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

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

Department of Engineering
University of Cambridge
Cambridge, UK

This document provides additional material supplementing the main manuscript. Section 1 gives definitions of width, depth, circumference and length measurements over the SMPL [11] body surface. Section 2 qualitatively corroborates the local controllability experiments presented in the main manuscript. Section 3 provides qualitative results comparing our measurement distribution prediction network against previously-proposed [16] SMPL shape coefficient (β) distribution predictors, using input images from our (i) synthetic evaluation dataset, (ii) SSP-3D [14], and (iii) two private datasets of tape-measured humans, which were named "A-Pose Subjects" and "Varying-Pose Subjects" in the main manuscript. Section 4 contains details regarding synthetic data generation and examples of synthetic training images and synthetic evaluation images used in the ablation studies presented in the main manuscript.

Finally, the attached video file `1340_varying_local_measurements.mp4` visualises the effect of smoothly increasing specific input measurement offsets from -5cm to 5cm, thereby demonstrating the semantic local controllability of our measurements-to-βs regressor.

1 Body Measurement Definitions

In the main manuscript, obtaining measurements from an SMPL T-pose body was abstracted away as an operation $\boldsymbol{m} = \text{measure}(\boldsymbol{\beta})$. In this section, Figures 1 and 2 give definitions of each of the 23 body measurements, in terms of the SMPL T-pose joint/vertex IDs used as endpoints (for widths, depths and lengths) or waypoints (for circumferences). Given the joint/vertex IDs, measurement values are obtained by simply computing the 3D Euclidean distance between the corresponding endpoints/waypoints for a T-pose body. For circumferences, the Euclidean distance between waypoints are summed along the circumference. Note that a T-pose SMPL body only depends on given shape coefficients $\boldsymbol{\beta}$ (see Equation 1 in the main manuscript). Thus, the operation $\boldsymbol{m} = \text{measure}(\boldsymbol{\beta})$ involves (i) generating T-pose joints and vertices from the input $\boldsymbol{\beta}$, (ii) gathering measurement endpoints/waypoints using the joint and vertex IDs given in Figure 2 and (iii) computing Euclidean distances between endpoints/waypoints and summing if needed.

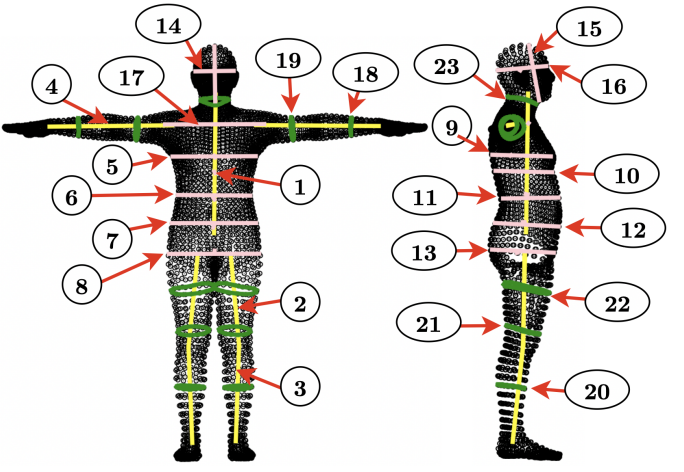

Figure 1: Front- and side-view visualisation of 23 measurement definitions over the SMPL [11] body surface. Please refer to Figure 2 for the semantic meaning of each measurement and the specific SMPL vertex/joint IDs used to define them. Colour key: Yellow = "Length" measurements defined using SMPL T-pose joints, Pink = "Width" / "Depth" measurements defined using SMPL T-pose vertices, Green = "Circumference" measurements defined using SMPL T-pose vertices.

Figures 1 and 2 visualise and define separate left/right limb measurements. However, locally controlling left/right measurements independently is challenging with the SMPL body model. The SMPL shape space is learnt using PCA applied to human body scans from CAESAR [13] and the majority of human bodies exhibit strong left/right symmetry, both in CAESAR and in the general population. Thus, we convert the separate left/right limb measurements defined in Figures 1 and 2 into single limb measurements by taking the mean of the left and right sides.

