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

This document provides additional material supplementing the main manuscript. Section 1 contains details regarding training data generation, evaluation protocols and probabilistic shape combination. Section 2 discusses qualitative results on the SSP-3D [3] and 3DPW [4] datasets, as well as providing examples from our private evaluation dataset of tape-measured humans.

1. Implementation Details

Training. Table 1 lists the data augmentation methods used to bridge the synthetic-to-real domain gap during synthetic training data generation, along with associated hyperparameter values. Table 2 lists additional hyperparameter values not given in the main manuscript.

Uncertainty Visualisation. Figures 2 and 3 in this supplementary material, as well as several figures in the main manuscript, visualise per-vertex prediction uncertainties. These are computed from the predicted SMPL [2] pose and shape parameter distributions by i) sampling 100 SMPL parameter vectors from the predicted distributions, ii) passing each of these samples through the SMPL function 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 mean for each vertex over all the samples, which is ultimately visualised in the scatter plots as a measure of uncertainty.

SSP-3D Evaluation Groups. SSP-3D [3] contains 311 images of 62 subjects, where subjects can have a different number of associated images. To evaluate our multi-input shape prediction method, the images for each subject were split into groups of *maximum* size equal to N, where N ranged from 1 to 5. For example, if a subject has 6 associated images and $N = 4$, the images would be split into two groups with 4 and 2 images respectively. Splitting/group assignment was done after random shuffling of the images to prevent sequential images with similar poses/global orientations from always being in the same group.

Tape measurement normalisation by height. There is an inherent ambiguity between 3D subject size/scale and distance from camera. Since the true camera location relative to the 3D subject (and the focal length) is unknown, it is not possible to estimate the absolute size of the subject given an image. This is accounted for by the PVE-T-SC [3] metric used to evaluate shape prediction accuracy on synthetic data and SSP-3D in the main manuscript. For our evaluation dataset of tape-measured humans (see Figure 1), scale correction is done using the subject's height. The height of the predicted SMPL human can be determined by computing the neutral-pose mesh (i.e. pose parameters/joint rotations set to 0) and measuring the y-axis distance between the top of the head and bottom of the feet. The ratio between the subject's true height and this predicted height is then used to scale all the predicted body measurements derived from the neutral-pose mesh.

Augmentation	**Hyperparameter**	**Value**
Body part occlusion	Occlusion prob.	0.1
2D joints L/R swap	Swap prob.	0.1
Half-image occlusion	Occlusion prob.	0.05
2D joints removal	Removal prob.	0.1
2D joints noise	Noise range	[-8, 8] pixels
2D vertices noise	Noise range	[-10, 10] mm
Occlusion box	Probability, Size	0.5, 48 pixels

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

Hyperparameter	**Value**
Shape parameter sampling mean	0
Shape parameter sampling var.	2.25
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
Proxy representation dimensions	256 × 256 pixels
2D joint confidence threshold	0.025

Table 2: List of hyperparameter values not provided in the main manuscript.

Probabilistic shape combination. The main manuscript presents our method to probabilistically combine individual body shape distributions, $p(\boldsymbol{\beta}|\mathbf{X}_n)$ for $n = 1, ..., N$, into a final distribution $p(\boldsymbol{\beta}|\{\mathbf{X}_n\}_{n=1}^N)$. The full derivation is given below:

$$\begin{aligned} p(\boldsymbol{\beta}|\{\mathbf{X}_n\}_{n=1}^N) &\propto p(\{\mathbf{X}_n\}_{n=1}^N|\boldsymbol{\beta})p(\boldsymbol{\beta}) \\ &= \left(\prod_{n=1}^N p(\mathbf{X}_n|\boldsymbol{\beta})\right)p(\boldsymbol{\beta}) \\ &\propto \frac{\prod_{n=1}^N p(\boldsymbol{\beta}|\mathbf{X}_n)}{p(\boldsymbol{\beta})^{N-1}} \\ &\propto \prod_{n=1}^N p(\boldsymbol{\beta}|\mathbf{X}_n). \end{aligned} \tag{1}$$

The first and third lines use Bayes' theorem. The second line follows from the conditional independence assumption

$(\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j)|\boldsymbol{\beta}$ for $i, j \in \{1, ..., N\}$ and $i \neq j$. This assumption is reasonable because only the subject's body shape is fixed across inputs - hence, the inputs are independent given the body shape parameters. The final line follows from assuming an (improper) uniform prior over the shape parameters $p(\boldsymbol{\beta}) = 1$.

