Probabilistic 3D Human Shape and Pose Estimation from Multiple Unconstrained Images in the Wild

Akash Sengupta
University of Cambridge
as2562@cam.ac.uk

Ignas Budvytis
University of Cambridge
ib255@cam.ac.uk

Roberto Cipolla
University of Cambridge
rc10001@cam.ac.uk

Abstract

This paper addresses the problem of 3D human body shape and pose estimation from RGB images. Recent progress in this field has focused on single images, video or multi-view images as inputs. In contrast, we propose a new task: shape and pose estimation from a group of multiple images of a human subject, without constraints on subject pose, camera viewpoint or background conditions between images in the group. Our solution to this task predicts distributions over SMPL body shape and pose parameters conditioned on the input images in the group. We probabilistically combine predicted body shape distributions from each image to obtain a final multi-image shape prediction. We show that the additional body shape information present in multi-image input groups improves 3D human shape estimation metrics compared to single-image inputs on the SSP-3D dataset and a private dataset of tape-measured humans. In addition, predicting distributions over 3D bodies allows us to quantify pose prediction uncertainty, which is useful when faced with challenging input images with significant occlusion. Our method demonstrates meaningful pose uncertainty on the 3DPW dataset and is competitive with the state-of-the-art in terms of pose estimation metrics.

1. Introduction

3D human body shape and pose estimation from RGB images is a challenging problem with potential applications in augmented and virtual reality, healthcare and fitness technology and virtual retail. Recent solutions have focused on three types of inputs: i) single images [7, 48, 19, 27, 28, 57, 36, 45, 38, 41, 50], ii) video [26, 20, 47, 40, 16] with temporal constraints on pose, camera viewpoint and background conditions and iii) multi-view images [32, 46] with a fixed subject pose captured from multiple viewpoints. In contrast, we aim to estimate 3D body shape and pose from a group of images of the same human subject without any constraints on the subject's pose, camera viewpoint or back-
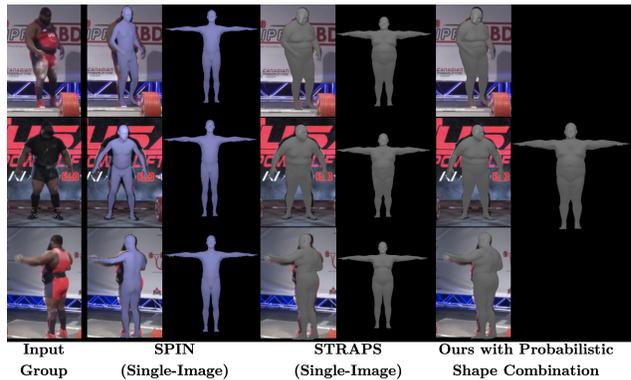

Figure 1: Example shape and pose predictions from a group of input images. Probabilistic shape combination results in a more accurate body shape estimate than both individual single-image predictions (visualised here from SPIN [27] and STRAPS [45]) and naively-averaged single-image predictions, as our experiments show in Section 5.

ground conditions between the images, as illustrated in Figure 1. This task is motivated by the intuition that multiple images of the same subject should contain additional visual information about their body shape compared to a single image, regardless of whether the subject's pose or surrounding environment change between images. A suitable shape and pose estimator should leverage this information to improve shape prediction accuracy over single-image methods.

We present a probabilistic body shape and pose estimation method from a group of unconstrained images of the same subject. Inference occurs in three stages (see Figure 2). First, we predict a proxy representation from each input image in the group, consisting of the subject's silhouette and 2D joint location heatmaps, using off-the-shelf segmentation and 2D keypoint detection CNNs [14, 13, 25, 53]. Then, each proxy representation is passed through a 3D distribution prediction network that outputs a probability distribution over SMPL [33] body shape and pose parameters conditioned on the input representation. Lastly, body shape distributions from each input image are prob-

abilistically combined to procure a final shape prediction. This yields a better estimate of the subject's body shape than current single-image body shape and pose estimators [19, 27, 28, 57, 45], which may be inaccurate or inconsistent, as shown in Figure 1.

Moreover, most single-image body model parameter regressors [19, 27, 38, 57, 45, 54, 10] do not consider the uncertainty associated with each pose parameter estimate. If certain body parts are occluded or out-of-frame in the input image, the estimator can only guess about the pose parameters corresponding to these body parts. Such situations further motivate our approach of predicting a distribution over body pose, since the variance of the distribution quantifies the uncertainty associated with each pose parameter prediction, as shown in Figures 3 and 4.

Training body model parameter regressors to accurately predict body shape is challenging due to the lack of suitable training datasets of in-the-wild images paired with *accurate and diverse* body shape labels. Collecting such data is practically difficult, particularly for our proposed task of shape estimation from a group of unconstrained images. Recent works [45, 46, 51] propose using synthetic input-label pairs to overcome the lack of suitable training datasets. We adopt the same synthetic training approach as STRAPS [45] to train our 3D distribution prediction network, but extend the data augmentations used to bridge the gap between synthetic and real inputs. In particular, our synthetic training data better models occluded and out-of-frame body parts in silhouettes and joints such that the domain gap to real occluded data is smaller. This allows our method to estimate pose prediction uncertainty and also results in improved single-input pose prediction metrics on challenging evaluation datasets, such as 3DPW [52].

In summary, our main contributions are as follows:

- We propose a novel task: predicting body shape from a group of images of the same human subject, without imposing any constraints on subject pose, camera viewpoint or backgrounds between the images.

