Hierarchical Kinematic Probability Distributions for 3D Human Shape and Pose Estimation from 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 an RGB image. This is often an ill-posed problem, since multiple plausible 3D bodies may match the visual evidence present in the input - particularly when the subject is occluded. Thus, it is desirable to estimate a distribution over 3D body shape and pose conditioned on the input image instead of a single 3D reconstruction. We train a deep neural network to estimate a hierarchical matrix-Fisher distribution over relative 3D joint rotation matrices (i.e. body pose), which exploits the human body's kinematic tree structure, as well as a Gaussian distribution over SMPL body shape parameters. To further ensure that the predicted shape and pose distributions match the visual evidence in the input image, we implement a differentiable rejection sampler to impose a reprojection loss between ground-truth 2D joint coordinates and samples from the predicted distributions, projected onto the image plane. We show that our method is competitive with the state-of-the-art in terms of 3D shape and pose metrics on the SSP-3D and 3DPW datasets, while also yielding a structured probability distribution over 3D body shape and pose, with which we can meaningfully quantify prediction uncertainty and sample multiple plausible 3D reconstructions to explain a given input image. Code is available at https://github.com/akashsengupta1997/HierarchicalProbabilistic3DHuman.

1. Introduction

3D human body shape and pose estimation from an RGB image is a challenging computer vision problem, partly due to its under-constrained nature wherein multiple 3D human bodies may explain a given 2D image, especially when the subject is significantly occluded, as is common for in-the-wild images. Several recent works [55, 19, 26, 25, 47, 65, 12, 38, 14, 41, 40, 56, 36, 53] use deep neural networks to regress a single body shape and pose solution, which can result in impressive 3D body reconstructions given sufficient visual evidence in the input image. However, when visual evidence of the subject's shape and pose is obscured, e.g. due to occluding objects or self-occlusions, a single solution does not fully describe the space of plausible 3D reconstructions. In contrast, we aim to estimate a *structured probability distribution* over 3D body shape and pose, conditioned on the input image, thereby allowing us to sample any number of plausible 3D reconstructions and quantify prediction uncertainty over the 3D body surface, as shown in Figure 1.

Figure 1. 3D reconstruction samples and per-vertex uncertainty corresponding to the predicted hierarchical shape and pose distributions computed from the given input images.

We use the SMPL body model [33] to represent human shape and pose. Identity-dependent body shape is parameterised by coefficients of a PCA basis - hence, a simple multivariate Gaussian distribution over the shape parameters is suitable. Body pose is parameterised by relative 3D joint rotations along the SMPL kinematic tree, which may be represented using rotation matrices. Regressing rotation matrices using neural networks is non-trivial, since they lie in $SO(3)$, a non-linear 3D manifold with a different topology to $\mathbb{R}^{3\times3}$ or $\mathbb{R}^9$, the space in which unconstrained neural network outputs lie. However, one can define probability density functions over the Lie group $SO(3)$, such as the matrix-Fisher distribution [34, 11, 21], the parameter of which is an element of $\mathbb{R}^{3\times3}$ and may be easily regressed with a neural network [35]. We propose a hierarchical probability distribution over relative 3D joint rotations along the SMPL kinematic tree, wherein the probability density func-

tion of each joint's relative rotation matrix is a matrix-Fisher distribution conditioned on the parents of that joint in the kinematic tree. We train a deep neural network to predict the parameters of such a distribution over body pose, alongside a Gaussian distribution over SMPL shape.

Moreover, to ensure that 3D bodies sampled from the predicted distributions match the 2D input image, we implement a reprojection loss between predicted samples and ground-truth visible 2D joint annotations. To allow for the backpropagation of gradients through the sampling operation, we present a differentiable rejection sampler for matrix-Fisher distributions over relative 3D joint rotations.

Finally, a key obstacle for SMPL body shape regression from in-the-wild images is the lack of training datasets with *accurate* and *diverse* body shape labels [47]. To overcome this, we follow [47, 53, 41, 48] and utilise synthetic data, randomly generated on-the-fly during training. Inspired by [7], we use convolutional edge filters to close the large synthetic-to-real gap and show that using edge-based inputs yields better performance than commonly-used silhouette-based inputs [47, 53, 48, 41], due to improved robustness and capacity to retain visual shape information.

In summary, our main contributions are as follows:

- Given an input image, we predict a novel hierarchical matrix-Fisher distribution over relative 3D joint rotation matrices, whose structure is explicitly informed by the SMPL kinematic tree, alongside a Gaussian distribution over SMPL shape parameters.
- We present a differentiable rejection sampler to sample any number of plausible 3D reconstructions and quantify prediction uncertainty over the body surface. This enables a reprojection loss between predicted samples and ground-truth coordinates of visible 2D joints, further ensuring that the predicted distributions are consistent with the input image.
- We use simple convolutional edge filters to improve the random synthetic training framework used by [47, 48]. Edge filtering is a computationally-cheap and robust method for closing the domain gap between synthetic RGB training data and real RGB test data.

2. Related Work

This section reviews approaches to monocular 3D human body shape and pose estimation, as well as deep-learning-based methods for probabilistic rotation estimation.

Monocular 3D shape and pose estimation methods can be classified as optimisation-based or learning-based. Optimisation-based approaches fit a parametric 3D body model [33, 1, 39, 18] to 2D observations, such as 2D keypoints [5, 29], silhouettes [29] or body part segmentations [63], by optimising a suitable cost function. These methods do not require expensive 3D-labelled training data, but are sensitive to poor intialisations and noisy observations.

Learning-based approaches can be further split into model-free or model-based. Model-free methods use deep networks to directly output human body vertex meshes [26, 36, 65, 64, 8], voxel grids [56] or implicit surfaces [45, 46] from an input image. In contrast, model-based methods [19, 47, 38, 12, 55, 14, 41, 40, 61] regress 3D body model parameters [39, 33, 18, 1], which give a low-dimensional representation of a 3D human body. To overcome the lack of in-the-wild 3D-labelled training data, several methods [19, 61, 26, 12, 14] use diverse 2D-labelled data as a source of weak supervision. [25] extends this approach by incorporating optimisation into their model training loop, lifting 2D labels to self-improving 3D labels. These approaches often result in impressive 3D pose predictions, but struggle to accurately predict a diverse range of body shapes, since 2D keypoint supervision only provides a sparse shape signal. Shape prediction accuracy may be improved using synthetic training data [47, 53, 41, 48] consisting of synthetic input proxy representations (PRs) paired with ground-truth body shape and pose. PRs commonly consist of silhouettes and 2D joint heatmaps [47, 41, 48], necessitating accurate silhouette segmentations [24, 15] at test-time, which is not guaranteed for challenging in-the-wild inputs. Other methods [56] pre-train on synthetic RGB inputs [57] and then fine-tune on the scarce and limited-shape-diversity real 3D training data available [16, 58], to avoid over-fitting to artefacts in low-fidelity synthetic data. In contrast, we utilise edge-based PRs, hence dropping the reliance on accurate segmentation networks without requiring fine-tuning on real data or high-fidelity synthetic data.

3D human shape and pose distribution estimation. Early optimisation-based 3D pose estimators [50, 51, 52, 9, 10] specified a cost function corresponding to the posterior probability of 3D pose given 2D observations and analysed its multi-modal structure due to ill-posedness. Strategies to sample multiple 3D poses with high posterior probability included cost-covariance-scaled [50] and inverse-kinematics-based [52] global search and local refinement, as well as cost-function-modifying MCMC [51]. Recently, several learning-based methods [49, 31, 17, 59, 37] predict multi-modal distributions over 3D joint locations conditioned on 2D inputs, using Bayesian mixture of experts [49], mixture density networks [31, 4, 37] or normalising flows [59, 44]. Our method extends beyond 3D joints and predicts distributions over human pose *and shape*. This has been addressed by Biggs *et al*. [3], who predict a categorical distribution over a set of SMPL [33] parameter hypotheses. Sengupta *et al*. [48] estimate an independent Gaussian distribution over both SMPL shape and joint rotation vectors. In contrast, we note that 3D rotations lie in $SO(3)$, motivating our hierarchical matrix-Fisher distribution.