2 Local Controllability

The main manuscript analyses the local controllability of the proposed measurements-to-βs regressor. In particular, we quantitatively show that regressing from body measurements to 10 SMPL shape coefficients (βs) results in poor controllability, wherein an input offset of +5cm applied to a specific measurement results in large undesired output offsets to several other measurements. This is because the 10-dimensional SMPL shape space is not expressive enough to allow for fine-grained local control of body shape. A significant quantitative improvement in local controllability is observed when the number of SMPL βs is increased from 10 to 70. Figure 3 demonstrates this qualitatively, by visualising the effect of input measurement offsets on SMPL bodies when regressing 10 βs and 70 βs. Using only 10 βs may result in either (i) non-local output offsets and unrealistic output body shapes (Figure 3, top row, highlighted by red arrows) or (ii) zero output offsets and unchanged output body shapes (Figure 3, top row, column 5). On the other hand, the measurement-to-70-βs regressor yields *realistic and local* body shapes offsets, which match the desired inputs.

Meas. ID	Meas. Name	IDs of SMPL Joint Endpoints / Vertex Endpoints	Meas. ID	Meas. Name	IDs of SMPL Joint Endpoints / Vertex Endpoints
1	Torso Length	(0, 15)	13	Hip Depth	(3141, 3145)
2	Thigh Length	Left: (1, 4), Right: (2, 5)	14	Head Width	(368, 3872)
3	Calf Length	Left: (4, 7), Right: (5, 8)	15	Head Height	(412, 3058)
4	Arm Length	Left: (16, 20), Right: (17, 21)	16	Head Depth	(457, 3165)
5	Chest Width	(738, 4226)	17	Shoulder Width	(1509, 4982)
6	Stomach Width	(1323, 4804)	18	Forearm Circum.	Left: (1557, 1558, 1587, 1554, 1553, 1727, 1583, 1584, 1689, 1687, 1686, 1590, 1591, 1548, 1547, 1551) Right: (5027, 5028, 5020, 5018, 5017, 5060, 5061, 5157, 5156, 5159, 5053, 5054, 5196, 5024, 5023, 5057)
7	Abdomen Width	(1794, 5256)	19	Upper Arm Circum.	Left: (628, 627, 789, 1311, 1315, 1379, 1378, 1394, 1393, 1389, 1388, 1233, 1232, 1385, 1381, 1382, 1397, 1396) Right: (4117, 4277, 4791, 4794, 4850, 4851, 4865, 4866, 4862, 4863, 4716, 4717, 4859, 4856, 4855, 4870, 4871, 4116)
8	Hip Width	(3129, 6550)	20	Calf Circum.	Left: (1074, 1077, 1470, 1094, 1095, 1473, 1465, 1466, 1108, 1111, 1530, 1089, 1086, 1154, 1372) Right: (4583, 4580, 4943, 4561, 4560, 4845, 4640, 4572, 4573, 5000, 4595, 4594, 4940, 4938, 4946)
9	Chest Depth	(3015, 3076)	21	Lower Thigh Circum.	Left: (1041, 1147, 1171, 1172, 1029, 1030, 1167, 1033, 1034, 1035, 1037, 1036, 1038, 1040, 1039, 1520, 1042) Right: (4528, 4632, 4657, 4660, 4515, 4518, 4653, 4519, 4520, 4521, 4522, 4523, 4524, 4525, 4526, 4991, 4527)
10	Underbust Depth	(1329, 3017)	22	Upper Thigh Circum.	Left: (910, 1365, 907, 906, 957, 904, 905, 903, 901, 962, 898, 899, 934, 935, 1453, 964, 909, 910) Right: (4397, 4396, 4838, 4393, 4392, 4443, 4391, 4390, 4388, 4387, 4448, 4386, 4385, 4422, 4421, 4926, 4449)
11	Stomach Depth	(3502, 3509)	23	Neck Circum	(3050, 3839, 3796, 3797, 3662, 3663, 3810, 3718, 3719, 3723, 3724, 3768, 3918, 460, 423, 257, 212, 213, 209, 206, 298, 153, 150, 285, 284, 334)
12	Abdomen Depth	(3507, 3159)			

Figure 2: Semantic meaning of each measurement visualised in Figure 1, along with SMPL joint/vertex IDs used to define them. Joint/vertex IDs correspond to endpoints for "Width", "Depth" and "Length" measurements, and waypoints for "Circum." measurements. These specific 23 measurements were chosen to sufficiently constrain the body surface, such that a full T-pose body mesh can be recovered from just 23 measurements.