- We present a solution to the proposed task which predicts a distribution over 3D human body shape and pose parameters conditioned on the input images in the group. Body shape distributions from each image are probabilistically combined to yield a final body shape estimate which leverages multi-image shape information, resulting in a more accurate body shape estimate compared to single-input methods.

- To the best of our knowledge, our method is the first to output uncertainties alongside associated SMPL [33] shape and pose parameter predictions, which are shown to be useful when input images contain occluded or out-of-frame body parts.

- We extend the synthetic training framework introduced by [45] to better model occlusion and missing body parts, allowing our synthetically-trained distribution prediction neural network to yield better 3D shape and pose metrics.

2. Related Work

This section discusses recent approaches to 3D human shape and pose estimation from single images, multi-view images and video.

Single-image shape and pose estimation methods can be classified into 2 categories: optimisation-based and learning-based. Optimisation-based methods fit a parametric 3D body model [33, 39, 2, 18] to 2D observations (e.g. 2D joints [7, 39], surface landmarks [30], silhouettes [30] or part segmentations [55]) via optimisation. They can accurately estimate 3D poses without requiring expensive 3D-labelled datasets, however they are susceptible to poor initialisation and tend to be slow at test-time.

Learning-based methods can be further classified as model-free or model-based. Model-free approaches directly predict a 3D body representation from an image, such as a voxel occupancy grid [50], vertex mesh [28, 36, 56] or implicit surface representation [44]. Model-based approaches [19, 57, 38, 41, 48, 10, 12] predict the parameters of a 3D body model [33, 39, 2, 18], which provides a useful prior over human body shape. Several methods [19, 28, 12, 54] overcome the scarcity of in-the-wild 3D-labelled training data by incorporating weak supervision with datasets of labelled 2D keypoints. [27] extends this further by integrating optimisation into the model training loop to lift 2D keypoint labels to self-improving 3D pose and shape labels. Such approaches predict impressive 3D poses but fail to predict accurate body shapes (particularly for non-average humans) since 2D keypoints do not densely inform shape. Recently, [45] used random synthetic training data to overcome data scarcity and demonstrated improved shape predictions.

Video shape and pose estimation methods may be classified similarly to their single-image counterparts. Optimisation-based video methods [3, 16, 1] extend single-image optimisation over time, while learning-based video methods [20, 26, 47, 49, 40] modify single-image predictors to take sequences of frames as inputs. However, video inputs allow these methods to enforce consistent body shapes and smooth motions across frames, e.g. using motion discriminators [11], optical flow [1, 49], or texture consistency [40]. Learning-based video methods also overcome 3D data scarcity by incorporating weak 2D supervision [26, 20], or with self-supervision enforcing visual consistency between frames [49, 40, 1]. Nevertheless, current methods are unable to predict accurate body shapes, particularly for non-average humans.

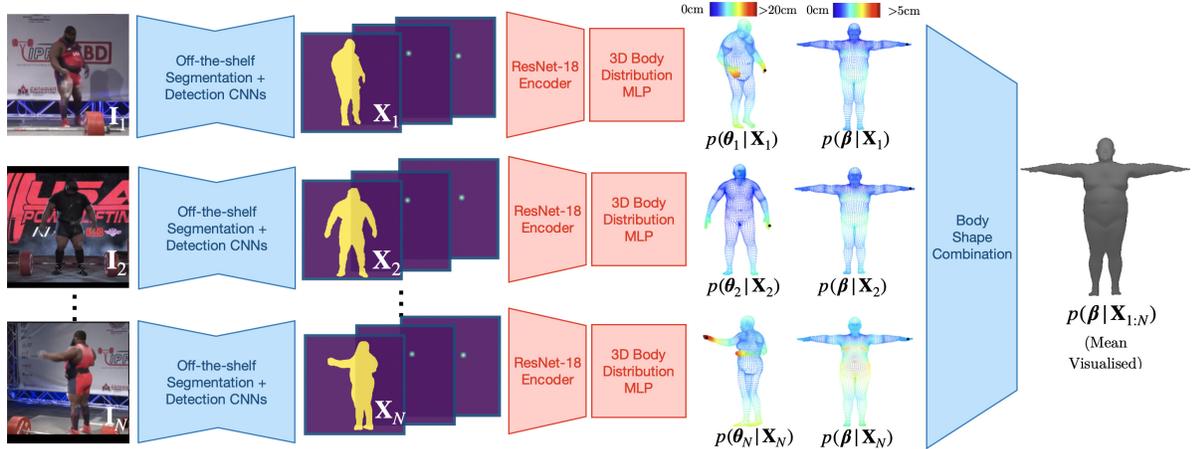**Multi-view shape and pose estimation.** [32] extends the

Figure 2: Overview of our shape and pose distribution prediction network. Each image $\mathbf{I}_n$ in the input group is converted into a silhouette and joint proxy representation $\mathbf{X}_n$, which is passed through a distribution prediction network to obtain multivariate distributions over SMPL [33] shape and pose parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_n$, conditioned on the input. Shape distributions from each individual input are probabilistically combined to form a multi-input shape distribution. The encoder and distribution MLP are trained using randomly-generated synthetic data [45]. The per-vertex uncertainty visualisations (in cm) are obtained by sampling SMPL parameters from the predicted distributions, computing the SMPL vertex mesh for each sample and determining the average Euclidean distance from the mean for each vertex. Black dots indicate left hands.

iterative regressor of [19] to predict body model parameters from multiple input images of the same subject in a fixed pose, captured from varying camera angles. They use synthetic data to overcome data scarcity, resulting in more accurate body shape estimates, particularly under clothing. [46] uses synthetic data to learn to predict body model parameters from A-pose silhouettes.