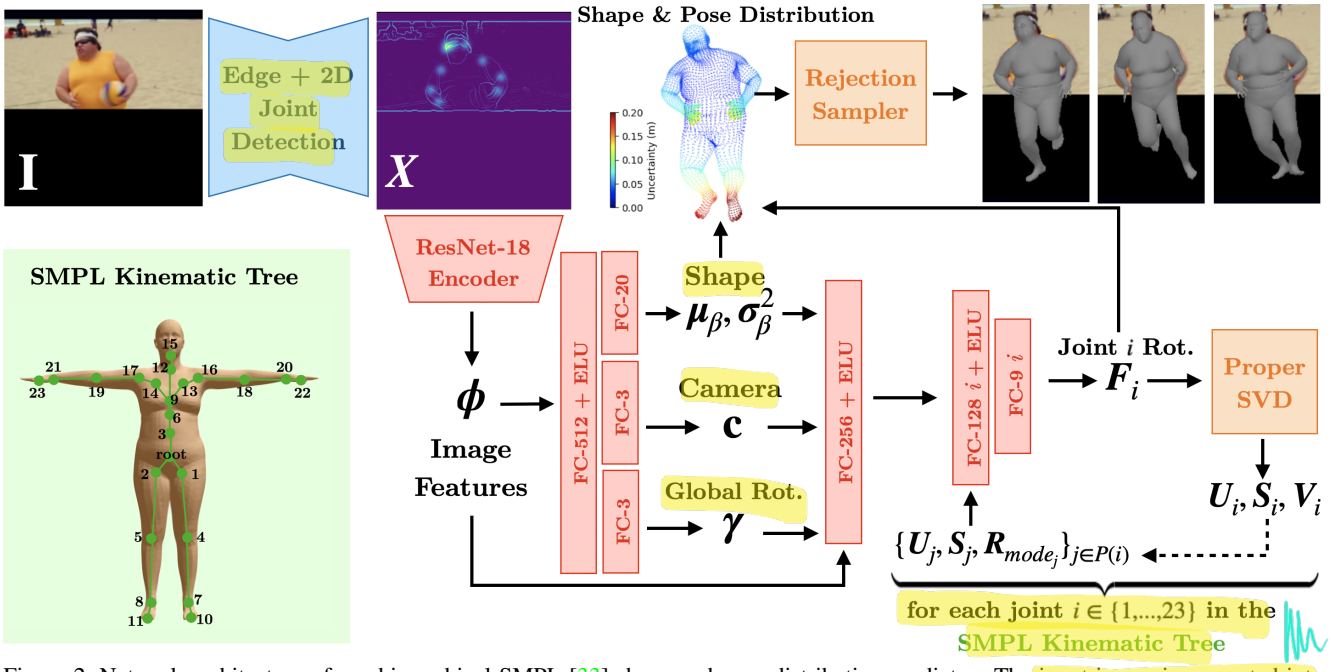

Figure 2. Network architecture of our hierarchical SMPL [33] shape and pose distribution predictor. The input image is converted into an edge-and-joint-heatmap proxy representation, which is passed through the prediction network to produce distributions over shape parameters and relative 3D joint rotation matrices. Rejection sampling is used to sample 3D reconstructions from the predicted distributions.

Rotation distribution estimation via deep learning. Prokudin *et al*. [42] use biternion networks to predict a mixture-of-von-Mises distribution over object pose angle. Gilitschenski *et al*. [13] use a Bingham distribution over unit quaternions to represent orientation uncertainty. However, these works have to enforce constraints on the parameters of their predicted distributions (e.g. positive semi-definiteness). To overcome this, Mohlin *et al*. [35] train a deep network to regress a matrix-Fisher distribution [34, 11, 21] over 3D rotation matrices. We adapt this approach to define our hierarchical matrix-Fisher distribution over relative 3D joint rotation matrices.

3. Method

This section provides an overview of SMPL [33] and the matrix-Fisher distribution [11, 21, 34], presents our structured, hierarchical pose and shape distribution estimation architecture and discusses the loss functions used to train it.

3.1. SMPL model

SMPL [33] is a parametric 3D human body model. Identity-dependent body shape is represented by shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, which are coefficients of a PCA body shape basis. Body pose is defined by the relative 3D rotations of the bones formed by the 23 body (i.e. non-root) joints in the SMPL kinematic tree. The rotations may be represented using rotation matrices $\{\boldsymbol{R}_i\}_{i=1}^{23}$, where $\boldsymbol{R}_i \in SO(3)$. We parameterise the global rotation (i.e. rotation of the root joint) in axis-angle form by $\boldsymbol{\gamma} \in \mathbb{R}^3$. A differentiable function $\mathcal{S}(\{\boldsymbol{R}_i\}_{i=1}^{23}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ maps the input pose and shape parameters to an output vertex mesh $\boldsymbol{V} \in \mathbb{R}^{6890\times 3}$. 3D joint locations, for L joints of interest, are obtained as $\boldsymbol{J}^{3D} = \mathcal{J}\boldsymbol{V}$ where $\mathcal{J} \in \mathbb{R}^{L\times 6890}$ is a linear vertex-to-joint regression matrix.

3.2. Matrix-Fisher distribution over $SO(3)$

The 3D special orthogonal group may be defined as $SO(3) = \{\boldsymbol{R} \in \mathbb{R}^{3\times 3} | \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}, \det(\boldsymbol{R}) = 1\}$. The matrix-Fisher distribution [11, 21, 34] defines a probability density function over $SO(3)$, given by

$$p(\boldsymbol{R}|\boldsymbol{F}) = \frac{1}{c(\boldsymbol{F})}\exp(\mathrm{tr}(\boldsymbol{F}^T\boldsymbol{R})) = \mathcal{M}(\boldsymbol{R};\boldsymbol{F}) \quad (1)$$

where $\boldsymbol{F} \in \mathbb{R}^{3\times 3}$ is the matrix parameter of the distribution, $c(\boldsymbol{F})$ is the normalising constant and $\boldsymbol{R} \in SO(3)$. We present some key properties of the matrix-Fisher distribution below, but refer the reader to [30, 35] for further details, visualisations and a method for approximating the intractable normalising constant and its gradient w.r.t. $\boldsymbol{F}$.

The properties of $\mathcal{M}(\boldsymbol{R};\boldsymbol{F})$ can be described in terms of the singular value decomposition (SVD) of $\boldsymbol{F}$, denoted by $\boldsymbol{F} = \boldsymbol{U}'\boldsymbol{S}'\boldsymbol{V}'^T$, with $\boldsymbol{S}' = \mathrm{diag}(s'_1, s'_2, s'_3)$. $\boldsymbol{U}'$ and $\boldsymbol{V}'$ are orthonormal matrices, but they may have a determinant of -1 and thus are not necessarily elements of $SO(3)$. There-

fore, a *proper* SVD [30] $\boldsymbol{F} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ is used, where

$$\begin{aligned} \boldsymbol{U} &= \boldsymbol{U}'\text{diag}(1, 1, \det(\boldsymbol{U}')) \\ \boldsymbol{V} &= \boldsymbol{V}'\text{diag}(1, 1, \det(\boldsymbol{V}')) \\ \boldsymbol{S} &= \text{diag}(s_1, s_2, s_3) = \text{diag}(s_1', s_2', \det(\boldsymbol{U}'\boldsymbol{V}')s_3') \end{aligned} \tag{2}$$

which ensures that $\boldsymbol{U}, \boldsymbol{V} \in SO(3)$. Then, the *mode* of the distribution is given by [30]

$$\boldsymbol{R}_{\text{mode}} = \underset{\boldsymbol{R} \in SO(3)}{\arg\max}\, p(\boldsymbol{R}|\boldsymbol{F}) = \boldsymbol{U}\boldsymbol{V}^T. \tag{3}$$

The columns of $\boldsymbol{U}$ define the distribution's *principal axes* of rotation (analogous to the principal axes of a multivariate Gaussian distribution), while the proper singular values in $\boldsymbol{S}$ give the *concentration* of the distribution for rotations about the principal axes [30]. Specifically, the concentration along rotations of $\boldsymbol{R}_{\text{mode}}$ about the i-th principal axis (i-th column of $\boldsymbol{U}$) is given by $s_j + s_k$ for $(i, j, k) \in \{(1, 2, 3), (2, 3, 1), (3, 1, 2)\}$. The concentration of the distribution may be different about each principal axis, allowing for axis-dependent rotation uncertainty modelling.

3.3. Proxy representation computation

Given an input RGB image $\mathbf{I}$, we first compute a proxy representation $\boldsymbol{X}$ (see Figure 2), consisting of an edge-image concatenated with joint heatmaps. Comparisons with silhouette- and RGB-based representations are given in Section 5.1. Edge-images are obtained with Canny edge detection [6]. 2D joint heatmaps are computed using HRNet-W48 [54], and joint predictions with low confidence scores (< 0.6) are thresholded out. The edge-image and joint heatmaps are stacked along the channel dimension to produce $\boldsymbol{X} \in \mathbb{R}^{H \times W \times (L+1)}$. Proxy representations [47, 41] are used to close the domain gap between synthetic training images and real test-time RGB images, since synthetic proxy representations are more similar to their real counterparts than synthetic RGB images are to real RGB images.