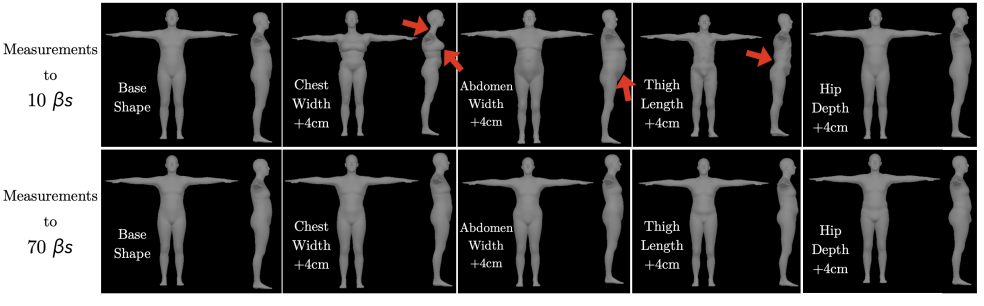

Figure 3: Visualisation of the effect of +4cm input measurement offsets applied to a base body shape. Measurement offsets are mapped to SMPL β offsets using the proposed measurements-to-βs linear regressor, using 10 SMPL βs (top) and 70 SMPL βs (bottom). Regressing only 10 βs may result in (i) non-local output offsets and unrealistic output body shapes (top row, highlighted by red arrows) or (ii) zero output offsets and unchanged output body shapes (top row, hip depth in column 5). Regressing 70 βs yields *realistic and local* body shape offsets, which match the desired inputs.

3 Qualitative Results

Figures 4, 5 and 6 compare results from our proposed measurement distribution predictor and SMPL β distribution predictors, both using 70 βs (i.e. the number of shape coefficients output by our measurements-to-βs regressor), as well as 10 βs as proposed by Sengupta *et al*. [16]. We conclude that predicting Gaussian distributions over semantic body mea-

surements allows for meaningful predictions of *local* aleatoric [3, 7] shape uncertainty that models ambiguities in the input images related to the subject's pose, camera viewpoint and occlusion, which is not possible with independent Gaussian distributions over *global* SMPL βs. Improved local shape uncertainty quantification yields better body shape estimates after probabilistic combination, as demonstrated in the main manuscript.

4 Synthetic Data Generation

Following [14, 15, 16, 17], we adopt a synthetic training framework as a means of overcoming the lack of body shape diversity in common datasets for 3D pose and shape estimation from images [5, 10, 19]. In particular, we use a edge-and-joint-heatmap proxy representation [15], to bridge the domain gap between low-fidelity synthetic training inputs and real test inputs.

Examples of synthetic RGB images, and corresponding edge-image + 2D joint heatmap proxy representations, are given in Figure 7. They are generated on-the-fly during training by sampling a random SMPL shape, SMPL pose, clothing texture and background image for each training iteration, and rendering using a light-weight renderer [12].

SMPL poses (i.e. 3D joint rotations) and global body rotations are randomly selected from the training splits of UP-3D [10], 3DPW [19] and Human3.6M [5]. SMPL shapes are obtained in two stages: (i) base body shape coefficients are randomly sampled from $\mathcal{N}(\beta_i; 0, 1.25^2)$ for $i \in \{1, 2, ..., |\boldsymbol{\beta}|\}$, and (ii) measurement offsets from the base body (for each of the 23 measurements listed in Figure 2) are randomly sampled from $\mathcal{N}(m_j; 0, 0.02^2)$ (units of metres), converted into shape coefficient offsets using the measurements-to-βs regressor and added to the random base body shape. Step (ii) acts as random body *measurement* augmentation, and is crucial when learning to estimate measurement uncertainties.