2. Experimental Results

Evaluation using ground-truth vs predicted inputs. The synthetic training data augmentations listed in Table 1 and the main manuscript are used to increase the robustness of our distribution prediction neural network to noisy and occluded test data, as demonstrated in Figure 3. However, the synthetic-to-real domain gap still persists, as evidenced by Table 3, which compares body shape and pose prediction metrics when using ground-truth, synthetic ground-truth and predicted input proxy representations. A significant improvement in both body shape and pose metrics is observed when using synthetic inputs, instead of predicted inputs. This is mostly because predicted input silhouettes and 2D joints can be very inaccurate in cases with challenging poses, significant occlusion or occluding humans, such that the synthetic training data augmentations are not sufficient. Moreover, synthetic SMPL human silhouettes are not clothed, while silhouette predictors generally classify clothing pixels as part of the human body. This is particularly detrimental to body shape prediction metrics when subjects are dressed in loose clothing, as can be seen in Figure 3 (left side, rows 3 and 4), where our method tends to over-estimate the subject's body proportions.

SSP-3D qualitative results. Figure 2 shows qualitative results, particularly focusing on shape prediction, on groups of input images from SSP-3D [3] corresponding to subjects with a wide range of body shapes. The first column in each cell shows the input images in the group. The second column shows the predicted SMPL [2] body (rendered) for each *individual* image, obtained by passing the mean of predicted SMPL parameter distributions through the SMPL function. The third and fourth columns visualise the 3D per-vertex uncertainty (or variance) in the individual SMPL shape distribution predictions (in a neutral pose i.e. pose parameters/joint rotations set to 0). The fifth column shows the *combined* body shape prediction, which are obtained by probabilistically combining the individual shape distributions.

In particular, note the relationship between challenging poses with significant self-occlusion (e.g. right side, row 4 of Figure 2) and uncertainty in the predicted SMPL shape distribution.

3DPW qualitative results. Figure 3 shows qualitative results, particularly focusing on pose prediction, using single-image inputs from 3DPW [4]. The first column on each side shows the input images. The second column shows the corresponding silhouette and joint heatmap proxy representation predictions. The third column shows the predicted SMPL [2] body (rendered) for each image, obtained by passing the mean of predicted SMPL parameter distributions through the SMPL function. The fourth column visualises the 3D per-vertex uncertainty (or variance) in the SMPL pose and shape distribution predictions (per-vertex uncertainties are mostly due to pose variance rather than shape).

Figure 1: Example images from our private dataset of humans with body measurements obtained using a tape measure or 3D body scanners. The subjects' body pose, clothing, surrounding environment and camera viewpoints vary between images.

Specifically, note the large uncertainties of vertices belonging to body parts that are invisible in the image (and corresponding proxy presentations), either due to occluding objects, self-occlusion or being out-of-frame. Furthermore, large uncertainties also occur when the proxy representation prediction is highly-degraded, such as left side, row 7 of Figure 3.