3.4. Body shape and pose distribution prediction

Our goal is to predict a probability distribution over relative 3D joint rotations $\{\boldsymbol{R}_i\}_{i=1}^{23}$ and SMPL shape parameters $\boldsymbol{\beta}$ conditioned upon a given input proxy representation $\boldsymbol{X}$. We also predict deterministic estimates of the global body rotation $\boldsymbol{\gamma}$ and weak-perspective camera parameters $\mathbf{c} = [s, t_x, t_y]$, representing scale and xy translation.

Since $\boldsymbol{\beta}$ represents the linear coefficients of a PCA shape-space, a Gaussian distribution with a diagonal covariance matrix is suitable [48],

$$p(\boldsymbol{\beta}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_\beta(\boldsymbol{X}), \text{diag}(\boldsymbol{\sigma}_\beta^2(\boldsymbol{X})) \tag{4}$$

where the mean $\boldsymbol{\mu}_\beta$ and variances $\boldsymbol{\sigma}_\beta^2$ are functions of $\boldsymbol{X}$.

The matrix-Fisher distribution (Equation 1) may be naively used to define a distribution over 3D joint rotations

$$p(\boldsymbol{R}_i|\boldsymbol{X}) = \mathcal{M}(\boldsymbol{R}_i; \boldsymbol{F}_i(\boldsymbol{X})) \tag{5}$$

for $i \in \{1, 2, ..., 23\}$. Here, each joint is modelled *independently* of all the other joints. Thus, the matrix parameter of the i-th joint, $\boldsymbol{F}_i$, is a function of the input $\boldsymbol{X}$ only.

To predict the parameters of this naive, independent distribution over 3D joint rotations, in addition to the shape distribution parameters, global body rotation and weak-perspective camera, we learn a function f_{indep} mapping the input $\boldsymbol{X}$ to the set of desired outputs $Y = \{\{\boldsymbol{F}_i\}_{i=1}^{23}, \boldsymbol{\mu}_\beta, \boldsymbol{\sigma}_\beta^2, \boldsymbol{\gamma}, \mathbf{c}\}$, where f_{indep} is represented by a deep neural network with weights $\mathbf{W}_{\text{indep}}$.

However, the independent matrix-Fisher distribution in Equation 5 does not model SMPL 3D joint rotations faithfully, since the rotation of each part/bone is defined *relative* to its parent joint in the SMPL kinematic tree. Hence, a distribution over the i-th rotation matrix $\boldsymbol{R}_i$ conditioned on the input $\boldsymbol{X}$ should be informed by the distributions over all its parent joints $P(i)$, as well as the global body rotation $\boldsymbol{\gamma}$, to enable the distribution to match the 2D visual pose evidence present in $\boldsymbol{X}$. Furthermore, 3D joints in the SMPL rest-pose skeleton are dependent upon the shape parameters $\boldsymbol{\beta}$, while the mapping from 3D to the 2D image plane is given by the camera model. Hence, a distribution over $\boldsymbol{R}_i$ given $\boldsymbol{X}$ should also consider the predicted shape mean $\boldsymbol{\mu}_\beta$ and variance $\boldsymbol{\sigma}_\beta^2$, as well as the predicted camera $\mathbf{c}$. This is similar to the rationale behind the deterministic iterative/hierarchical predictors in [19, 12], except we model these relationships in a probabilistic sense, by defining

$$\begin{aligned} p\big(\boldsymbol{R}_i|\boldsymbol{X}, \{\boldsymbol{F}_j\}_{j \in P(i)}, \boldsymbol{\gamma}, \boldsymbol{\mu}_\beta, \boldsymbol{\sigma}_\beta^2, \mathbf{c}\big) &= \mathcal{M}(\boldsymbol{R}_i; \boldsymbol{F}_i) \\ \boldsymbol{F}_i = f_i\big(\boldsymbol{X}, \{(\boldsymbol{U}_j, \boldsymbol{S}_j, \boldsymbol{R}_{\text{mode}_j})\}_{j \in P(i)}, \boldsymbol{\gamma}, \boldsymbol{\mu}_\beta, \boldsymbol{\sigma}_\beta^2, \mathbf{c}\big) & \end{aligned} \tag{6}$$

for $i \in \{1, 2, ..., 23\}$. Now, the matrix parameter of the i-th joint is a function of all its parent distributions, represented by the principal axes $\boldsymbol{U}_j$, singular values $\boldsymbol{S}_j$ and modes $\boldsymbol{R}_{\text{mode}_j} = \boldsymbol{U}_j\boldsymbol{V}_j^T$ for $j \in P(i)$, as well as the shape distribution $\{\boldsymbol{\mu}_\beta, \boldsymbol{\sigma}_\beta^2\}$, global rotation $\boldsymbol{\gamma}$, camera parameters $\mathbf{c}$ and the input $\boldsymbol{X}$. Note that the parent distributions are themselves functions of *their* respective parent joints, while $\boldsymbol{\gamma}, \boldsymbol{\mu}_\beta, \boldsymbol{\sigma}_\beta^2$ and $\mathbf{c}$ are all functions of $\boldsymbol{X}$.

To predict the parameters of the hierarchical matrix-Fisher distribution in Equation 6, we propose a hierarchical neural network architecture f_{hier}, with weights $\mathbf{W}_{\text{hier}}$ (Figure 2). When considered as a black-box, f_{hier} yields the same set of outputs Y as f_{indep}. However, f_{hier} utilises the iterative hierarchical architecture presented in Figure 2, which amounts to multiple streams of fully-connected layers, each following one "limb" of the kinematic tree. In contrast, f_{indep} predicts pose similarly to shape, camera and global rotation parameters, using a single stream of fully-connected layers. We compare the naive independent formulation with the hierarchical formulation in Section 5.1.

3.5. Loss functions

Distribution prediction networks are trained with a synthetic dataset $\{\boldsymbol{X}^n, (\{\boldsymbol{R}_i^n\}_{i=1}^{23}, \boldsymbol{\beta}^n, \boldsymbol{\gamma}^n)\}_{n=1}^N$ (Section 4).

Negative log-likelihood (NLL) loss on distribution parameters. The NLL corresponding to the Gaussian body shape distribution (Equation 4) is given by:

$$\mathcal{L}_{\beta\text{-NLL}} = -\sum_{n=1}^{N} \log \mathcal{N}\big(\boldsymbol{\beta}^n; \boldsymbol{\mu}_\beta(\boldsymbol{X}^n), \text{diag}(\boldsymbol{\sigma}_\beta^2(\boldsymbol{X}^n))\big). \tag{7}$$

The NLL corresponding to the matrix-Fisher distribution over relative 3D joint rotations is defined as [35]:

$$\begin{aligned}\mathcal{L}_{R\text{-NLL}} &= -\sum_{n=1}^{N} \log \mathcal{M}(\boldsymbol{R}_i^n; \boldsymbol{F}_i^n) \\ &= \sum_{n=1}^{N} \log c(\boldsymbol{F}_i^n) - \text{tr}(\boldsymbol{F}_i^{nT} \boldsymbol{R}_i^n)\end{aligned} \tag{8}$$

for $i \in \{1, 2, ..., 23\}$, where $\boldsymbol{F}_i^n$ may be obtained via the independent or hierarchical matrix-Fisher models presented above. Intuitively, the trace term pushes the predicted distribution mode $\boldsymbol{R}_{\text{mode}_i}^n$ (Equation 3) towards the target $\boldsymbol{R}_i^n$, while the log normalising constant acts as a regulariser, preventing the singular values of $\boldsymbol{F}_i^n$ from getting too large [35]. All predicted distribution parameters are dependent on the model weights, $\mathbf{W}_{\text{indep}}$ or $\mathbf{W}_{\text{hier}}$, which are learnt in a maximum likelihood framework aiming to minimise the joint shape and pose NLL: $\mathcal{L}_{\text{NLL}} = \mathcal{L}_{\beta\text{-NLL}} + \mathcal{L}_{R\text{-NLL}}$.