Contrary to the above approaches, our method estimates shape and pose from a group of images without *any* temporal or absolute constraints on the subject's pose, camera viewpoint or background between images.

Shape and pose distribution estimation. While substantial progress has been made in predicting probability distributions using neural networks [4, 37, 6, 22, 43, 42, 31, 35], prediction of distributions over 3D human shape *and* pose is still under-explored. Recently, [36] used lixel-based 1D heatmaps to quantify uncertainty in predicted 3D human mesh vertex locations. [5] predicted a categorical distribution over multiple SMPL hypotheses given an ambiguous image. In contrast, we aim to explicitly output separable uncertainties per predicted pose and shape parameter, since the shape uncertainties, specifically, are used for shape prediction from multiple unconstrained images.

3. Method

This section provides a brief overview of the SMPL parametric human body model [33], presents our three-stage method for probabilistic body shape and pose estimation from a group of unconstrained images of the same human subject (illustrated in Figure 2) and finally discusses the

synthetic training framework and loss functions used.

3.1. SMPL model

SMPL [33] provides a differentiable function $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ which takes pose parameters $\boldsymbol{\theta}$, global rotation $\boldsymbol{\gamma}$ and identity-dependent body shape parameters $\boldsymbol{\beta}$ as inputs and outputs a vertex mesh $\mathbf{V} \in \mathbb{R}^{6890 \times 3}$. $\boldsymbol{\theta} \in \mathbb{R}^{69}$ and $\boldsymbol{\gamma} \in \mathbb{R}^3$ represent axis-angle rotation vectors for 23 SMPL body joints and the root joint respectively. $\boldsymbol{\beta} \in \mathbb{R}^{10}$ represents coefficients of a PCA body shape basis. Given the vertex mesh $\mathbf{V}$, 3D joint locations may be obtained using a linear regressor, $\mathbf{J}^{3D} = \mathcal{J}\mathbf{V}$ where $\mathcal{J} \in \mathbb{R}^{L \times 6890}$ is a regression matrix for L joints of interest.

3.2. Proxy representation computation

Given a group of N RGB input images $\{\mathbf{I}_n\}_{n=1}^N$ of the same subject, we first compute proxy representations $\{\mathbf{X}_n\}_{n=1}^N$. DensePose [13] is used to obtain body part segmentations, which are converted into silhouettes. Keypoint-RCNN from Detectron2 [14, 53] is used to obtain 2D joint locations, which are converted into Gaussian heatmaps, and associated confidence scores. Heatmaps corresponding to joint detections with confidence scores less than a threshold $t = 0.025$ are set to 0. Thresholding is essential for modelling uncertainty in 3D pose predictions as it typically removes invisible 2D joints from the input representations. The predicted silhouette and joint heatmaps from each image are stacked along the channel dimension to form each proxy representation $\mathbf{X}_n \in \mathbb{R}^{H \times W \times (L+1)}$.

The use of silhouette and joint heatmap representations

as inputs instead of RGB images is inspired by [41, 45] and allows us to train our distribution prediction network using a simple synthetic training framework (see Section 3.5), overcoming the lack of shape diversity in current datasets. We follow [45] and use simple silhouettes and 2D joint heatmaps as our proxy representation, instead of more complex alternatives (e.g. part segmentations or IUV maps), since this leads to a smaller synthetic-to-real domain gap which is more readily bridged by data augmentation [45].

3.3. Body shape and pose distribution prediction

We aim to estimate probability distributions [37] over SMPL pose parameters $\{\boldsymbol{\theta}_n\}_{n=1}^N$ (which are free to change between inputs) and the subject's identity-dependent shape $\boldsymbol{\beta}$, both conditional upon $\{\mathbf{X}_n\}_{n=1}^N$. We assume simple multivariate Gaussian distributions

$$p(\boldsymbol{\theta}_n|\mathbf{X}_n) = \mathcal{N}(\boldsymbol{\theta}_n; \boldsymbol{\mu}_\theta(\mathbf{X}_n), \boldsymbol{\Sigma}_\theta(\mathbf{X}_n))$$
$$p(\boldsymbol{\beta}|\mathbf{X}_n) = \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_\beta(\mathbf{X}_n), \boldsymbol{\Sigma}_\beta(\mathbf{X}_n)). \tag{1}$$

Covariance matrices are constrained to be diagonal, i.e. $\boldsymbol{\Sigma}_\theta(\mathbf{X}_n) = \mathrm{diag}(\boldsymbol{\sigma}_\theta^2(\mathbf{X}_n))$ and $\boldsymbol{\Sigma}_\beta(\mathbf{X}_n) = \mathrm{diag}(\boldsymbol{\sigma}_\beta^2(\mathbf{X}_n))$. Formally, $\boldsymbol{\sigma}_\theta^2(\mathbf{X}_n)$ and $\boldsymbol{\sigma}_\beta^2(\mathbf{X}_n)$ represent estimates of the heteroscedastic aleatoric uncertainty [9, 22] in the SMPL parameters explaining the input observations $\mathbf{X}_n$, which arises particularly due to occlusion.

We also predict deterministic estimates of the global rotations $\{\boldsymbol{\gamma}_n\}_{n=1}^N$ and weak-perspective camera parameters $\{\mathbf{c}_n\}_{n=1}^N$, where $\mathbf{c}_n = [s_n, t_n^x, t_n^y]$ representing scale and xy translation respectively. Global rotation and camera parameters are unconstrained across images.

