D models to ambiguous image data. In *NeurIPS*, 2020.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, October 2016.

[5] John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 8(6):679–698, 1986.

[6] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision (ECCV)*, 2020.

[7] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus), 2016.

[8] Armen Der Kiureghian and Ove Dalager Ditlevsen. Aleatoric or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009. ISSN 0167-4730. doi: 10.1016/j.strusafe.2008.06.020.

[9] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Kosecka, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *European Conference on Computer Vision (ECCV)*, 2020.

[10] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36 (7):1325–1339, July 2014.

[14] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[15] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeuRIPS)*, 2017.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[19] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[21] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] Junbang Liang and Ming C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia*, volume 34, pages 248:1–248:16. ACM, 2015.

[25] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020.

[26] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2018.

[27] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[28] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019.

[31] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[32] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[33] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

[34] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *Proceedings of the British Machine Vision Conference (BMVC)*, September 2020.

[35] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[36] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3d human shape and pose estimation from multiple unconstrained images in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[37] B. M. Smith, V. Chari, A. Agrawal, J. M. Rehg, and R. Sever. Towards accurate 3d human body reconstruction from silhouettes. In *International Conference on 3D Vision (3DV)*, pages 279–288, 2019. doi: 10.1109/3DV.2019.00039.

[38] Vince J. K. Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human shape and pose prediction. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[39] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5236–5246. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7108-self-supervised-learning-of-motion-capture.pdf.

[40] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[41] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[42] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[43] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[45] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Danet: Decompose-and-aggregate network for 3D human shape and pose estimation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 935–944, 2019.