Loss on global body rotation. We predict deterministic estimates of the global body rotation vectors $\hat{\boldsymbol{\gamma}}^n$, which are supervised using ground-truth global rotations $\boldsymbol{\gamma}^n$, with loss $\mathcal{L}_{\text{global}} = \sum_{n=1}^{N} \|\boldsymbol{R}(\boldsymbol{\gamma}_n) - \boldsymbol{R}(\hat{\boldsymbol{\gamma}}_n)\|_F^2$. $\boldsymbol{R}(\boldsymbol{\gamma}) \in SO(3)$ is the rotation matrix corresponding to $\boldsymbol{\gamma}$.

2D joints loss on samples. Applying $\mathcal{L}_{\text{NLL}}$ alone results in overly uncertain predicted 3D shape and pose distributions (see Section 5.1). To ensure that the predicted distributions match the visual evidence in the input $\boldsymbol{X}^n$, we impose a reprojection loss between ground-truth 2D joint coordinates (in the image plane) and predicted 2D joint samples, which are obtained by *differentiably sampling* 3D bodies from the predicted distributions and projecting to 2D using the predicted camera $\mathbf{c}^n = [s^n, t_x^n, t_y^n]$. Ground-truth 2D joints $\boldsymbol{J}_{\text{2D}}^n$ are computed from $\{\{\boldsymbol{R}_i^n\}_{i=1}^{23}, \boldsymbol{\beta}^n, \boldsymbol{\gamma}^n\}$ during synthetic training data generation (see Section 4).

We adapt the rejection sampler presented in [20] to sample from a matrix-Fisher distribution $\mathcal{M}(\boldsymbol{R}; \boldsymbol{F})$, modifying it to allow for backpropagation of gradients through the proposal sampling step (lines 5-7 in Algorithm 1). We refer the reader to [20] for further details about the rejection sampler. In short, to simulate a matrix-Fisher distribution with parameter $\boldsymbol{F} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$, we sample unit quaternions from a Bingham distribution [34] over the unit 3-sphere S^3, with Bingham parameter $\boldsymbol{A}$ computed from $\boldsymbol{S}$,

Algorithm 1: Differentiable Rejection Sampler

Input: $\boldsymbol{U}$, $\boldsymbol{S} = \text{diag}(s_1, s_2, s_3)$, $\boldsymbol{V}$, b
Output: $\hat{\boldsymbol{R}} \in SO(3)$ s.t. $\hat{\boldsymbol{R}} \sim \mathcal{M}(\boldsymbol{R}; \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T)$

1 $\boldsymbol{A} = \text{diag}(0, 2(s_2 + s_3), 2(s_1 + s_3), 2(s_1 + s_2))$
2 $\boldsymbol{\Omega} = \boldsymbol{I}_4 + \frac{2}{b}\boldsymbol{A}$
3 $M = \exp\left(\frac{b-4}{2}\right)\left(\frac{4}{b}\right)^2$
4 **repeat**
5 Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_4, \boldsymbol{I}_4)$
6 $\boldsymbol{y} = (\boldsymbol{\Omega}^{-1})^{\frac{1}{2}}\boldsymbol{\epsilon}$
7 Propose $\boldsymbol{x} = \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}$ s.t. $\boldsymbol{x} \in S^3$
8 Sample $w \sim \text{Unif}[0, 1]$
9 **until** $w < \frac{\exp(-\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})}{M(\boldsymbol{x}^T\boldsymbol{\Omega}\boldsymbol{x})^{-2}}$;
10 $\hat{\boldsymbol{Q}} = quaternion\_to\_matrix(\boldsymbol{x})$ s.t. $\hat{\boldsymbol{Q}} \in SO(3)$
11 **return** $\hat{\boldsymbol{R}} = \boldsymbol{U}\hat{\boldsymbol{Q}}\boldsymbol{V}^T$

and then convert the sampled quaternions into rotation matrices [20, 34] with the desired matrix-Fisher distribution. Rejection sampling is used to sample from the Bingham distribution, which has pdf $p_{\text{Bing}}(\boldsymbol{x}) \propto \exp(-\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})$ for $\boldsymbol{x} \in S^3$. The proposal distribution for the rejection sampler is an angular central Gaussian (ACG) distribution, with pdf $p_{\text{ACG}}(\boldsymbol{x}) \propto (\boldsymbol{x}^T\boldsymbol{\Omega}\boldsymbol{x})^{-2}$. The ACG distribution is easily simulated [20] by sampling from a zero-mean Gaussian distribution with covariance matrix $\boldsymbol{\Omega}^{-1}$ and normalising to unit-length (lines 5-7 in Algorithm 1). The re-parameterisation trick [23] is used to differentiably sample from this zero-mean Gaussian, thus allowing for backpropagation of gradients through the rejection sampler.

Algorithm 1 samples K sets of relative 3D joint rotation matrices $\{\{\hat{\boldsymbol{R}}_{i,k}^n\}_{i=1}^{23}\}_{k=1}^K$ from the corresponding distributions $\{\mathcal{M}(\boldsymbol{R}_i^n; \boldsymbol{F}_i^n)\}_{i=1}^{23}$. Furthermore, we differentiably sample K SMPL shape vectors from the predicted Gaussian distribution $\{\hat{\boldsymbol{\beta}}_k^n \sim \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_\beta(\boldsymbol{X}^n), \text{diag}(\boldsymbol{\sigma}_\beta^2(\boldsymbol{X}^n)))\}_{k=1}^K$, again using the re-parameterisation trick [23].

The body shape and 3D joint rotation samples are converted into 2D joint samples using the SMPL model and weak-perspective camera parameters

$$\hat{\boldsymbol{J}}_{\text{2D}_k}^n = s^n \Pi(\mathcal{JS}(\{\hat{\boldsymbol{R}}_{i,k}^n\}_{i=1}^{23}, \hat{\boldsymbol{\beta}}_k^n, \hat{\boldsymbol{\gamma}}^n)) + [t_x^n, t_y^n] \tag{9}$$

where $\Pi()$ is an orthographic projection. The reprojection loss applied between the predicted 2D joint samples and the visible target 2D joint coordinates is given by

$$\mathcal{L}_{\text{2D Samples}} = \sum_{n=1}^{N}\sum_{k=1}^{K} \|\boldsymbol{\omega}^n(\boldsymbol{J}_{\text{2D}}^n - \hat{\boldsymbol{J}}_{\text{2D}_k}^n)\|_2^2 \tag{10}$$

where the visibilities of the target joints are denoted by $\boldsymbol{\omega}^n \in \{0, 1\}^L$ (1 if visible, 0 otherwise).