Hence, we require a function mapping each input proxy representation $\mathbf{X}_n$ to the desired set of outputs $\mathbf{Y}(\mathbf{X}_n) = \{\boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\beta, \boldsymbol{\sigma}_\theta^2, \boldsymbol{\sigma}_\beta^2, \boldsymbol{\gamma}_n, \mathbf{c}_n\}$. This function is represented using a deep neural network f with learnable weights $\mathbf{W}$:

$$\mathbf{Y} = f(\mathbf{X}_n; \mathbf{W}). \tag{2}$$

f consists of a convolutional encoder for feature extraction followed by a simple multi-layer perceptron that predicts the set of outputs $\mathbf{Y}$, as illustrated in Figure 2. The network training procedure is detailed in Section 3.5.

3.4. Body shape combination

We combine the conditional body shape distributions output by f given each individual input, $p(\boldsymbol{\beta}|\mathbf{X}_n)$ for $n = 1, ..., N$, into a final distribution $p(\boldsymbol{\beta}|\{\mathbf{X}_n\}_{n=1}^N)$ that aggregates shape information across the input group. Formally,

$$p(\boldsymbol{\beta}|\{\mathbf{X}_n\}_{n=1}^N) \propto \prod_{n=1}^N p(\boldsymbol{\beta}|\mathbf{X}_n) \tag{3}$$

which follows from the conditional independence assumption $(\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j)|\boldsymbol{\beta}$ for $i, j \in \{1, ..., N\}$ and $i \neq j$. This is justifiable since we do not impose any relationship between
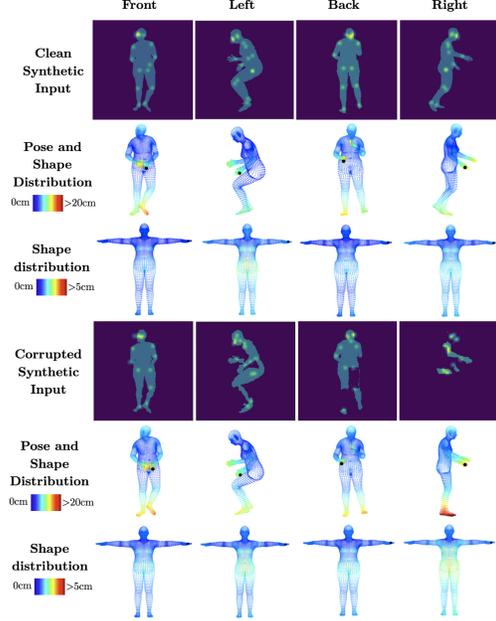

Figure 3: Clean and corrupted versions of an example group of inputs from our synthetic evaluation dataset, and corresponding (single-image) shape and pose distribution predictions. Black dots indicate left hands.

the subject's pose or camera viewpoint across inputs - only the body shape is fixed. Further details are in the supp. material. Since the product of Gaussians is an un-normalised Gaussian, $p(\boldsymbol{\beta}|\{\mathbf{X}_n\}_{n=1}^N) \propto \mathcal{N}(\boldsymbol{\beta}; \mathbf{m}, \mathbf{S})$ where

$$\mathbf{S} = \left(\sum_{n=1}^N \boldsymbol{\Sigma}_\beta^{-1}(\mathbf{X}_n) \right)^{-1}$$
$$\mathbf{m} = \mathbf{S} \left(\sum_{n=1}^N \boldsymbol{\Sigma}_\beta^{-1}(\mathbf{X}_n) \boldsymbol{\mu}_\beta(\mathbf{X}_n) \right). \tag{4}$$

The combined mean $\mathbf{m}$ is a final point estimate of the subject's body shape from the input group $\{\mathbf{X}_n\}_{n=1}^N$.

3.5. Network training

Loss functions. While inference occurs using a group of inputs, the model is trained with a dataset of *individual* input-label pairs, $\{\mathbf{X}_k, \{\boldsymbol{\theta}_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k\}\}_{k=1}^K$, with K i.i.d training samples. The negative log-likelihood is given by

$$\mathcal{L}_{\mathrm{NLL}} = - \sum_{k=1}^K \left(\log p(\boldsymbol{\theta}_k|\mathbf{X}_k) + \log p(\boldsymbol{\beta}_k|\mathbf{X}_k) \right)$$
$$\propto \sum_{k=1}^K \left(\sum_{i=1}^{69} \log(2\pi\sigma_{\theta_i}^2) + \frac{(\theta_{k_i} - \mu_{\theta_i})^2}{\sigma_{\theta_i}^2} \right.$$
$$\left. + \sum_{j=1}^{10} \log(2\pi\sigma_{\beta_j}^2) + \frac{(\beta_{k_j} - \mu_{\beta_j})^2}{\sigma_{\beta_j}^2} \right) \tag{5}$$

Input Image and Segmentation **Mean 3D Body Prediction** **Pose and Shape Distribution**

0cm ▮▮▮ >20cm

Figure 4: Predictions on single images from 3DPW [52]. 3D locations of invisible parts are uncertain due to large predicted variances for the corresponding pose parameters.

where μ_{θ_i}, μ_{β_i}, $\sigma^2_{\theta_i}$ and $\sigma^2_{\beta_i}$ represent elements of the predicted SMPL mean and variance vectors $\boldsymbol{\mu}_\theta(\mathbf{X}_k, \mathbf{W})$, $\boldsymbol{\mu}_\beta(\mathbf{X}_k, \mathbf{W})$, $\boldsymbol{\sigma}^2_\theta(\mathbf{X}_k, \mathbf{W})$ and $\boldsymbol{\sigma}^2_\beta(\mathbf{X}_k, \mathbf{W})$, which are output by the neural network f with weights $\mathbf{W}$. We maximise the log-likelihood of the model w.r.t $\mathbf{W}$ by minimising the loss function $\mathcal{L}_{\text{NLL}}$. Intuitively, each squared error term in Eqn. 5 is adaptively-weighted by the corresponding predicted variance [22]. This mitigates the ill-posed nature of a naive squared error loss on SMPL parameters when training inputs are occluded, since the network learns to predict large variances for the parameters corresponding to invisible body parts, thus down-weighting the respective squared error terms. Furthermore, adaptive weighting means that our network is able to train stably without additional "global" losses on 3D vertices or 3D joints, as is common in most other recent methods [45, 41, 27, 19, 57].

