HuManiFlow: Ancestor-Conditioned Normalising Flows on SO(3) Manifolds for Human Pose and Shape Distribution Estimation Supplementary Material

Akash Sengupta Ignas Budvytis Roberto Cipolla
University of Cambridge
{as2562, ib255, rc10001}@cam.ac.uk

This supplementary material contains further details regarding our probabilistic pose and shape prediction method, and presents additional quantitative and qualitative results and comparisons with other methods. Section A describes the model architecture and synthetic training data. It also validates our approach for point estimate computation, and investigates our radial tanh transform for compact distribution support. Section B discusses evaluation datasets and metrics, cropped evaluation dataset generation and directional variance visualisations. Finally, Section C presents comparisons with other methods.

A. Implementation Details

A.1. Model architecture

An overview of our model architecture is provided in Figure 2 of the main manuscript. Further details regarding the CNN encoder, shape/global rotation/camera MLP and per-body-part normalising flow modules are provided below.

We use a ResNet-18 CNN encoder [9], which takes a proxy representation input $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ and outputs a feature vector $\boldsymbol{\phi} \in \mathbb{R}^{512}$. The proxy representation consists of an edge-image and 2D keypoint heatmaps stacked along the channel dimension, with height $H = 256$, width $W = 256$ and channels $C = 18$. The choices of proxy representation and CNN encoder follow [24].

The input features $\boldsymbol{\phi}$ are passed through the shape/global rotation/camera MLP, which outputs the parameters of a Gaussian distribution over SMPL [21] shape, $\boldsymbol{\mu}_{\beta}, \boldsymbol{\sigma}^2_{\beta} \in \mathbb{R}^{10}$, as well as deterministic estimates of weak-perspective camera parameters $\boldsymbol{\pi} = [s, t_x, t_y] \in \mathbb{R}^3$ and the global body rotation $\mathbf{R}_{\text{glob}} \in SO(3)$. The latter is predicted using the continuous 6D rotation representation proposed by [29], then converted to a rotation matrix. The shape/global rotation/camera MLP has 1 hidden layer with 512 nodes and ELU activation [3], and an output layer with 29 nodes.

For each SMPL body-part $i \in \{1, \dots, 23\}$, our method outputs a normalising flow distribution over the body-part's relative rotation $\mathbf{R}_i \in SO(3)$. This is conditioned on the input features $\boldsymbol{\phi}$, camera estimate $\boldsymbol{\pi}$, global rotation estimate $\mathbf{R}_{\text{glob}}$, a shape vector sample $\boldsymbol{\beta}$ and ancestor body-part rotation samples $\{\mathbf{R}_j\}_{j \in A(i)}$, where $A(i)$ denotes the ancestors of body-part i in the SMPL kinematic tree. $\boldsymbol{\beta}$ is sampled differentiably from $\mathcal{N}(\boldsymbol{\mu}_{\beta}(X), \boldsymbol{\sigma}^2_{\beta}(X))$ using the re-parameterisation trick [13]. $\{\mathbf{R}_j\}_{j \in A(i)}$ are differentiably sampled from their own respective normalising flow rotation distributions. The conditioning variables $\{\boldsymbol{\phi}, \boldsymbol{\pi}, \mathbf{R}_{\text{glob}}, \boldsymbol{\beta}, \{\mathbf{R}_j\}_{j \in A(i)}\}$ are aggregated into a context vector $\mathbf{c}_i \in \mathbb{R}^{64}$ using a context generation MLP for each body-part i, as shown in Figure 2 of the main manuscript. Each context generation MLP has 1 hidden layer with 256 nodes and ELU activation [3], and an output layer with 64 nodes. Note: in Eqns. 8 and 10 in the main manuscript, we notationally replaced the conditioning variables $\boldsymbol{\phi}, \boldsymbol{\pi}$ and $\mathbf{R}_{\text{glob}}$ with $\mathbf{X}$ for simplicity, because each of these are deterministically obtained as functions of $\mathbf{X}$.

The conditional normalising flow distribution over each

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]
Body-part occlusion probability	0.1
2D joints L/R swap probability	0.1
Half-image occlusion probability	0.05
Extreme crop probability	0.1
2DKP occlusion probability	0.1
2DKP noise range	[-8, 8] pixels

Table 1. List of hyperparameter values associated with synthetic training data generation and augmentation. Body-part occlusion uses the 24 DensePose [8] parts. Joint L/R swap is done for shoulders, elbows, wrists, hips, knees, ankles.