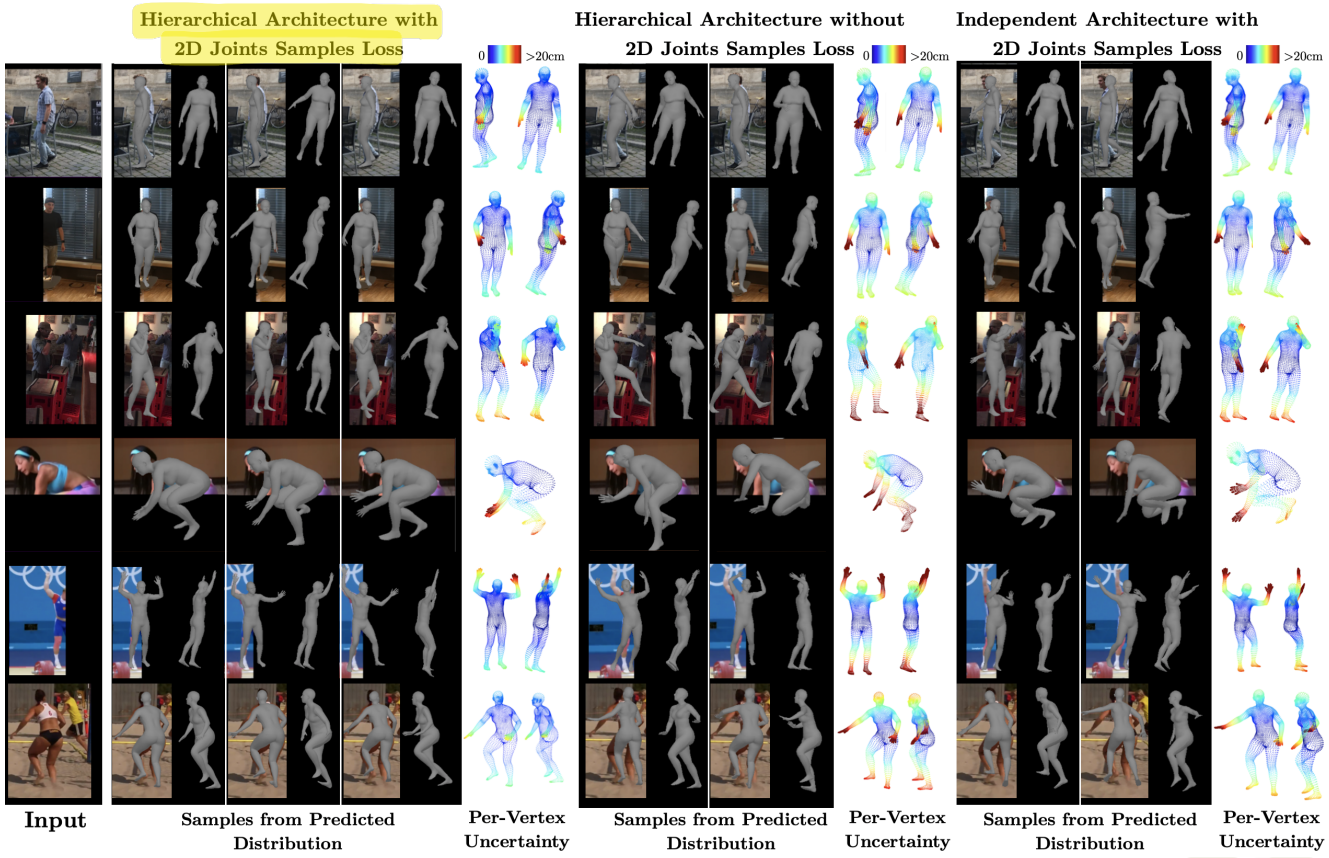

Figure 3. 3D reconstruction samples and per-vertex uncertainty corresponding to shape and pose distributions predicted using the hierarchical architecture with 2D samples loss (left), hierarchical architecture without 2D samples loss (centre) and independent architecture with 2D samples loss (right). Per-vertex uncertainty (in cm) is estimated by sampling 100 SMPL meshes from the predicted distributions and determining the average Euclidean distance from the sample mean for each vertex. Both the hierarchical architecture *and* the sample reprojection loss are required for predicted distributions to match the inputs, while demonstrating greater uncertainty for ambiguous parts.

4. Implementation Details

Synthetic training data. To train our 3D body shape and pose distribution prediction networks, we require a training dataset $\{\boldsymbol{X}^n, (\{\boldsymbol{R}_i^n\}_{i=1}^{23}, \boldsymbol{\beta}^n, \boldsymbol{\gamma}^n)\}_{n=1}^N$. We extend the synthetic training frameworks presented in [47, 48], which involve generating inputs and corresponding SMPL body shape and pose (i.e. 3D joint rotation) labels randomly and on-the-fly during training. In brief, for every training iteration, SMPL shapes $\boldsymbol{\beta}^n$ are randomly sampled from a prior Gaussian distribution while relative 3D joint rotations $\{\boldsymbol{R}_i^n\}_{i=1}^{23}$ and global rotation $\boldsymbol{\gamma}^n$ are chosen from the training sets of UP-3D [29], 3DPW [58] or Human3.6M [16]. These are converted into training inputs $\boldsymbol{X}^n$ and ground-truth 2D joint coordinates $\boldsymbol{J}^n$ using the SMPL model and a light-weight renderer [43]. Cropping, occlusion and noise augmentations are then applied to the synthetic inputs.

Previous synthetic training frameworks [47, 48, 53] often use silhouette-based training inputs. This necessitates accurate human silhouette segmentation at test-time, which may be challenging to do robustly. In contrast, our input representations consist of edge-images concatenated with 2D joint heatmaps. To generate edge-images, we first create synthetic RGB images by rendering textured SMPL meshes. For each training mesh, clothing textures are randomly chosen from [57, 2]. The textured SMPL mesh is rendered onto a background image (randomly chosen from LSUN [62]), using randomly-sampled lighting and camera parameters. Canny edge detection [6] is used to compute edge-images from the synthetic RGB images. We show in Section 5.1 that, despite the lack of photorealism in the synthetic RGB images, edge-filtering bridges the synthetic-to-real domain gap at test-time - and performs better than either silhouette-based or synthetic-RGB-based training inputs in our experiments. Examples of synthetic training samples are given in the supplementary material.

Training details. We use Adam [22] with a learning rate of 0.0001, batch size of 80 and train for 150 epochs. For stability, the 2D joints reprojection loss is only applied on the mode pose and shape (projected to 2D) in the first 50 epochs and not on the samples, which are supervised in the next 100 epochs. To boost 3D pose metrics, an MSE loss on

(Hierarch. consistently outperforms indep.)

Input Type	Architecture	2D Samples Loss	Synthetic Test Data			SSP-3D		3DPW
			MPJPE-SC	PVE-T-SC	2D Joint Err. Mode/Samples	PVE-T-SC	2D Joint Err. Mode/Samples	MPJPE-SC
Silh. + J2DHmap	Independent	No	84.9	12.8	7.2 / 11.6	14.3	6.0 / 11.9	93.0
RGB + J2DHmap	Independent	No	79.9	**11.3**	7.1 / 11.7	14.0	5.9 / 12.0	92.8
Edge + J2DHmap	Independent	No	85.8	12.9	7.5 / 12.0	13.7	5.9 / 11.8	88.4
Edge + J2DHmap	Independent	Yes	86.3	13.2	7.6 / 8.9	13.9	6.2 / 9.6	91.3
Edge + J2DHmap	Hierarchical	No	84.4	12.8	7.3 / 10.4	**13.6**	5.3 / 11.2	87.7
Edge + J2DHmap	Hierarchical	Yes	**79.1**	12.6	**6.7 / 6.9**	**13.6**	**4.8 / 6.9**	**84.7**

Table 1. Experiments investigating different input representations, hierarchical versus independent distribution prediction networks and the 2D samples reprojection loss, evaluated in terms of shape and pose prediction metrics on synthetic data, SSP-3D [47] and 3DPW [58].

the mode 3D joint locations is applied in the final 50 epochs.

Evaluation datasets. 3DPW [58] is used to evaluate 3D pose prediction accuracy. We report mean-per-joint-position-error after scale correction (MPJPE-SC) [47] and after Procrustes analysis (MPJPE-PA), both in mm. Both metrics are computed using the mode 3D joint coordinates of the predicted shape and pose distributions.

SSP-3D is primarily used to evaluate 3D body shape prediction accuracy, using per-vertex Euclidean error in a T-pose after scale-correction (PVE-T-SC) [47] in mm, computed with the mode 3D body shape from the predicted shape distribution. We also evaluate 2D joint prediction error (2D Joint Err. Mode/Samples) in pixels, computed using both the mode 3D body and 10 3D bodies randomly sampled from the predicted shape and pose distributions, projected onto the image plane using the camera prediction. 2D joint error is evaluated on *visible* target 2D joints only.

Finally, we use a synthetic test dataset for our ablation studies investigating different input representations. It consists of 1000 synthetic input-label pairs, generated in the same way as the synthetic training data, with poses sampled from the test set of Human3.6M. [16].

5. Experimental Results

This section investigates different input representations and the benefits of the 2D joints samples loss, compares independent and hierarchical distribution predictors and benchmarks our method against the state-of-the-art.

5.1. Ablation studies

Input proxy representation. Rows 1-3 in Table 1 compare different choices of input proxy representation: binary silhouettes, RGB images and edge-filtered images (each additionally concatenated with 2D joint heatmaps). The independent network architecture is used for all three input types. To investigate the synthetic-to-real domain gap, metrics are presented for synthetic test data, as well as real test images from SSP-3D and 3DPW. For the latter, silhouette segmentation is carried out with DensePose [15]. Using RGB-based input representations (row 2) results in the best 3D shape and pose metrics on synthetic data, which is reasonable since RGB contains more information than both silhouettes and edge-filtered images. However, metrics are significantly worse on real datasets, suggesting that the network has over-fitted to unrealistic artefacts present in low-fidelity (i.e. computationally cheap) synthetic RGB images. Silhouette-based input representations (row 1) also demonstrate a deterioration of 3D metrics on real test data compared to synthetic data, since they are heavily reliant upon accurate silhouettes, which are difficult to robustly segment in test images containing challenging poses or severe occlusions. Inaccurate silhouette segmentations critically impair the network's ability to predict 3D body pose and shape. In contrast, edge-filtering is a simpler and more robust operation than segmentation, but is still able to retain important shape information from the RGB image. Thus, edge-images (concatenated with 2D joint heatmaps) can better bridge the synthetic-to-real domain gap, resulting in improved metrics on real test inputs (row 3).