Our network also predicts deterministic estimates of the global rotations $\boldsymbol{\gamma}_k$. Predictions $\hat{\boldsymbol{\gamma}}_k$ are supervised by

$$\mathcal{L}_{\text{glob}} = \sum_{k=1}^{K} \|\mathbf{R}(\boldsymbol{\gamma}_k) - \mathbf{R}(\hat{\boldsymbol{\gamma}}_k)\|_F^2. \tag{6}$$

Input Groups **SPIN** **STRAPS** **Ours (Single-Image)** 0cm ▮▮▮ >5cm **Ours (Combined)**
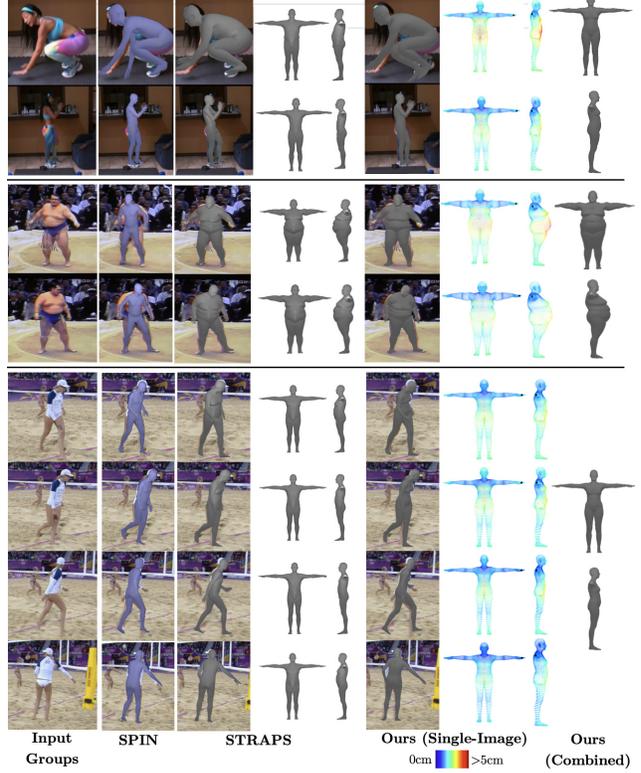
Figure 5: Example predictions on groups of images from SSP-3D [45]. Single-image predictions from SPIN [27], STRAPS [45] and our method are visualised, along with probabilistically combined body shapes from our method.

$\mathbf{R}(\boldsymbol{\gamma}) \in SO(3)$ is the rotation matrix corresponding to $\boldsymbol{\gamma}$.

Finally, our network estimates weak-perspective camera parameters $\mathbf{c}_k = [s_k, t^x_k, t^y_k]$, which are supervised using a 2D joint reprojection loss. Target 2D joint coordinates $\mathbf{J}_k \in \mathbb{R}^{L \times 2}$ are computed from $\{\boldsymbol{\theta}_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k\}$ during synthetic data generation (see Section 3.5). Predicted 2D joint coordinates are obtained by first differentiably sampling $\hat{\boldsymbol{\theta}}^i_k \sim p(\boldsymbol{\theta}_k|\mathbf{X}_k)$ and $\hat{\boldsymbol{\beta}}^i_k \sim p(\boldsymbol{\beta}_k|\mathbf{X}_k)$ using the reparameterisation trick [24], for $i = 1, ..., B$ samples. These are converted into 2D joint samples using the SMPL model and weak-perspective projection

$$\hat{\mathbf{J}}^i_k = s_k \Pi(\mathcal{J}\mathcal{M}(\hat{\boldsymbol{\theta}}^i_k, \hat{\boldsymbol{\beta}}^i_k, \hat{\boldsymbol{\gamma}}_k)) + [t^x_k, t^y_k] \tag{7}$$

where $\Pi()$ represents an orthographic projection. A squared error reprojection loss is imposed between the predicted 2D joint samples and the target 2D joints

$$\mathcal{L}_{\text{2D}} = \sum_{k=1}^{K} \sum_{i=1}^{B} \|\boldsymbol{\omega}_k(\mathbf{J}_k - \hat{\mathbf{J}}^i_k)\|_2^2 \tag{8}$$

where $\boldsymbol{\omega}_k \in \{0, 1\}^L$ denote the visibilities of the target joints (1 if visible, 0 otherwise), which are computed during synthetic data generation. We apply a reprojection loss

on *samples* from the predicted body shape and pose distributions, instead of only on the means of the distributions, because *any* 3D body sampled from the distributions must match the 2D joint locations present in the input $\mathbf{X}_k$.