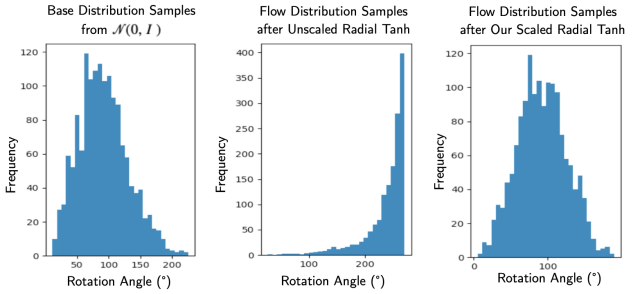

Figure 1. Comparison between the "unscaled" radial tanh transform proposed by [7] (Eqn. 3) and our "scaled" version (Eqn. 2), in terms of sample rotation angles (or axis-angle vector magnitudes) from a randomly-initialised (i.e. un-trained) normalising flow. The unscaled transform pushes samples from the base distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ towards the boundary of the desired support ball $B_r(\mathbf{0})$. Here, r is set to 1.5π rad (i.e. 270°). This results in unstable training, as shown in Figure 2. Our scaled transform mitigates this behaviour.

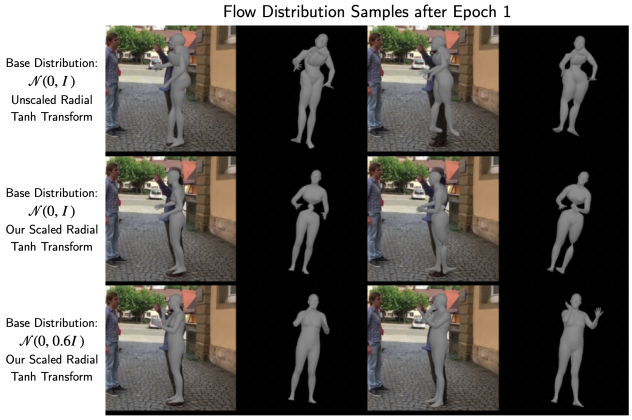

Figure 2. Effect of our "scaled" radial tanh transform (Eqn. 2), and reduced-variance base distribution, on SMPL pose and shape samples after training for 1 epoch. Samples using $\mathcal{N}(\mathbf{0}, \mathbf{I})$ for base distributions and the "unscaled" radial tanh transform proposed by [7] (Eqn. 3) are highly unhuman (row 1). Samples become much more realistic with the scaled transform and reduced-variance base distribution (row 2 and 3). This improves training stability.

body-part's rotation, $p_{SO(3)}(\mathbf{R}_i|\mathbf{c}_i)$, is formed by pushing a flow distribution over the corresponding axis-angle vector, $p_{\mathbb{R}^3}(\mathbf{v}_i|\mathbf{c}_i)$ for $\mathbf{v}_i \in \mathbb{R}^3 \cong \mathfrak{so}(3)$, onto $SO(3)$ using the exp map (detailed by Eqn. 6 in the main manuscript). We use Linear Rational Spline (LRS) normalising flows [6], which transform a simple base distribution into a complex density function with a series of LRS coupling layer diffeomorphisms. Specifically, let $\mathbf{z}_{k-1} \in \mathbb{R}^D$ be the input variable to the k-th coupling layer f_k^{LRS}, and let $\mathbf{z}_k \in \mathbb{R}^D$ be the output, such that $\mathbf{z}_k = f_k^{\text{LRS}}(\mathbf{z}_{k-1})$. The coupling layer [4, 5] splits the input into two parts $\mathbf{z}_{k-1}^{0:d}$ and $\mathbf{z}_{k-1}^{d:D}$. Then, the output variable is determined by

$$\begin{aligned} \mathbf{z}_k^{0:d} &= \mathbf{z}_{k-1}^{0:d} \\ \mathbf{z}_k^{d:D} &= g(\mathbf{z}_{k-1}^{d:D}; \mathbf{w}(\mathbf{z}_{k-1}^{0:d})) \end{aligned} \tag{1}$$