Clothing textures for the SMPL body are randomly selected from SURREAL [18] and MultiGarmentNet [1]. Background images are obtained from LSUN [21], which contains both indoor and outdoor scenes. Note that background images, intentionally, may contain other humans - this is important for the network to be robust against background humans in real in-the-wild test images.

The sampled SMPL shape, SMPL pose, clothing texture and background image are rendered into a synthetic RGB image using Pytorch3D [12]. Perspective camera translation is randomly sampled, along with Phong lighting parameters. 2D joint locations (and Gaussian heatmaps) are generated by projecting 3D SMPL joint locations onto the 2D image plane. Synthetic RGB images are converted to edge-images using Canny edge detection [2].

Finally, following [14, 16, 17], several data augmentation and corruption methods are applied to the synthetic edge-images and joint heatmaps, to further close the domain gap between training data and noisy test data. Hyperparameters associated with random data generation and augmentation are listed in Table 1.

The synthetic training inputs are paired with ground-truth SMPL pose parameters, body measurements, global body rotations and 2D joint locations, which are each obtained at some point in the synthetic input generation process, as detailed above. Body measurements are computed from sampled SMPL shape coefficients using the measure(.) operation defined in Section 1.

The ablation studies presented in the main manuscript use a synthetic evaluation dataset, which is rendered very similarly to synthetic training data. Examples are given in Figure 8.