Our overall loss function is given by $\mathcal{L} = \mathcal{L}_{\text{NLL}} + \lambda_{\text{glob}}\mathcal{L}_{\text{glob}} + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}}$ where $\lambda_{\text{glob}}, \lambda_{\text{2D}}$ are weighting terms.
Synthetic data generation. To train our distribution prediction network using the proposed losses, we require training data consisting of input proxy representations paired with target SMPL shape, pose and global rotation parameters, $\{\mathbf{X}_k, \{\boldsymbol{\theta}_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k\}\}_{k=1}^{K}$. We employ a similar synthetic training data generation process as [45]. In short, within each iteration of the training loop, $\boldsymbol{\theta}_k$ and $\boldsymbol{\gamma}_k$ are sampled from any suitable dataset with SMPL pose parameters [17, 30, 34, 52], while $\boldsymbol{\beta}_k$ are randomly sampled from a suitably high-variance Gaussian distribution to ensure body shape diversity. These are converted into synthetic silhouette and joint heatmap representations $\mathbf{X}_k$ and target 2D joint coordinates $\mathbf{J}_k$ using the SMPL model, a renderer [21] and randomly sampled perspective camera parameters. The clean synthetic inputs are corrupted to model the failure modes of the off-the-shelf detection and segmentation CNNs used at test-time, such as noisy keypoint locations, and occluded silhouettes. Examples are given in Figure 3.

We improve the data generation process of [45] in two ways. First, we significantly increase the severity of the occlusion and cropping augmentations to match the occlusions seen in challenging test datasets such as 3DPW (illustrated in Figure 4, first row). Second, we explicitly compute a joint visibility vector $\boldsymbol{\omega}_k$ (1 if visible, 0 otherwise) for each $\mathbf{J}_k$ and set the heatmaps corresponding to invisible joints to 0, unlike [45]. This is necessary for our distribution prediction network to learn to be uncertain about pose parameters corresponding to invisible body parts.

4. Implementation Details

Network Architecture. We use a ResNet-18 [15] encoder followed by a multi-layer perceptron (MLP) to predict SMPL parameter distributions. The MLP is comprised of one hidden layer with 512 neurons and ELU [8] activation and one output layer with 164 neurons, which predicts the set of outputs $\mathbf{Y}$. Predicted variances are forced to be positive using an exponential activation function.
Training dataset. Synthetic training data is generated by sampling SMPL pose parameters from the training sets of UP-3D [30], 3DPW [52], and Human3.6M [17] (Subjects 1, 5, 6, 7, 8).
Training details. We use Adam [23] with a learning rate of 1e-4 and a batch size of 120, and train for 100 epochs, which takes 1.5 days on a 2080Ti GPU. Inference runs at ~4fps, 90% of which is silhouette and joint prediction [13, 53].
Evaluation datasets. We use the test set of 3DPW to evaluate pose prediction accuracy. It consists of 35515 RGB im-

Input groups	Method	Synthetic Clean PVE-T-SC	Synthetic Corrupted PVE-T-SC
Single-Input	Ours	14.4	15.1
Quadruplets Front + L/R Side + Back	Ours + Mean	13.1	13.3
	Ours + PC	**13.0**	**12.8**
Pairs Front + L Side or Back + R Side	Ours + PC	13.5	13.4
Pairs Front + Back or L + R Side	Ours + PC	13.6	13.8

Table 1: PVE-T-SC (mm) results on synthetic data (see Figure 3) investigating: i) probabilistic shape combination (PC) versus simple averaging (Mean), ii) effect of increasing input group size from 1 to 2 (Pairs) to 4 (Quadruplets) and iii) effect of global rotation variation within pairs of inputs.

Max. input group size	Method	SSP-3D PVE-T-SC
1	Ours	15.2
2	Ours + PC	13.9
3	Ours + PC	13.6
4	Ours + PC	13.5
5	Ours + Mean	13.6
	Ours + PC	**13.3**

Table 2: PVE-T-SC (mm) results on SSP-3D [45] comparing i) probabilistic shape combination (PC) versus simple averaging (Mean) and ii) effect of increasing input group size from 1 to 5.

ages of 7 subjects with paired ground-truth SMPL parameters. We report mean per joint position error after scale correction (MPJPE-SC) and after Procrustes analysis (MPJPE-PA). We use the scale correction technique introduced in [45] to combat the ambiguity between subject scale and distance from camera. MPJPE-SC measures 3D joint error up to scale and MPJPE-PA measures 3D joint error up to scale and global rotation. We also report scale-corrected per-vertex Euclidean error in a T-pose (PVE-T-SC) on the SSP-3D dataset [45] to evaluate identity-dependent shape prediction accuracy. SSP-3D consists of 311 images of 62 subjects and pseudo-ground-truth SMPL parameters.

In addition, we evaluate body shape prediction accuracy on a private dataset consisting of 6 subjects (4 male, 2 female) with 4 RGB images of each and ground-truth body measurements obtained using a tape measure or full 3D body scanning technology. The subjects' body poses, clothing, surrounding environments and camera viewpoints vary between images. Example images are in the supplementary material.

Finally, we create a synthetic dataset for our experimen-

Max. input group size	Method	SSP-3D PVE-T-SC
1	HMR [19]	22.9
	GraphCMR [28]	19.5
	SPIN [27]	22.2
	DaNet [57]	22.1
	STRAPS [45]	15.9
	Ours	**15.2**
5	HMR [19] + Mean	22.9
	GraphCMR [28] + Mean	19.3
	SPIN [27] + Mean	21.9
	DaNet [57] + Mean	22.1
	STRAPS [45] + Mean	14.4
	Ours + Mean	13.6
	Ours + PC	**13.3**
Video	VIBE [26]	24.1

Table 3: Comparison with the state-of-the-art in terms of PVE-T-SC (mm) on SSP-3D [45]. Our method surpasses the state-of-the-art when using single-image inputs. Probabilistic shape combination (PC) outperforms simple averaging of predictions from other methods when using groups of up to 5 images, as well as video predictions from [26].