where $g(\,.\,; \mathbf{w}(\mathbf{z}_{k-1}^{0:d}))$ is an *element-wise* bijective and differentiable function (i.e. a diffeomorphism), whose parameters $\mathbf{w}$ depend on the first half of the input $\mathbf{z}_{k-1}^{0:d}$. Note that the Jacobian of f_k^{LRS} is lower-triangular, and thus $\det J_{f_k^{\text{LRS}}}$ is easily-computed as the product of the diagonal terms of $J_{f_k^{\text{LRS}}}$. For an LRS coupling layer [6], g is an element-wise spline transform. Each spline segment is a linear rational function of the form $\frac{ax+b}{cx+d}$. The parameters of g, $\mathbf{w}(\mathbf{z}_{k-1}^{0:d})$, are the parameters of each segment's linear rational function, and the locations of each segment's endpoints (or knots). These are obtained by passing $\mathbf{z}_{k-1}^{0:d}$ through an MLP. LRS coupling layers are able to model significantly more complex distributions [6] than affine [5] or additive [4] coupling layers with the same number of layers composed together.

For each body-part i, $p_{\mathbb{R}^3}(\mathbf{v}_i|\mathbf{c}_i)$ is implemented as an LRS-NF composed of 3 LRS coupling layer transforms, with a permutation following each coupling layer. Each layer's spline parameters are output by an MLP with 3 hidden layers that have 32 nodes each, and ELU [3] activations.

A.2. Radial tanh and sample angle regularisation

We must ensure that $p_{\mathbb{R}^3}(\mathbf{v}_i|\mathbf{c}_i)$ has compact support, i.e. $p_{\mathbb{R}^3}(\mathbf{v}_i|\mathbf{c}_i) = 0$ for $\mathbf{v}_i \notin B_r(\mathbf{0})$ where $B_r(\mathbf{0})$ is an open ball of radius $\pi < r < 2\pi$. We choose $r = 1.5\pi$. Towards this end, we use a radial tanh transform [7], $t : \mathbb{R}^3 \to B_r(\mathbf{0})$ as the last layer of each body-part's normalising flow transform, as discussed in the main manuscript. This is reproduced here for convenience:

$$t(\mathbf{x}) = r \tanh\left(\frac{\|\mathbf{x}\|}{r}\right) \frac{\mathbf{x}}{\|\mathbf{x}\|}. \tag{2}$$

This is slightly different to the original transform proposed in [7], which is given by

$$t'(x) = r \tanh\left(\|\mathbf{x}\|\right) \frac{\mathbf{x}}{\|\mathbf{x}\|} \tag{3}$$

i.e. the argument of the $\tanh$ in our transform is scaled by $1/r$. This scaling is highly beneficial for training the body pose normalising flows. To see why, consider the behaviour of t and t' when $\|\mathbf{x}\|$ is small, such that $\tanh \|\mathbf{x}\| \approx \|\mathbf{x}\|$. Then, $t(\mathbf{x}) \approx \mathbf{x}$ while $t'(\mathbf{x}) \approx r\mathbf{x}$. Notably, t does not significantly affect points which have small magnitude, and thus are already well within the desired support $B_r(\mathbf{0})$. In contrast, t' increases the magnitude of these points by a factor of r, pushing them towards the boundary of $B_r(\mathbf{0})$. This is illustrated by Figure 1, which visualises the histogram of sample magnitudes drawn from a base distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and the histogram of magnitudes after these samples are passed through a randomly-initialised flow ending in a radial tanh transform. Note that sample magnitudes correspond to rotation angles, as samples from $p_{\mathbb{R}^3}(\mathbf{v}_i|\mathbf{c}_i)$ are axis-angle vectors. Comparing the sample magnitude histogram for the