Hierarchical architecture and reprojection loss on 2D joints samples. Figure 3 and rows 3-6 in Table 1 compare the independent and hierarchical distribution prediction architectures (f_{indep} and f_{hier}) presented in Section 3.4, both with and without the reprojection loss on sampled 2D joints ($\mathcal{L}_{\text{2D Samples}}$) from Section 3.5. When $\mathcal{L}_{\text{2D Samples}}$ is not applied, the shape and pose distributions predicted by both the independent and hierarchical network architectures do not consistently match the the input image, as evidenced by the significant gap between the visible 2D joint error computed using the distributions' modes versus samples drawn from the distributions (in rows 3 and 5 of Table 1) on both synthetic test data and SSP-3D [47]. This implies that the predicted distributions are overly uncertain about parts of the subject's body that are visible and unambiguous in the input image. The visualisations corresponding to the hierarchical architecture trained without $\mathcal{L}_{\text{2D Samples}}$ in Figure 3 (centre) further demonstrate that the predicted samples often do not match the input image, particularly at the extreme ends of the body. This results in significant undesirable per-vertex uncertainty over unambiguous body parts.

Applying $\mathcal{L}_{\text{2D Samples}}$ to the independent network f_{indep} partially alleviates the mismatch between inputs and predicted samples, as shown by Figure 3 (right) and row 4 in

Method	3DPW		
	MPJPE	MPJPE-SC	MPJPE-PA
HMR [19]	130.0	102.8	76.7
GraphCMR [26]	119.9	102.0	70.2
SPIN [25]	96.9	89.4	59.0
Pose2Mesh [8]	89.2	-	58.9
I2L-MeshNet [36]	93.2	77.5	57.7
Biggs *et al.* [3]	93.8	-	59.9
DaNet [65]	85.5	76.4	54.8
HybrIK [32]	**80.0**	-	**48.8**
HMR (unpaired) [19]	-	126.3	92.0
Kundu *et al.* [28]	153.4	-	89.8
STRAPS [47]	-	99.0	66.8
Sengupta *et al.* [48]	-	90.9	61.0
Ours w. Detectron2 [60]	96.2	84.7	59.2
Ours w. HRNet-W48 [54]	**84.9**	**73.0**	**53.6**

Table 2. Comparison with SOTA in terms of MPJPE, MPJPE-SC and MPJPE-PA (all mm) on 3DPW [58]. Methods in the top half require training images paired with 3D ground-truth, methods in the bottom half do not. For our method, we present metrics using both Detectron2 [60] and HRNet [54] as 2D joint detectors for proxy representation computation, to enable a fair comparison with past methods [47, 48] that used Detectron2.

Table 1, where the mode versus sample 2D joint error gap has reduced. However, training with $\mathcal{L}_{\text{2D Samples}}$ deteriorates the independent architecture's mode pose prediction metrics (MPJPE-SC and 2D Joint Err. Mode in row 3 vs 4 of Table 1) on both synthetic and real test data. This is because f_{indep} naively models each joint's relative rotation independently of its parents' rotations (Equation 5); however, to predict realistic human pose samples that match the visible input, each joint's rotation distribution *must* be informed by its parents. $\mathcal{L}_{\text{2D Samples}}$ attempts to force predicted samples to match the input despite this logical inconsistency, which causes a trade-off between mode and sample pose prediction metrics, particularly worsening MPJPE-SC.

In contrast, applying $\mathcal{L}_{\text{2D Samples}}$ to the hierarchical network f_{hier} improves metrics corresponding to both mode and sample predictions, as shown by row 6 in Table 1. Now, each SMPL joint's relative rotation distribution is conditioned on all its parents' distributions (Equation 6). Thus, $\mathcal{L}_{\text{2D Samples}}$ and $\mathcal{L}_{\text{NLL}}$ work in conjunction in enabling predicted hierarchical distributions (and samples) to match the visible input, while yielding improved 3D metrics. Figure 3 (left) exhibits such visually-consistent samples and demonstrates greater prediction uncertainty for ambiguous parts. Note that uncertainty can arise even without occlusion in a monocular setting, e.g. due to depth ambiguities [50, 52] as shown by the left arm samples in the last row of Figure 3. Further visual results are in the supplementary material.

Max. input set size	Method	SSP-3D PVE-T-SC
1	HMR [19]	22.9
	GraphCMR [26]	19.5
	SPIN [25]	22.2
	DaNet [65]	22.1
	STRAPS [47]	15.9
	Sengupta *et al.* [48]	15.2
	Ours	**13.6**
5	HMR [19] + Mean	22.9
	GraphCMR [26] + Mean	19.3
	SPIN [25] + Mean	21.9
	DaNet [65] + Mean	22.1
	STRAPS [47] + Mean	14.4
	Sengupta *et al.* [48] + Mean	13.6
	Sengupta *et al.* [48] + Prob. Comb.	13.3
	Ours + Mean	12.2
	Ours + Prob. Comb.	**12.0**

Table 3. Comparison with SOTA in terms of PVE-T-SC (mm) on SSP-3D [47]. Top half: single-input, bottom half: multi-input. Prob. Comb. denotes the multi-input probabilistic combination approach proposed in [48].

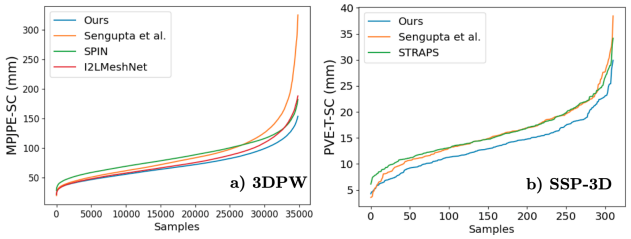

Figure 4. Comparison with SOTA using sorted per-sample distributions of a) MPJPE-SC on 3DPW and b) PVE-T-SC on SSP-3D.

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

Shape prediction. Table 3 evaluates 3D body shape metrics on SSP-3D [47] for single image inputs and multi-image input sets, which we evaluate using both mean and probabilistic combination methods from [48]. Our network surpasses the state-of-the-art [48], mainly due to our use of an edge-based proxy representation, instead of the silhouette-based representations used in [47] and [48]. These methods rely on accurate human silhouettes, which may be difficult to compute at test-time, as discussed in Section 5.1, while our method does not have such dependencies. However, our method may result in erroneous shape predictions when the subject is wearing loose clothing which obscures body shape, in which case the shape prediction over-estimates the subject's true proportions (see rows 1-2 in Figure 3).

Pose prediction. Table 2 evaluates 3D pose metrics on 3DPW [58]. Our method is competitive with the state-of-the-art and surpasses other methods that do not require 3D-labelled training images [47, 48, 28, 19]. Figure 4(a) shows that our method performs well for most test examples in 3DPW, even matching pose-focused approaches that do not

attempt to accurately predict diverse body shapes [36, 25]. However, some images in 3DPW contain significant occlusion, which can lead to noisy 2D joint heatmaps in the proxy representations, resulting in poor 3D pose metrics as shown by the right end of the curve in Figure 4(a).

Further quantitative comparison with other shape and pose distribution/multi-hypothesis prediction approaches is given in the supplementary material.

6. Conclusion

In this paper, we have proposed a probabilistic approach to the ill-posed problem of monocular 3D human shape and pose estimation, motivated by the fact that multiple 3D bodies may explain a given 2D image. Our method predicts a novel hierarchical matrix-Fisher distribution over relative 3D joint rotations and a Gaussian distribution over SMPL [33] shape parameters, from which we can sample any number of plausible 3D reconstructions. To ensure that the predicted distributions match the input image, we have implemented a differentiable rejection sampler to impose a loss between predicted 2D joint samples and ground-truth 2D joint coordinates. Our method is competitive with the state-of-the-art in terms of pose metrics on 3DPW, while surpassing the state-of-the-art for shape accuracy on SSP-3D.

Acknowledgements. We thank Dr. Yu Chen (Metail), Mr. Jim Downing (Metail), Dr. David Bruner (SizeStream) and Dr. Delman Lee (TAL Apparel) for providing body shape evaluation data and supporting this research.

Figure 5. Examples of synthetic training and validation data rendered on-the-fly during model training. Synthetic RGB images are converted into edge-filtered images and 2D joint heatmaps, which act as the input to the distribution prediction network presented in the main manuscript. The synthetic RGB images are computationally-cheap and far from photorealistic. However, edge detection [6] is able to significantly bridge the synthetic-to-real domain gap, as can be seen by comparing the synthetic edge-images with real edge-images in Figures 6 and 7 of the supplementary material.