Method	3DPW MPJPE-SC	MPJPE-PA
HMR [19]	102.8	71.5
GraphCMR [28]	102.0	70.2
SPIN [27]	89.4	59.2
I2L-MeshNet [36]	-	57.7
Biggs et al. [5]	-	55.6
DaNet [57]	**82.4**	**54.8**
HMR (unpaired) [19]	126.3	92.0
Kundu et al. [29]	-	89.8
STRAPS [45]	99.0	66.8
Ours	90.9	61.0

Table 4: Comparison with the state-of-the-art in terms of MPJPE-SC and MPJPE-PA (both mm) on 3DPW [52]. Methods in the top half require training images paired with 3D ground-truth, methods in the bottom half do not.

tal analysis. It consists of 1000 synthetic humans with randomly sampled SMPL body shapes, each posed in 4 different SMPL poses sampled from Human3.6M [17] subjects 9 and 11. Global orientations are set such that the camera is facing the human's front, back, left or right. A group of 4 clean synthetic inputs and 4 corrupted inputs is rendered for each human, where the corruptions used are the same as the data augmentations applied during training. Examples are given in Figure 3.

5. Experimental Results

In this section, we present our ablation studies, where we investigate uncertainty predictions, compare probabilis-
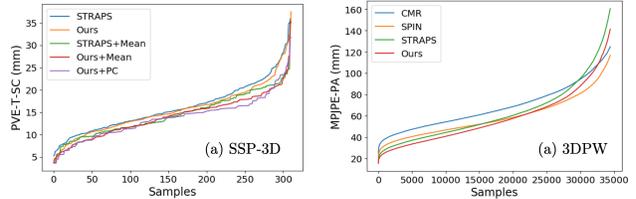

Figure 6: Comparison with other methods using sorted distributions of a) PVE-T-SC per SSP-3D evaluation sample and b) MPJPE-PA per 3DPW evaluation sample.

tic shape combination with simple averaging and explore the effects of varying input group sizes and global rotation variation within groups. We also compare our method to other approaches in terms of shape and pose accuracy.

5.1. Ablation studies

Pose and shape uncertainty. SMPL pose and shape prediction uncertainties are represented by the predicted variances σ_θ^2 and σ_β^2. Rows 2 and 5 of Figure 3 illustrate pose prediction uncertainty on clean and corrupted synthetic inputs. Heavily-occluded inputs (e.g. the corrupted input in column 4), result in large predicted variances for the pose parameters corresponding to the occluded joints, while predicted variances are smaller for visible joints. This behaviour is replicated on real inputs from the 3DPW dataset, as shown by Figure 4 where the network is uncertain about the 3D locations of occluded and out-of-frame body parts.

Figure 3 also showcases shape parameter prediction uncertainty on synthetic data, in rows 3 and 6. The network is more uncertain about body shape when the subject is heavily occluded and/or in a challenging pose, seen by comparing the sitting pose in column 2 with the standing poses in columns 1 and 3. This behaviour is also seen on real inputs in Figure 5, e.g. by comparing the crouching pose in row 1 with the standing pose in row 2.

Body shape combination method. We compare probabilistic body shape combination (from Section 3.4) with a simpler heuristic combination, where we obtain combined body shape estimates from a group of inputs $\{\mathbf{X}_n\}_{n=1}^N$ by simply averaging (i.e. taking the mean of) the shape distribution means $\{\boldsymbol{\mu}_\beta(\mathbf{X}_n)\}_{n=1}^N$. Rows 3-4 in Table 1 show that better shape estimation metrics are attained using probabilistic combination versus simple averaging on synthetic input quadruplets (examples in Figure 3). This is replicated on groups of real inputs from SSP-3D, as shown in Table 2, row 5 versus row 6. Since probabilistic combination may be interpreted as uncertainty-weighted averaging (Eqn. 4), these experiments suggest that inaccurate mean body shape predictions are generally accompanied by large prediction uncertainty, and subsequently down-weighted during probabilistic combination. This may explain why probabilistic

Group size	Method	RMSE					
		C	S	H	B	F	T
1	SPIN [27]	6.9	8.0	6.6	6.9	2.5	5.3
	STRAPS [45]	6.7	5.3	4.3	3.9	1.8	3.7
	Ours	4.9	4.7	5.5	4.2	1.8	3.9
4	SPIN [27] + Mean	6.5	8.1	6.4	6.7	2.4	5.1
	STRAPS [45] + Mean	6.1	4.2	4.0	**3.2**	1.7	3.3
	Ours + Mean	3.4	3.9	3.8	4.9	**1.6**	3.1
	Ours + PC	**3.1**	**3.8**	**2.7**	5.0	1.7	**2.8**

Table 5: Comparison with the state-of-the-art in terms of tape measurement RMSE (cm) on our private shape evaluation dataset. Errors are reported for the chest (C), stomach (S), hips (H), biceps (B), forearms (F) and thighs (T).

combination actually gives better shape metrics when evaluating on corrupted synthetic inputs compared to clean inputs in Table 1, since heavy input corruption results in inaccurate but highly-uncertain shape estimates.

Input group size. Table 1 also investigates the effect of the input group size, evaluated on our synthetic dataset, by comparing single inputs (i.e. group size of 1) with body shape combination applied to pairs and quadruplets (i.e. input group sizes of 2 and 4). Body shape metrics are significantly improved when using pairs compared to single images, suggesting that probabilistic combination is successfully using shape information from the multiple inputs. A smaller improvement is seen when using quadruplets versus pairs. Table 2 shows that increasing the input group size on real data (from SSP-3D) also results in a consistent but diminishing improvement in shape prediction metrics.