Method	3DPW				3DPW 70% Cropped				3DPW 50% Cropped			
	MPJPE (mm)		MPJPE-PA (mm)		MPJPE (mm)		MPJPE-PA (mm)		MPJPE (mm)		MPJPE-PA (mm)	
	Point	Sample Min.	Point	Sample Min.	Point	Sample Min.	Point	Sample Min.	Point	Sample Min.	Point	Sample Min.
HMR [12]	130.0	-	76.7	-	177.4	-	96.6	-	214.6	-	120.2	-
GraphCMR [16]	119.9	-	70.2	-	120.8	-	74.5	-	205.2	-	119.7	-
SPIN [15]	96.9	-	59.0	-	108.5	-	63.9	-	196.7	-	130.5	-
PARE [14]	74.5	-	46.5	-	**80.0**	-	**50.6**	-	121.8	-	79.7	-
HybrIK [19]	**74.1**	-	**45.0**	-	91.9	-	60.7	-	187.5	-	153.0	-
3D Multibodies [2]	93.8	74.6 (20.5%)	59.9	48.3 (19.4%)	110.5	80.9 (**26.8%**)	67.7	51.1 (24.5%)	190.6	98.4 (**48.4%**)	120.3	64.7 (**46.2%**)
Sengupta *et al.* [25]	97.1	84.4 (13.1%)	61.1	52.1 (14.7%)	99.8	86.1 (13.7%)	62.7	52.2 (16.7%)	144.7	125.5 (13.3%)	93.6	76.1 (18.7%)
ProHMR [17]	97.0	81.5 (16.0%)	59.8	48.2 (19.4%)	99.4	84.1 (15.4%)	62.1	50.0 (19.5%)	143.8	123.3 (14.3%)	85.4	68.8 (19.4%)
HierProbHuman [24]	84.9	70.9 (16.5%)	53.6	43.8 (18.3%)	94.2	78.4 (16.8%)	61.6	49.5 (19.6%)	126.9	101.8 (19.8%)	87.0	67.7 (22.2%)
HuManiFlow	**83.9**	**65.1** (**22.4%**)	**53.4**	**39.9** (**25.3%**)	**93.5**	**71.6** (23.4%)	**60.7**	**44.6** (**26.5%**)	**116.4**	86.9 (25.3%)	**78.2**	**54.9** (29.8%)

Table 2. Comparison between recent deterministic (top half) and probabilistic (bottom half) pose and shape predictors in terms of accuracy on the 3DPW dataset [27], as well as 50% and 70% cropped versions of 3DPW (see Section B.2 for cropping details). %s are decreases in MPJPE(-PA) from the point-estimate to the minimum sample value computed over 100 samples. Our method, HuManiFlow, is more accurate than all current probabilistic methods. Point estimates from HuManiFlow are competitive with the state-of-the-art deterministic methods, particularly on more ambiguous and challenging cropped images.

"unscaled" transform [7] t' and our "scaled" version t demonstrates that t' pushes points towards the boundary of $B_r(\mathbf{0})$ (with $r = 1.5\pi$).

Large sample rotation angles from un-trained (i.e. randomly-initialised) flow distributions lead to unstable training. This is because human body-parts rarely have large rotations in natural poses, which is particularly true for torso joints in the SMPL kinematic tree. We empirically found that the flow models struggled to recover from initially too-large rotation angles during training, when using the unscaled transform t'. This is shown in Figure 2, where t' results in extremely unhuman pose samples (row 1) after 1 training epoch, while our scaled transform t gives more realistic samples. In other words, t acts as a rotation angle regulariser, making use of prior domain knowledge about human body-parts often having small rotations in natural poses.

To further regularise sample rotation angles, we replace the typical base distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with $\mathcal{N}(\mathbf{0}, 0.6\mathbf{I})$, which has reduced variance. This results in reasonable pose samples during early training, as shown by Figure 2, row 3.

A.3. Point estimate validity

Eqn. 12 in the main manuscript describes our approach to obtaining a point estimate $(\{\mathbf{R}_i^*\}_{i=1}^{23}, \boldsymbol{\beta}^*)$ from the complex joint distribution over SMPL pose and shape $p_{\text{joint}}(\{\mathbf{R}_i\}_{i=1}^{23}, \boldsymbol{\beta}|\mathbf{X})$ predicted by our method. The point estimate is not, in general, the actual mode of p_{joint}. However, we empirically verify that $(\{\mathbf{R}_i^*\}_{i=1}^{23}, \boldsymbol{\beta}^*)$ typically has high likelihood under p_{joint}, using the test set of 3DPW [27]. To do so, we compute the log-likelihood of the point estimate for each test input, $p_{\text{joint}}(\{\mathbf{R}_i^*\}_{i=1}^{23}, \boldsymbol{\beta}^*|\mathbf{X})$, as well as the maximum sample log-likelihood $\max_{n=1,\dots,N} p_{\text{joint}}(\{\mathbf{R}_i^n\}_{i=1}^{23}, \boldsymbol{\beta}^n|\mathbf{X})$ for $N = 1000$ total samples. The maximum sample log-likelihood is subtracted from the point estimate log-likelihood for each test input, and the histogram of these log-likelihood deltas