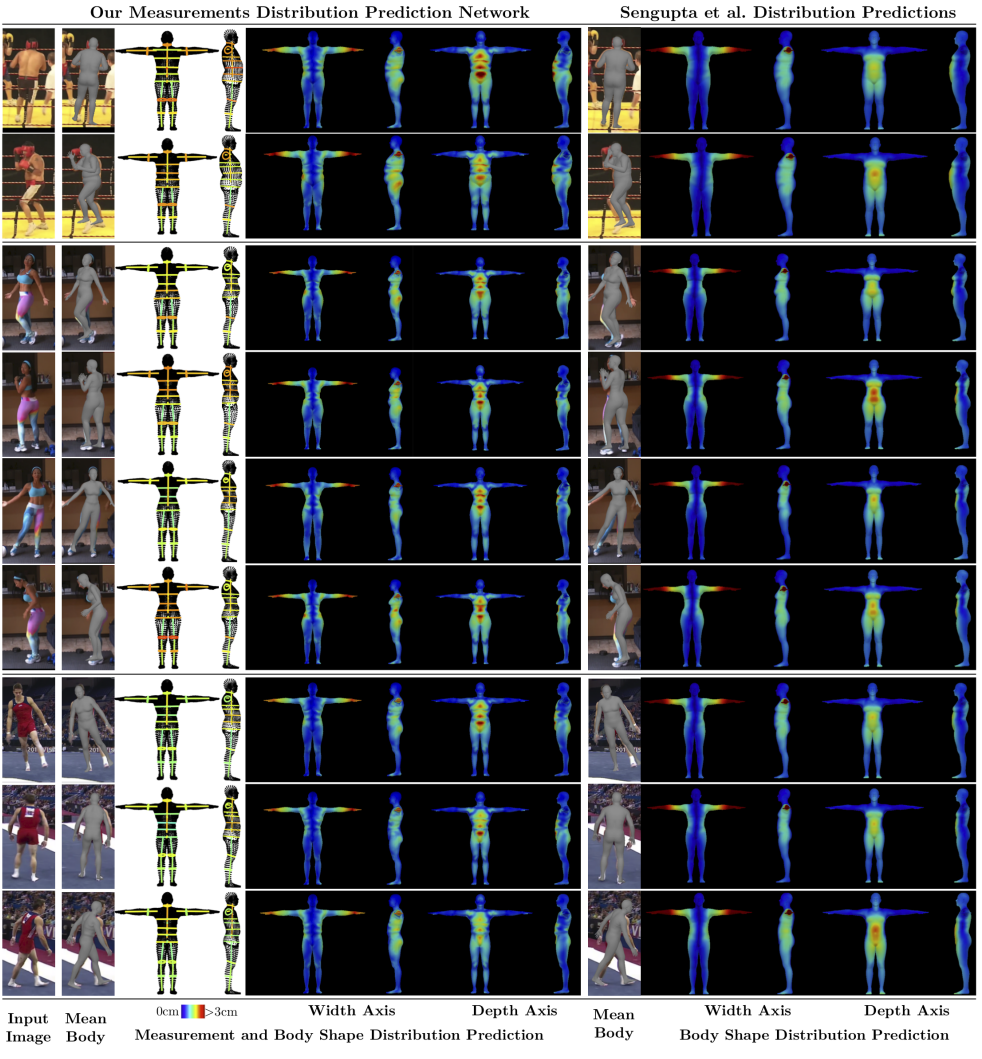

Figure 4: Comparison between our predicted measurement distributions and SMPL β distributions from Sengupta *et al.* [16] on images from SSP-3D [14]. Note that [16] is the previous state-of-the-art approach in terms of body shape metrics on SSP-3D. Similar to Figure 3 in the main manuscript, this figure demonstrates that Gaussian measurement distributions exhibit meaningful *local* shape uncertainty arising from varying camera angles, challenging poses and self-occlusions. In contrast, independent Gaussian SMPL β distributions from [16] cannot model such local shape uncertainty, since βs control global deformations over the whole body surface. Instead, predicted shape uncertainty increases globally over the whole body when the input contains a challenging pose or self-occlusion (compare the bottom 2 rows, columns 10-13). Such global uncertainty is less useful for downstream tasks as it does not specify which body-parts have high prediction uncertainty, only that the network is uncertain as a whole. Thus, probabilistically combining measurement distributions yields better shape metrics than [16], as shown in Tables 2 and 3 in the main manuscript.