Global rotation variation. To investigate whether variation in global rotation of the subject between inputs is correlated with shape prediction accuracy, we split each group of 4 inputs from our synthetic dataset into groups of 2 in two ways: (front, left) + (back, right) and (front, back) + (left, right). We expect the latter split to be less informative for shape prediction as the pairs contain more redundant visual shape information. This is corroborated by the experiments labelled "Pairs" in Table 1, where the former split yields better shape metrics, particularly for corrupted inputs where the amount of visual shape information in each individual input is lower.

5.2. Comparison with the state-of-the-art

Shape prediction. Our method surpasses the state-of-the-art on SSP-3D in the single-input case (group size of 1), as shown in Table 3 and Figure 5. The improvement over the similar synthetic training method in STRAPS [45] is primarily due to our improved training data augmentations. When using groups of multiple images as inputs (with group size = 5), probabilistic combination outperforms simple averaging of predictions from all other methods. The dis-

tribution of errors per SSP-3D sample, shown in Figure 6, suggests that probabilistic combination particularly improves errors for challenging samples, where the uncertainty weighting is more meaningful.

Table 5 compares tape measurement errors computed using shape predictions from our method and competitors on a private dataset. Probabilistic combination results in the lowest measurement errors for large body parts, such as the chest, stomach, waist and hips. However, it is less accurate on smaller body parts (e.g. biceps and forearms), which are significantly obscured by clothing. In general, our method may over-estimate measurements for subjects with loose clothing, since silhouette-based inputs don't distinguish between clothing and the human body.

Pose prediction. While we focus on body shape estimation, Table 4 shows that our method is competitive with the state-of-the-art on 3DPW, and surpasses other methods that do not require training images paired with 3D labels. Figure 6 demonstrates that our method does well on low-to-medium difficulty samples, but struggles with the most challenging ones, which typically exhibit very severe occlusions leading to degraded proxy representations. Nevertheless, we outperform STRAPS [45] on these challenging samples due to improved data augmentation and the adaptive loss weighting discussed in Section 3.5, which results in a more stable improvement of pose metrics during training.

6. Conclusion

In this paper, we have proposed the novel task of human body shape estimation from a group of images of the same subject, without imposing any constraints on body pose, camera viewpoint or backgrounds. Our solution predicts multivariate Gaussian distributions over SMPL [33] body shape and pose parameters conditioned on the input images. We probabilistically combine predicted body shape distributions from each image to obtain a final multi-image shape prediction, and experimentally show that probabilistically combined estimates are more accurate than both individual single-image predictions, as well as naively-averaged single-image predictions, when evaluated on SSP-3D and a private dataset of tape-measured humans. Furthermore, predicting distributions over SMPL parameters allows us to estimate the heteroscedastic aleatoric uncertainty associated with pose predictions, which is useful when faced with input images containing occluded or out-of-frame body parts. Future work can consider using (i) clothed and textured synthetic data to further close the synthetic-to-real domain gap, and (ii) more expressive predicted distributions than the simple Gaussians proposed in this paper.

References

[1] Thiemo Alldieck, Marc Kassubeck, Bastian Wandt, Bodo Rosenhahn, and Marcus Magnor. Optical flow-based 3D human motion estimation from monocular video. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*, 2017.

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. In *ACM Transactions on Graphics (TOG) - Proceedings of SIGGRAPH*, volume 24, pages 408–416, 2005.

[3] 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.

[4] Eric Baum and Frank Wilczek. Supervised learning of probability distributions by neural networks. In *Advances in Neural Information Processing Systems (NeuRIPS)*, 1988.

[5] 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 *Advances in Neural Information Processing Systems (NeuRIPS)*, 2020.

[6] Christopher M. Bishop. Mixture density networks. Technical report, 1994.

[7] 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)*, Oct. 2016.

[8] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[9] Armen Der Kiureghian and Ove Dalager Ditlevsen. Aleatoric or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009.

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

[11] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. MoVi: A Large Multipurpose Motion and Video Dataset, 2020.

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

[13] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[15] 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.

[16] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 421–430, 2017.

[17] 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.

[18] 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.

[19] 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.

[20] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3D mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[22] 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.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[24] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[25] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[26] 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.

[27] 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.

[28] 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.

[29] Jogendra Nath Kundu, Mugalodi Rakesh, Varun Jampani, Rahul M Venkatesh, and R. Venkatesh Babu. Appearance consensus driven self-supervised human mesh recovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[30] 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.

[31] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3D human pose estimation with mixture density network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] 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.

[33] 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.

[34] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[35] David Mohlin, Josephine Sullivan, and Gérald Bianchi. Probabilistic orientation estimation with matrix fisher distributions. In *Advances in Neural Information Processing Systems (NeuRIPS)*, 2020.

[36] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[37] D. A. Nix and A. S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of IEEE International Conference on Neural Networks (ICNN)*, volume 1, pages 55–60 vol.1, 1994.

[38] 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.

[39] 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.

[40] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[41] 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.

[42] Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Sept. 2018.

[43] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[44] 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.

[45] 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)*, 2020.

[46] B. M. Smith, V. Chari, A. Agrawal, J. M. Rehg, and R. Sever. Towards accurate 3D human body reconstruction from silhouettes. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 279–288, 2019.

[47] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV*, 2019.

[48] 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.

[49] 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.

[50] 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.

[51] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[52] 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.

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

[54] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[55] 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.

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

[57] 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.