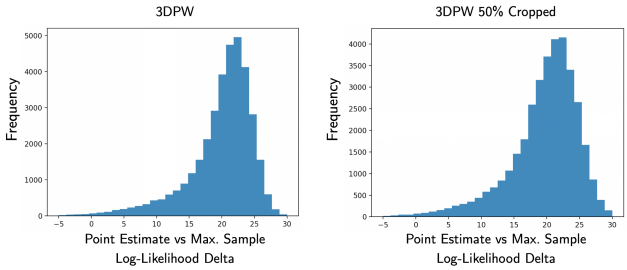

Figure 3. Histogram of differences between point estimate log-likelihoods and maximum sample log-likelihoods for 1000 samples, computed on the test set of 3DPW [27] and its cropped version. Point estimates generally have higher log-likelihoods than the most likely samples, as the log-likelihood differences are typically positive, confirming that the approximate method for obtaining point estimates presented in Eqn. 12 of the main manuscript is suitable.

is visualised in Figure 3. The point estimates generally have higher log-likelihood under p_{joint} than the most likely samples, corroborating their validity and usefulness when a single pose and shape solution is required. A qualitative comparison between point estimates from our approach and pose and shape predictions from the state-of-the-art deterministic methods [14, 19] is given in Figure 4.

A.4. Synthetic training data

We train our pose and shape distribution prediction method using the synthetic training data generation pipeline proposed by [24]. A brief overview is given below, but we refer the reader to [24] for details. Moreover, hyperparameters related to synthetic data generation and augmentation are given in Table 1.

Synthetic edge-and-keypoint-heatmap proxy representations are rendered on-the-fly during training, using ground-truth SMPL [21] pose parameters, and randomly sampled

SMPL shape parameters, camera extrinsics, lighting, backgrounds and clothing. Ground-truth pose parameters are obtained from the training splits of 3DPW [27], UP-3D [18] and Human3.6M [10], giving a total of 91106 training poses, as well as 33347 validation poses from the corresponding validation splits. SMPL shape parameters are sampled from $\mathcal{N}(\boldsymbol{\beta}; \mathbf{0}, 1.25\mathbf{I})$. RGB clothing textures are obtained from SURREAL [26] and MultiGarmentNet [1], which contain 917 training textures and 108 validation textures. Random backgrounds are obtained from a subset of LSUN [28] with 397582 training backgrounds and 3000 validation backgrounds. On-the-fly rendering of training inputs is done using Pytorch3D [22], with a perspective camera model and Phong shading. Camera and lighting parameters are randomly sampled, with hyperparameters given in Table 1.

To bridge the synthetic-to-real domain gap, synthetic proxy representations are augmented using random body-part occlusion, 2D keypoint occlusion, noise and swapping, and extreme cropping, as detailed in Table 1.

B. Evaluation Details

B.1. Datasets and metrics

The 3DPW [27] and SSP-3D [23] datasets are used for our ablation studies and comparison with other human pose and shape estimation approaches. The test set of 3DPW consists of 35515 images of 2 subjects paired with ground-truth SMPL parameters and 2D keypoint locations. SSP-3D consists of 311 images of 62 subjects with diverse body shapes, paired with pseudo-ground-truth SMPL parameters and 2D keypoint locations.

As discussed in Section 4 of the main manuscript, we use mean-per-joint-position-error (MPJPE) and MPJPE after Procrustes analysis (MPJPE-PA) to evaluate the accuracy of our method. Both MPJPE and MPJPE-PA are in units of mm. Following [12, 15–17], MPJPE and MPJPE-PA are computed using the 14 LSP joint convention [11]. Sample-input consistency is quantified using 2D keypoint reprojection error (or 2DKP Error) between GT *visible* 2DKPs and 2DKPs computed from predicted samples, for which we use the 17 COCO keypoint convention [20]. 2DKP Error is in units of pixels, assuming a 256×256 input image. Sample diversity is measured using the spread (i.e. average Euclidean distance from the mean) of 3D visible/invisible keypoints, which is denoted as 3DKP Spread. This is also computed using the 17 COCO keypoints. 3DKP Spread is in units of mm. We recognise that 3DKP Spread is flawed as a diversity metric, since the average Euclidean distance from the mean of 3DKPs may be too simplistic to accurately reflect the diversity of 3D body pose (i.e. body-part rotation) samples, particularly when evaluating highly complex multi-modal distributions. Future work can investigate improved diversity metrics for body pose distributions.