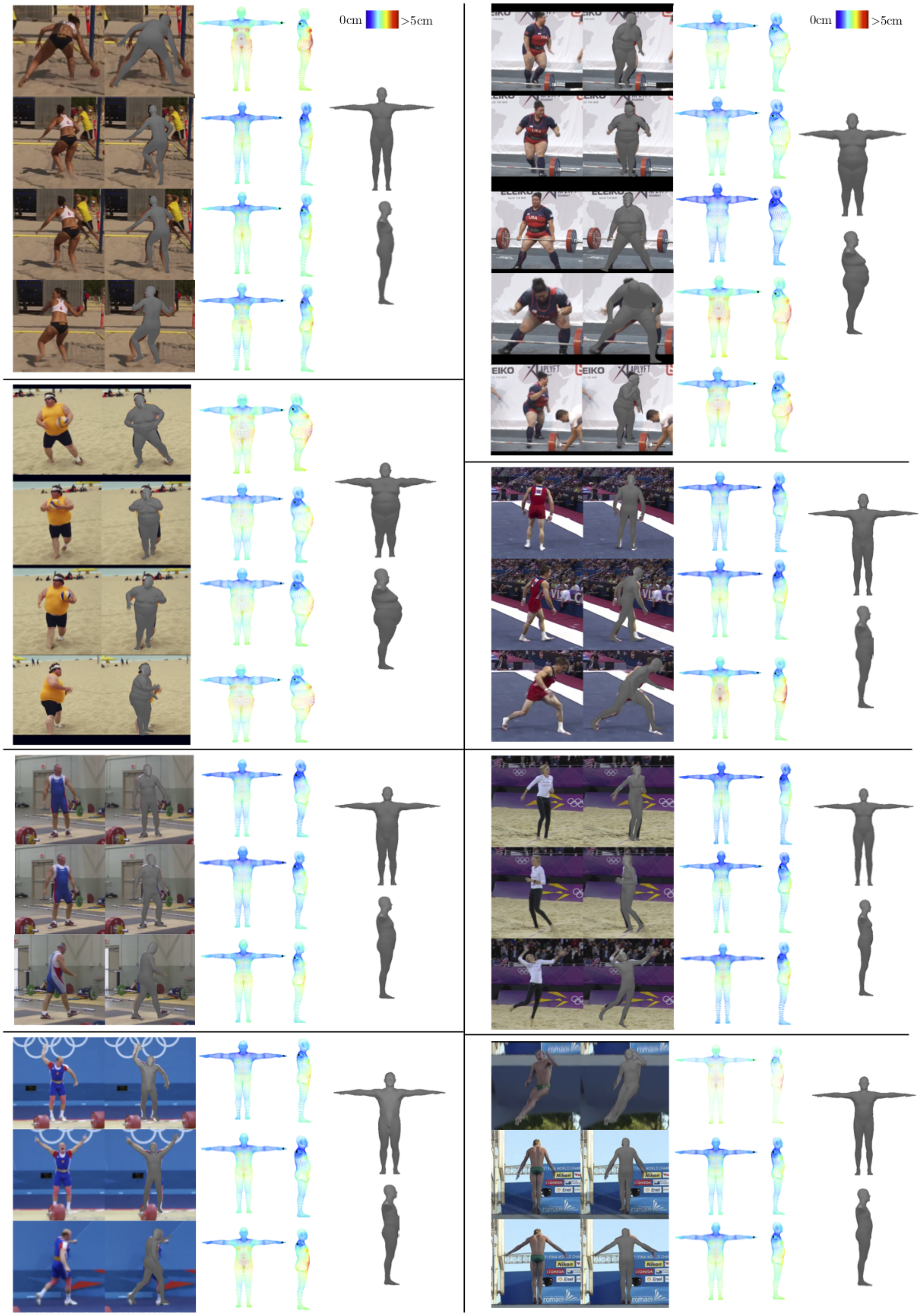

Figure 2: Qualitative results on groups of input images from SSP-3D [3]. Black dots indicate left hands. Within each cell: 1st column is group of input images, 2nd column is predicted SMPL body, 3rd and 4th columns show 3D per-vertex uncertainty in the SMPL *shape* distribution prediction, 5th column is the probabilistically-combined body shape. Challenging poses lead to large shape prediction uncertainty.