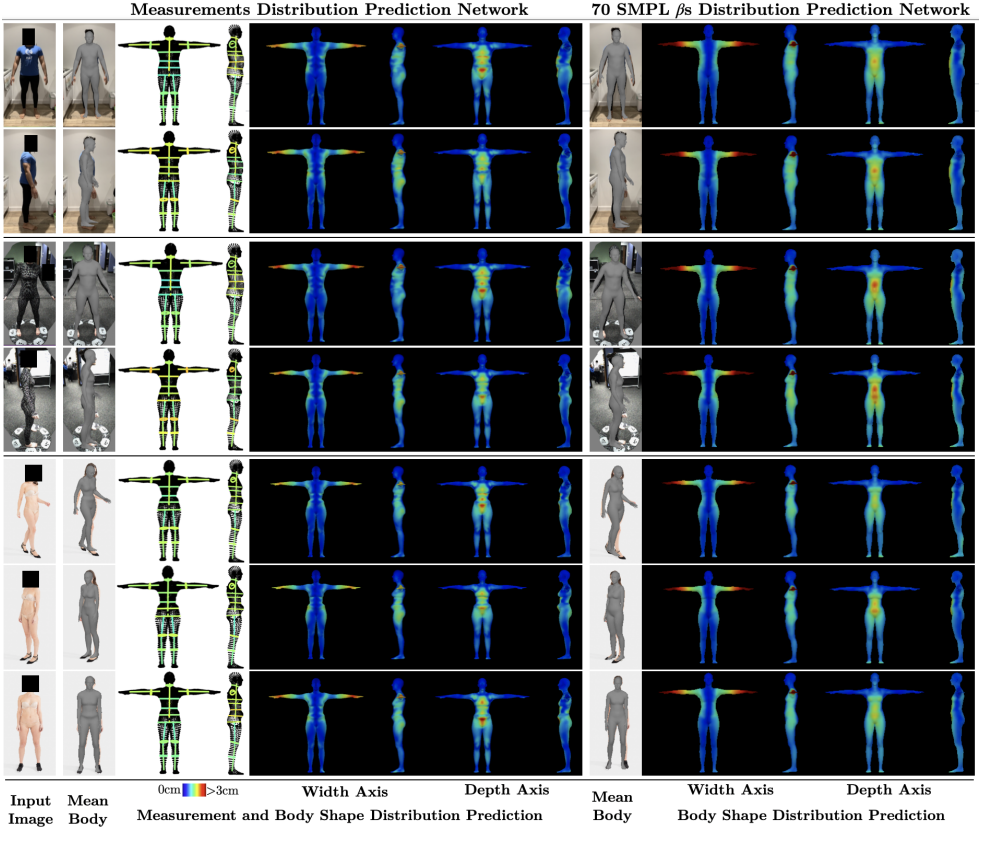

Figure 5: Comparison between measurement distribution predictions and SMPL 70 β distribution predictions on images from our two private datasets of tape-measured humans, "A-Pose Subjects" (rows 1-4) and "Varying-Pose Subjects" (rows 5-7). This figure further corroborates that predicting measurement distributions leads to meaningful local shape uncertainty when given images with different global body orientations, i.e. front-facing images result in lower predicted uncertainties for width measurements (columns 3, 5, 6), while side-facing images result in lower uncertainties for depth measurements (columns 4, 7, 8). On the other hand, uncertainty predictions from the SMPL 70 β distribution network cannot model such local ambiguities in the input image due to variations in camera viewpoint.

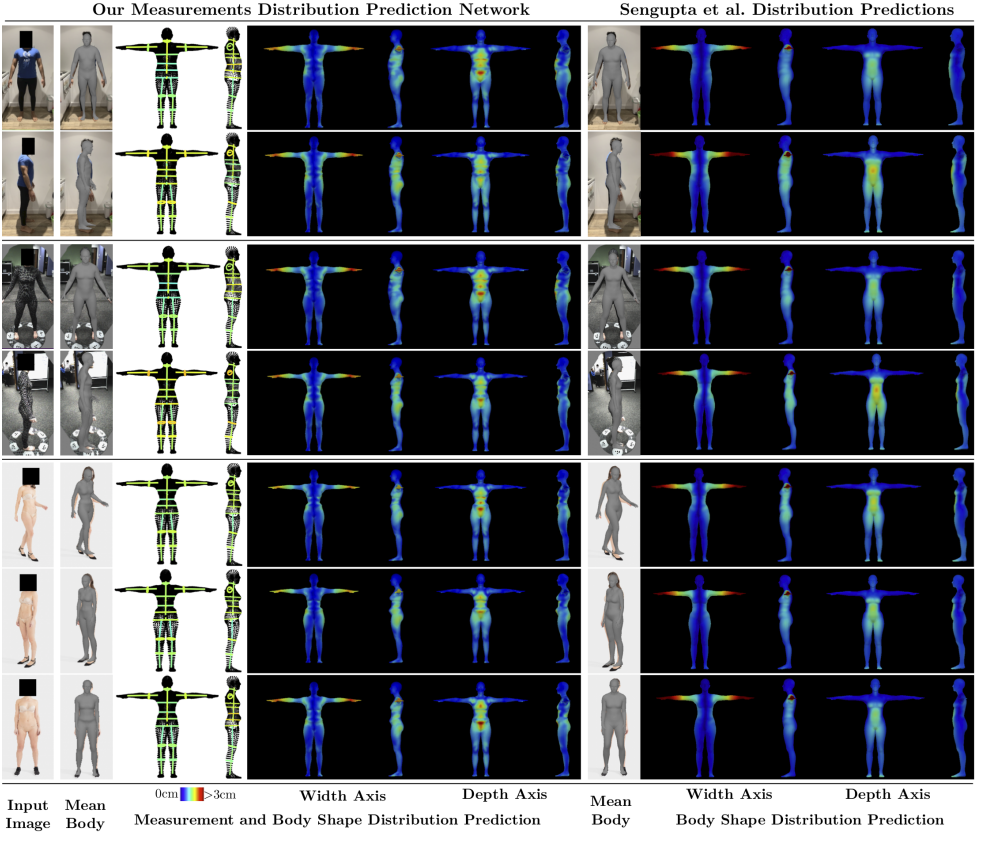

Figure 6: Comparison between our predicted Gaussian measurement distributions and SMPL β distributions from Sengupta *et al.* [16] on images from our two private datasets of tape-measured humans, "A-Pose Subjects" (rows 1-4) and "Varying-Pose Subjects" (rows 5-7). Similar to Figure 5, measurement distribution prediction results in intuitive local shape uncertainty when given images with different global body orientations, while predictions from the SMPL β distribution network of [16] do not appear to be related to the shape information present in the input (as dictated by the subject's global body orientation, pose or occlusions).