Dataset	**Method**	*Consistency* 2DKP Error Point / Samples	*Diversity* 3DKP Spread Vis. / Invis.
3DPW 70% Cropped	Sengupta *et al.* [25]	7.6 / 9.9	39.7/ 97.2
	3D Multibodies [2]	8.1 / 11.7	66.5 / 125.9
	ProHMR [17]	8.1 / 9.2	32.0 / 60.1
	HierProbHuman [24]	**7.2** / 9.5	41.8 / 102.3
	HuManiFlow	**7.2** / **8.6**	41.9 / 116.9
SSP-3D 70% Cropped	Sengupta *et al.* [25]	9.8 / 14.3	60.2 / 131.6
	3D Multibodies [2]	10.5 / 15.1	85.1 / 160.1
	ProHMR [17]	9.0 / 10.2	37.7 / 64.4
	HierProbHuman [24]	7.0 / 9.8	55.0 / 107.1
	HuManiFlow	**6.9** / **8.6**	46.9 / 123.3

Table 3. Comparison between probabilistic pose and shape predictors in terms of sample-input consistency and sample diversity on 70% cropped versions of 3DPW [27] and SSP-3D [23]. Our method, HuManiFlow, yields the most input-consistent samples (lowest visible 2DKP error) with reasonable diversity (3DKP spread).

B.2. Cropped dataset generation

To evaluate our method on highly ambiguous and challenging test inputs, we generate cropped versions of 3DPW [27] and SSP-3D [23]. Cropped test images are computed from the (already pre-processed) full-view test images by (i) centering at approximately the midpoint of the subject's torso, (ii) taking a square crop with dimensions given by $\alpha\%$ of the full-view image dimensions 256×256, and (iii) resizing back to 256×256. In the main manuscript, we used $\alpha = 50\%$ for all experiments with cropped data. However, the cropping percentage may be varied to evaluate our method on images with different levels of ambiguity. In Section C of this supplementary material, we present additional results with $\alpha = 70\%$. Examples of 50% and 70% cropped test images are given in Figure 6.

B.3. Directional variance visualisation

Figures 1 and 3 in the main manuscript, and Figure 6 in this supplementary material, visualise the per-vertex directional variance of samples drawn from predicted SMPL pose and shape distributions. For a given input image, per-vertex directional variance is computed by (i) drawing N samples from the predicted distribution $\{\{\mathbf{R}_i^n\}_{i=1}^{23}, \boldsymbol{\beta}^n\}_{n=1}^N$, (ii) passing each of these through the SMPL [21] model to obtain N vertex meshes $\{\mathbf{V}^n\}_{n=1}^N$ (where each $\mathbf{V}^n \in \mathbb{R}^{6890\times 3}$) and (iii) computing the variance (more specifically, the standard deviation) of each vertex along each of the x/y/z directions, or axes. We use $N = 100$. Note that the coordinate axes are aligned with the image plane, such that the x-axis represents the horizontal direction on the image, the y-axis represents the vertical direction on the image and the z-axis represents depth perpendicular to the image.