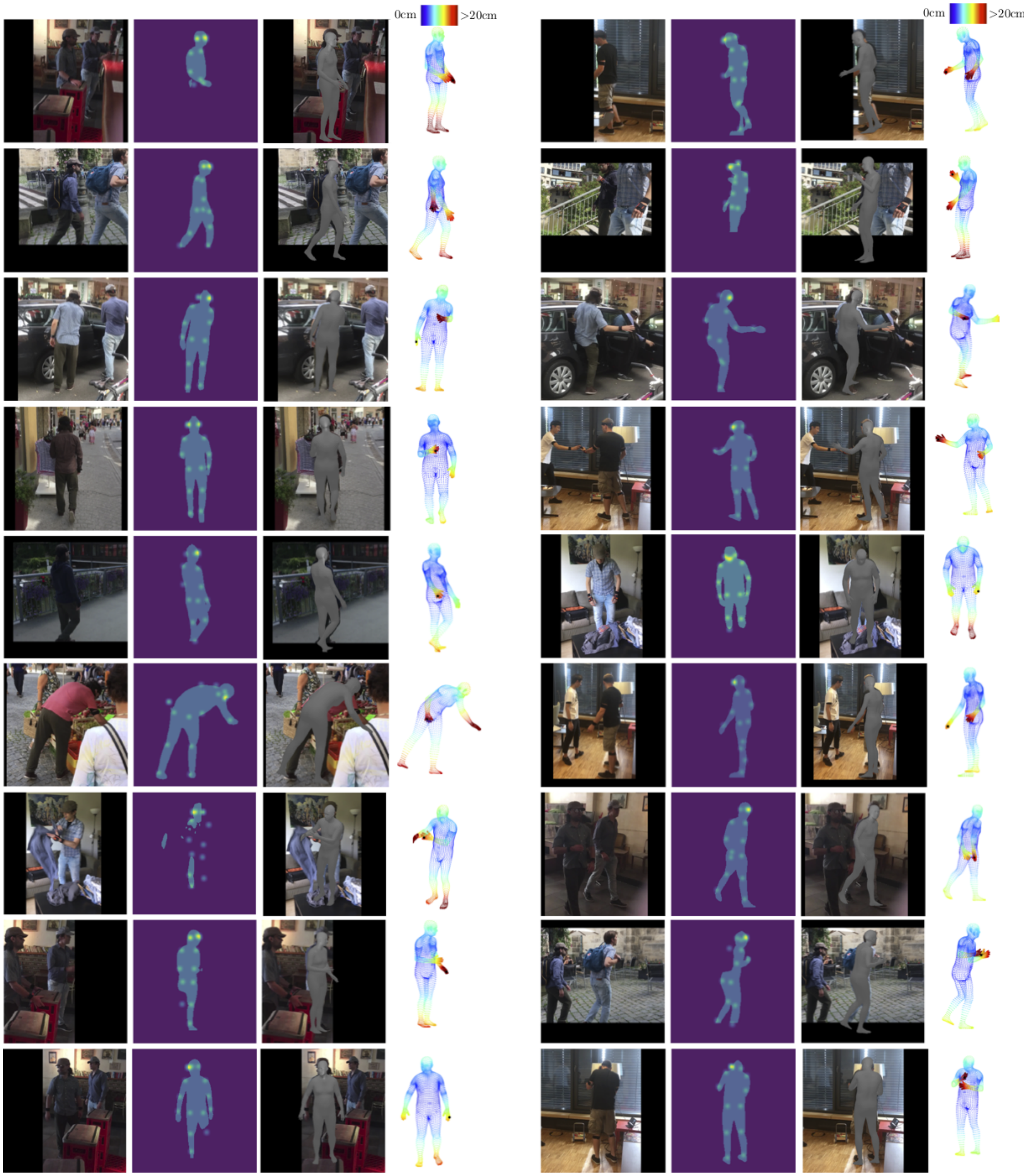

Figure 3: Qualitative results using single-image inputs from 3DPW [4]. Black dots indicate left hands. On each side: 1st column is input image, 2nd column is predicted proxy representation, 3rd column is predicted SMPL body and 4th column is 3D per-vertex uncertainty in the SMPL pose and shape distribution prediction. Vertices of occluded and out-of-frame body parts have higher prediction uncertainties.

Input	3DPW		SSP-3D	
	MPJPE-SC	MPJPE-PA	PVE-PA	PVE-T-SC
GT Synthetic Silh. + 2D Joint Heatmaps	**64.3**	**45.7**	**52.9**	**10.1**
GT Silh. + 2D Joint Heatmaps	-	-	69.9	14.4
Predicted Silh. + 2D Joint Heatmaps	90.9	61.0	71.4	15.2

Table 3: Comparison between ground-truth (GT), synthetic ground-truth and predicted input silhouettes and 2D joints, in terms of MPJPE-SC and MPJPE-PA (both in mm) on 3DPW [4], as well as PVE-PA and PVE-T-SC (both in mm) on SSP-3D [3]. Predicted silhouettes are obtained using DensePose [1] and predicted 2D joint coordinates and confidences (for thresholding) are obtained using Keypoint-RCNN from Detectron2 [5]. Synthetic ground-truth inputs are obtained by rendering the SMPL [2] body mesh labels given by SSP-3D and 3DPW, using ground-truth camera parameters, into silhouette and 2D joint input representations.

References

[1] 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, 5

[2] 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, 5

[3] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. 1, 2, 3, 5

[4] 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. 1, 2, 4, 5

[5] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5