Supplementary Material: Hierarchical Kinematic Probability Distributions for 3D Human Shape and Pose Estimation from Images in the Wild

Section 7 in this supplementary material contains implementation details, particularly regarding synthetic training data generation and per-vertex uncertainty visualisation. Section 8 discusses qualitative results on the SSP-3D [47] and 3DPW [58] datasets, and compares distribution predictions on images with versus without artificial occlusions. Table 5 compares several recent multi-hypothesis 3D human shape and pose estimation approaches.

7. Implementation Details

7.1. Synthetic Training Data

Our shape and pose distribution prediction neural networks are trained using synthetic training data, consisting of edge-and-joint-heatmap inputs paired with ground truth SMPL [33] shape and pose parameters. Inputs are rendered on-the-fly during model training using randomly sampled camera extrinsics, lighting, backgrounds and clothing textures. Examples of synthetic training and validation data are given in Figure 5. Note how each body pose may be paired with a different body shape, clothing, camera and background, as well as occlusion and noise augmentations. Thus, we are able to render highly diverse training data on-the-fly during training, enabling the network to see a new pose/shape/clothing/camera/background combination in each training iteration.

Our synthetic RGB images (Figure 5) are computationally cheap but clearly far from photorealistic, resulting in a large synthetic-to-real domain gap. However, simple edge detection [6] is able to significantly reduce this gap [7], motivating the use of edge-filtered images as part of our input proxy representation. We found that noisy edge detections (as seen in Figure 5) retained sufficient visual shape and pose information, and efforts to produce clean edge-images (e.g. hysteresis-based edge tracking or further hyperparameter tuning) did not improve performance.

The required body shape, pose, clothing and backgrounds are obtained as follows. For training, ground-truth SMPL 3D joint rotation matrices are sampled from the training splits of 3DPW [58] and UP-3D [29], as well as Human3.6M [16] subjects 1, 5, 6, 7 and 8, giving a total of 91106 training poses. Validation poses are sampled from the 3DPW/UP-3D validation splits and Human3.6M subjects 9 and 11, resulting in 33347 validation poses. SMPL body shape parameters are randomly sampled from $\mathcal{N}(\beta_i; 0, 1.25^2)$ for $i = 1, ..., 10$ [47]. RGB clothing textures for the SMPL body mesh are selected from SURREAL [57] and MultiGarmentNet [2], resulting in 917 training textures and 108 validation textures. Backgrounds are obtained from LSUN [62], which contains a collection of diverse indoor and outdoor scenes. We sample from 397582 different training backgrounds and 3000 different validation backgrounds. Note that background training images may contain other humans, which is intentional and essential for robustness against test images with multiple people. The network learns to focus on the person corresponding to the input joint heatmaps and ignore persons in the background.

Hyperparameter	Value
Shape parameter sampling mean	0
Shape parameter sampling std.	1.25
Cam. translation sampling mean	(0, -0.2, 2.5) m
Cam. translation sampling var.	(0.05, 0.05, 0.25) m
Cam. focal length	300.0
Lighting ambient intensity range	[0.4, 0.8]
Lighting diffuse intensity range	[0.4, 0.8]
Lighting specular intensity range	[0.0, 0.5]
Bounding box scale factor range	[0.8, 1.2]
Proxy representation dimensions	256×256 pixels

Table 4. List of hyperparameter values associated with synthetic training data generation.

Textured SMPL meshes are rendered with Pytorch3D [43], using a perspective camera model and Phong shading. Camera and lighting parameters are randomly sampled, with sampling hyperparameters given in Table 4. Generated images are cropped around the rendered body using a square bounding box, where the bounding box size is randomly scaled by a factor in range (0.8, 1.2).

To further bridge the gap synthetic-to-real gap, we implement random occlusion, body part removal, 2D joint removal and 2D joint noise augmentations during training. Hyperparameters associated with data augmentations are given in Table 6.

7.2. Visualisation of Per-Vertex Uncertainty

Figures 6, 7 and 8 in this supplementary material, as well as several figures in the main manuscript, visualise per-vertex 3D location uncertainties corresponding to the predicted shape and 3D joint rotation distributions. These are computed by i) sampling 100 shape parameter vectors and relative 3D joint rotations (for the entire kinematic tree) from the predicted distributions, ii) passing each of these samples through the SMPL function [33] to get the corresponding vertex meshes, iii) computing the mean location of each vertex over all the samples and iv) determining the average Euclidean distance from the sample mean for each vertex over all the samples, which is ultimately visualised in the vertex scatter plots as a measure of per-vertex 3D location uncertainty.

Method	3DPW								SSP-3D			
	MPJPE				MPJPE-PA				PVE-T-SC			
Number of Samples:	1	5	10	25	1	5	10	25	1	5	10	25
Biggs *et al*. [3]	93.8	82.2	79.4	75.8	59.9	57.1	56.6	55.6	-	-	-	-
Sengupta *et al*. [48]	97.1	95.8	93.1	89.7	61.1	59.4	58.2	56.5	15.2	14.8	13.6	11.9
ProHMR [27]	-	-	-	-	59.8	56.5	54.6	52.4	-	-	-	-
Ours (Independent) w. HRNet [54]	88.3	85.0	82.6	78.5	56.6	54.5	52.8	50.2	13.9	12.9	12.0	10.3
Ours (Hierarchical) w. HRNet [54]	**84.9**	**81.6**	**79.0**	**75.1**	**53.6**	**51.4**	**49.6**	**47.0**	**13.6**	**12.3**	**11.3**	**9.8**

Table 5. Comparison with other 3D human shape and pose distribution/multi-hypothesis estimation methods. Following Biggs *et al*. [3], we report body shape (PVE-T-SC) and pose (MPJPE and MPJPE-PA) metrics computed using the minimum error sample out of a set of n predicted samples for each test image, for $n \in \{1, 5, 10, 25\}$. This is motivated by the fact that the single ground-truth 3D body only represents *one plausible 3D solution out of many* (for ambiguous images), which may not be the same as the mode of our predicted shape and pose distribution. The improvement in 3D shape and pose metrics with increasing number of samples shows that our predicted distribution is able to model the 3D ground-truth as a possible sample.

Augmentation	Hyperparameter	Value
Body part occlusion	Occlusion probability	0.1
2D joints L/R swap	Swap probability	0.1
Half-image occlusion	Occlusion probability	0.05
2D joints removal	Removal probability	0.1
2D joints noise	Noise range	[-8, 8] pixels
Occlusion box	Probability, Size	0.5, 48 pixels

Table 6. List of synthetic training data augmentations and their associated hyperparameter values. Body part occlusion uses the 24 DensePose [15] parts. Joint L/R swap is done for shoulders, elbows, wrists, hips, knees, ankles.

8. Qualitative Results

Figure 7 presents results on artificially occluded images from SSP-3D [47]. In particular, note that i) occluded/invisible body parts result in increased 3D location uncertainty for corresponding vertices and ii) 3D body samples from the predicted distributions match the visible body parts in the 2D image, while invisible body part samples are more diverse. However, occluded sample diversity is still somewhat limited and samples tend to be clustered around the mode predictions, which is a weakness of our method. This may be alleviated by predicting multi-modal distributions over 3D shape and pose in future work. Figure 7 also illustrates our method's ability to predict a range of body shapes, owing to the synthetic training framework used.

Figure 6 presents results on the test split of 3DPW [58]. Again, note the increased uncertainty and sample diversity for occluded and out-of-frame body parts, and the reprojection consistency between predicted samples and the visible bodies in the images. Results on 3DPW highlight another key challenge for future work: when faced with baggy/loose clothing, our method tends to over-estimate the subject's body proportions. This is because our synthetic training data does not model the shape of clothing on the human body surface, but only its texture. Future work could focus on using synthetic *clothed* humans for training.

Figure 8 compares shape and pose distribution predictions on images from SSP-3D with versus without artificial occlusions, further corroborating that ambiguous parts result in greater uncertainty and more diverse 3D samples. However, it is again apparent that sample diversity for highly ambiguous parts is more limited than expected, as samples tend to be closely clustered around the mode prediction.

Note that uncertainty does not only arise from occlusion - depth ambiguities are prevalent when estimating 3D pose from a monocular 2D image [50, 52]. This is demonstrated in the non-occluded images in Figure 8 (left), by the left arm samples in rows 1 and 5 and the right arm in row 4.

Figure 6. 3D reconstruction samples and per-vertex uncertainties corresponding to shape and relative 3D joint rotation distributions predicted from 3DPW images[58]. The selected images exhibit self-occlusion and out-of-frame body parts, which result in greater 3D location uncertainty for vertices belonging to ambiguous parts.