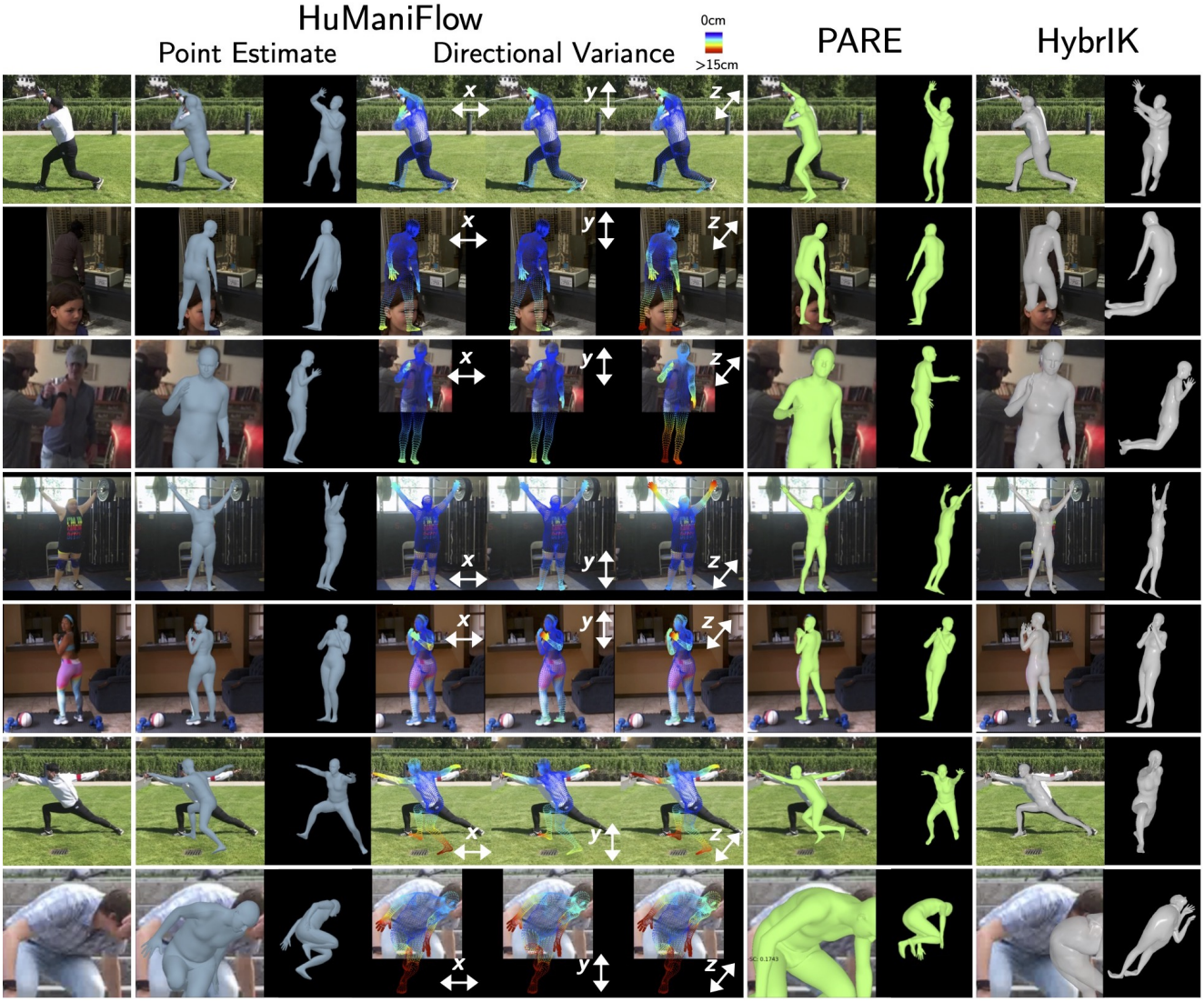

Figure 4. Qualitative comparison between point estimates from our probabilistic method (HuManiFlow) and the state-of-the-art single-solution (i.e. deterministic) SMPL predictors PARE [14] and HybrIK [19]. HybrIK gives highly accurate solutions on less-ambiguous images, but struggles with occlusion and truncation. Point estimates from HuManiFlow and PARE perform similarly, but predicting a distribution over pose and shape allows HuManiFlow to additionally estimate prediction uncertainty, which is visualised as directional per-vertex variance. The bottom two rows show some failure cases of our method, when faced with very challenging poses or extreme truncation. The estimated uncertainty is very high for these inputs, which may be used as a signal to discount the predictions as inaccurate.

C. Experimental Results

Results on 70% cropped images. The main manuscript reported results on the 3DPW [27] and SSP-3D [23] datasets, and 50% cropped versions that are more ambiguous. In Tables 2 and 3 of this supplementary material, we report additional results on 70% cropped versions. Comparing this with the results in the main manuscript illustrates the behaviour of pose and shape prediction methods as ambiguity increases due to greater cropping/truncation. In Table 2, we reproduce some of the metrics previously presented in the main manuscript (on 3DPW and 3DPW 50% Cropped), for convenience. Figure 6 presents a qualitative comparison between our method and other probabilistic pose and shape predictors using original, 70% and 50% cropped images.