Figure 7: Examples of synthetic RGB training images and corresponding edge-image + 2D joint heatmap proxy representations. Images within each group of 3 use the same pose (selected from common 3D SMPL pose datasets [5, 10, 19]) but different random body shapes, clothing textures, backgrounds and lighting parameters. The synthetic RGB images are computationally cheap and far from photorealistic - however, edge-filtering significantly reduces the synthetic-to-real domain gap.

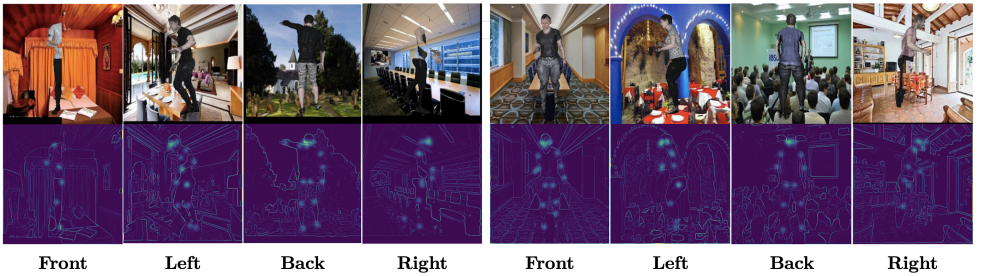

Figure 8: Examples of synthetic RGB evaluation images and corresponding edge-image + 2D joint heatmap proxy representations used for the ablation studies presented in the main manuscript. Each of the 1000 random body shapes in the synthetic evaluation dataset are posed in 4 different configurations, facing forwards, left, backwards and right.

Hyperparameter	Value
Camera translation sampling mean	(0, -0.2, 2.5) metres
Camera translation sampling variance	(0.05, 0.05, 0.25) metres
Camera 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)
Proxy representation dimensions	256×256 pixels
Bounding box scale factor range	(0.8, 1.2)
Body part occlusion probability (divided into 24 DensePose [4] parts)	0.1
2D joints L/R swap probability (for shoulders, elbows, wrists, hips, knees, ankles)	0.1
Vertical/horizontal half occlusion probability	0.05/0.05
2D joint heatmap removal probability	0.1
2D joint heatmap location noise range	[-8, 8] pixels

Table 1: Hyperparameter values associated with random synthetic training data generation and augmentation.

Method	Single-Image Inference Time (ms)
GraphCMR [9]	33
HMR [6]	**30**
SPIN [8]	**30**
DaNet [22]	160
STRAPS [14]	250
Sengupta *et al.* [16]	250
Ours	140

Table 2: Comparison of single-image inference run-times (in milliseconds) for different 3D shape and pose estimation approaches. Methods in the top half do not use proxy representation inputs, while methods in the bottom half do. Proxy representations enable the use of synthetic training data (by closing the synthetic-to-real domain gap), thus overcoming the lack of real training data with accurate and diverse body shape labels. However, proxy representation computation during inference significantly increases run-time. Our approach is faster than recent approaches that use silhouette-based proxy representations [14, 16], since edge-detection is a less-intensive operation than deep-learning-based segmentation. Most (90%) of our inference time is due to 2D joint detection [20].

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)*. IEEE, Oct 2019.

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

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] K. Robinette, S. Blackwell, H. Daanen, M. Boehmer, S. Fleming, T. Brill, D. Hoeferlin, and D. Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) Final Report AFRL-HE- WP-TR-2002-0169. Technical report, US Air Force Research Laboratory, 2002.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.