Figure 7. 3D reconstruction samples and per-vertex uncertainties corresponding to shape and relative 3D joint rotation distributions predicted from SSP-3D images[47]. The images are artificially occluded, resulting in greater 3D location uncertainty for vertices belonging to ambiguous parts.

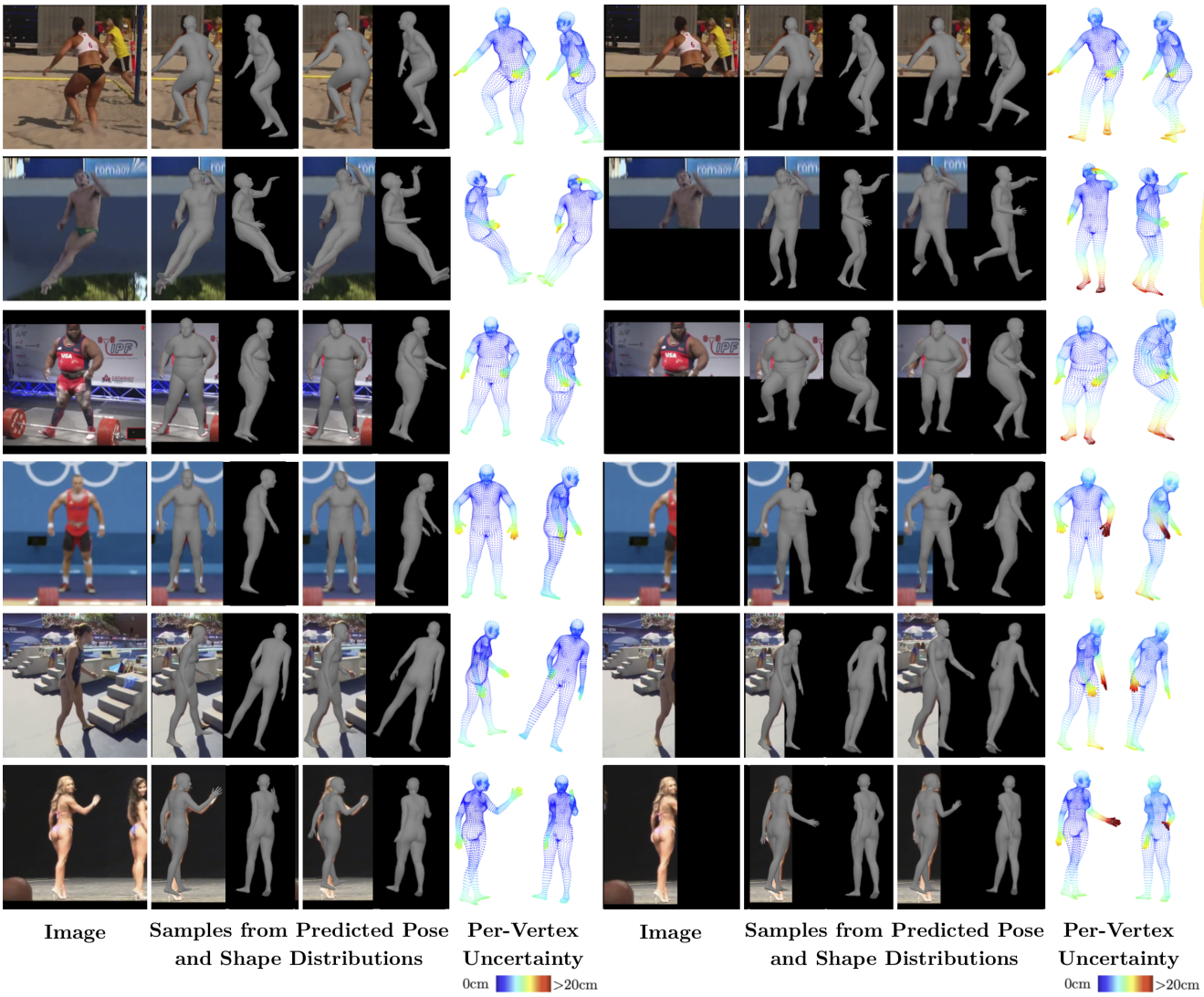

Figure 8. Comparison between 3D samples and per-vertex uncertainties obtained using artificially occluded versus non-occluded input images from SSP-3D [47]. Ambiguous parts have greater prediction uncertainty.

References

[1] 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. 2

[2] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3D people from images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 6, 10

[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. 2, 8, 11

[4] Christopher M. Bishop. Mixture density networks. Technical report, 1994. 2

[5] 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. 2

[6] John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 8(6):679–698, 1986. 4, 6, 9, 10

[7] J. Charles, S. Bucciarelli, and R. Cipolla. Real-time screen reading: reducing domain shift for one-shot learning. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. 2, 10

[8] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 8

[9] Kiam Choo and D.J. Fleet. People tracking using hybrid monte carlo filtering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 321–328 vol.2, 2001. 2

[10] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000. 2

[11] Thomas D. Downs. Orientation statistics. *Biometrika*, 59(3):665–676, 12 1972. 1, 3

[12] 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. 1, 2, 4

[13] Igor Gilitschenski, Roshni Sahoo, Wilko Schwarting, Alexander Amini, Sertac Karaman, and Daniela Rus. Deep orientation uncertainty learning based on a bingham loss. In *International Conference on Learning Representations*, 2020. 3

[14] 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. 1, 2

[15] 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. 2, 7, 11

[16] 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. 2, 6, 7, 10

[17] Ehsan Jahangiri and Alan L. Yuille. Generating multiple diverse hypotheses for human 3D pose consistent with 2D joint detections. In *IEEE International Conference on Computer Vision (ICCV) Workshops (PeopleCap)*, 2017. 2

[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. 2

[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. 1, 2, 4, 8

[20] John T. Kent, Asaad M. Ganeiber, and Kanti V. Mardia. A new method to simulate the Bingham and related distributions in directional data analysis with applications, 2013. 5

[21] C. G. Khatri and K. V. Mardia. The Von Mises-Fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):95–106, 1977. 1, 3

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 6

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014. 5

[24] 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. 2

[25] 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. 1, 2, 8, 9

[26] 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. 1, 2, 8

[27] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 11

[28] 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. 8

[29] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the Peo-

ple: Closing the loop between 3D and 2D human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6, 10

[30] T. Lee. Bayesian attitude estimation with the matrix fisher distribution on so(3). *IEEE Transactions on Automatic Control*, 63(10):3377–3392, 2018. 3, 4

[31] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3D human pose estimation with mixture density network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[32] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021. 8

[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. 1, 2, 3, 9, 10

[34] K. V. Mardia and P. E. Jupp. *Directional statistics*. Wiley, 2000. 1, 3, 5

[35] David Mohlin, Josephine Sullivan, and Gérald Bianchi. Probabilistic orientation estimation with matrix fisher distributions. In *Advances in Neural Information Processing Systems*, volume 33, 2020. 1, 3, 5

[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. 1, 2, 8, 9

[37] Tuomas P. Oikarinen, Daniel C. Hannah, and Sohrob Kazerounian. GraphMDN: Leveraging graph structure and deep learning to solve inverse problems. *CoRR*, abs/2010.13668, 2020. 2

[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. 1, 2

[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. 2

[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. 1, 2

[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. 1, 2, 4

[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. 3

[43] 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. 6, 10

[44] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. 2

[45] 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. 2

[46] 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 (CVPR)*, June 2020. 2

[47] 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. 1, 2, 4, 6, 7, 8, 10, 11, 13, 14

[48] 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. 2, 4, 6, 8, 11

[49] C. Sminchisescu, A. Kanaujia, Zhiguo Li, and D. Metaxas. Discriminative density propagation for 3D human motion estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 390–397 vol. 1, 2005. 2

[50] Cristian Sminchisescu and Bill Trigg. Covariance scaled sampling for monocular 3D body tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001. 2, 8, 11

[51] Cristian Sminchisescu and Bill Trigg. Hyperdynamics importance sampling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2002. 2

[52] Cristian Sminchisescu and Bill Trigg. Kinematic jump processes for monocular 3D human tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 2, 8, 11

[53] 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)*, 2019. 1, 2, 6

[54] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 4, 8, 11

[55] 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. 1, 2

[56] 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. 1, 2

[57] 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. 2, 6, 10

[58] 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. 2, 6, 7, 8, 10, 11, 12

[59] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3D human pose estimation with normalizing flows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[60] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 8

[61] 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. 2

[62] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6, 10

[63] 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. 2

[64] 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. 2

[65] 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. 1, 2, 8