Qualitative comparison with deterministic methods. Figure 4 compares point estimates from our method with the state-of-the-art single-solution (i.e. deterministic) monocular

Method			*Accuracy*	*Consistency*	*Diversity*
			MPJPE-PA	**2DKP Error**	**3DKP**
Distribution	**Conditioning**	**3DKP Losses**	**Point / Sample Min.**	**Point / Samples**	**Vis. / Invis.**
Matrix-Fisher	MF Parameters	Yes	53.6 / 43.8	**5.0** / 7.2	47.6 / 101.4
Matrix-Fisher	MF Parameters	No	55.8 / 46.4	5.3 / 8.4	49.1 / 128.7
Matrix-Fisher	Rotations	No	54.0 / 43.4	5.1 / 6.8	**51.4** / **131.7**
SO(3) Flow	Rotations	No	**53.4** / **39.9**	5.1 / **6.2**	42.8 / 116.0

Figure 5. **Conditioning on rotation samples vs Matrix-Fisher parameters**, evaluated on 3DPW. Row 1 is HierProbHumans [24]. Row 4 is HuManiFlow. Row 2 is HierProbHumans trained with the same losses as HuManiFlow - i.e. no point-based 3DKP losses. This increases diversity, but accuracy and consistency suffer. Row 3 improves these and maintains diversity, by changing HierProbHumans to *condition on rotations*. This suggests that rotation-conditioning without point-based losses performs best. All models have the same backbone and no. of parameters (approx.), and are trained on the same data.

SMPL prediction approaches [14, 19]. Figure 4 also illustrates some failure cases of our method (bottom two rows) due to challenging poses and extreme truncation. We note that our approach also tends to over-estimate body shape when the subject is wearing baggy clothes, which is due to the low-fidelity synthetic training data pipeline we adopt from [24].

Our ablation models vs. competing methods. Table 1 in the main manuscript presents our ablation study comparing several different SMPL pose distribution modelling approaches. Some of these ablation ablation models are, in fact, very similar to previously proposed probabilistic SMPL prediction methods. Specifically, the Gaussian distribution over full-body concatenated axis-angles (row 1 of Table 1) is similar to [25]. The normalising flow distribution over full-body concatenated axis-angles (row 3 of Table 1) is similar to ProHMR [17]. However, we use linear rational spline coupling layers [6], which are more expressive than the additive coupling layers [4] used by [17]. Moreover, we do not use a 6D rotation representation [29] for distribution prediction, to avoid the need for an orthogonality-enforcing loss, and take into account the non-Euclidean structure of $SO(3)$. Finally, the Matrix-Fisher distribution over body-part rotations (row 7 of Table 1) is similar to HierProbHumans [24]. However, as noted in the main manuscript, [24] conditions body-part rotations on ancestor *distribution parameters*, while we condition directly on ancestor *rotations*. Our approach is more akin to a usual autoregressive model, and allows our method to be more input-consistent. For fairness, we re-train HierProbHumans with the same losses as HuManiFlow, and report results in Figure 5.

Figure 6. Qualitative comparison between our method (HuManiFlow), ProHMR [17] and 3D Multibodies [2] on original, 50% cropped and 70% cropped images (cropping details given in Section B.2). HuManiFlow yields more *diverse* pose and shape samples than ProHMR, and more *input-consistent* samples than 3D Multibodies. The directional variance visualisation shows that HuManiFlow captures prediction uncertainty due to depth ambiguity (z-axis), occlusions and truncations (all-axes) in a more interpretable manner than [17] and [2].

References

[1] 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. 4

[2] 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. 3, 4, 7

[3] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus), 2016. 1, 2

[4] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. In *ICLR Workshop Track*, 2015. 2, 6

[5] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2

[6] Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. Invertible generative modeling using linear rational splines. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4236–4246, 2020. 2, 6

[7] L. Falorsi, P. de Haan, T.R. Davidson, and P. Forré. Reparameterizing distributions on lie groups. *22nd International Conference on Artificial Intelligence and Statistics (AISTATS-19)*, 2019. 2, 3

[8] 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. 1

[9] 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. 1

[10] 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. 4

[11] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of Computer Vision and Pattern Recognition (CVPR) 2011*, 2011. 4

[12] 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. 3, 4

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014. 1

[14] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 11127–11137, Oct. 2021. 3, 5, 6

[15] 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. 3, 4

[16] 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. 3, 4

[17] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 3, 4, 6, 7

[18] 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. 4

[19] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 5, 6

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 4

[21] 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, 3, 4

[22] 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. 4

[23] 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. 4, 5

[24] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical Kinematic Probability Distributions for 3D Human Shape and Pose Estimation from Images in the Wild. In *Proceedings of the International Conference on Computer Vision*, October 2021. 1, 3, 4, 6

[25] 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. 3, 4, 6

[26] 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. 4

[27] 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. 3, 4, 5

[28] 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. 4

[29] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 6