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

This document provides additional material supplementing the main manuscript. Section 1 contains implementation details, particularly regarding synthetic training data generation and per-vertex uncertainty visualisation. Section 2 discusses qualitative results on the SSP-3D [9] and 3DPW [13] datasets, and compares distribution predictions on images with versus without artificial occlusions.

1. Implementation Details

1.1. 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 [7] 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 1. 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.

The synthetic RGB images shown in Figure 1 are computationally cheap to render but clearly far from photorealistic, resulting in a large synthetic-to-real domain gap. However, simple edge detection [2] is able to significantly reduce this gap [3], motivating the use of edge-filtered images as part of our input proxy representation. Furthermore, we found that noisy edge detections (as seen in Figure 1) 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 [13] and UP-3D [6], as well as Human3.6M [5] 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 $i = 1, ..., 10$ [9]. RGB clothing textures for the SMPL body mesh are selected from SURREAL [12] and MultiGarmentNet [1], resulting in 917 training textures and 108 validation textures. Backgrounds are obtained from LSUN [14], 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 1. List of hyperparameter values associated with synthetic training data generation.

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 2. List of synthetic training data augmentations and their associated hyperparameter values. Body part occlusion uses the 24 DensePose [4] parts. Joint L/R swap is done for shoulders, elbows, wrists, hips, knees, ankles.

Textured SMPL meshes are rendered with Pytorch3D [8], using a perspective camera model and Phong shading. Camera and lighting parameters are randomly sampled, with sampling hyperparameters given in Table 1. 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 between synthetic data and real test data, which may exhibit significant occlusions and noise, 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 2.

Figure 1. 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 [2] 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 2 and 3.

1.2. Visualisation of Per-Vertex Uncertainty

Figures 2, 3 and 4 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 [7] 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.

2. Qualitative Results

Figure 3 presents results on artificially occluded images from SSP-3D [9]. 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 3 also illustrates our method's ability to predict a range of body shapes, owing to the synthetic training framework used.

Figure 2 presents results on the test split of 3DPW [13].

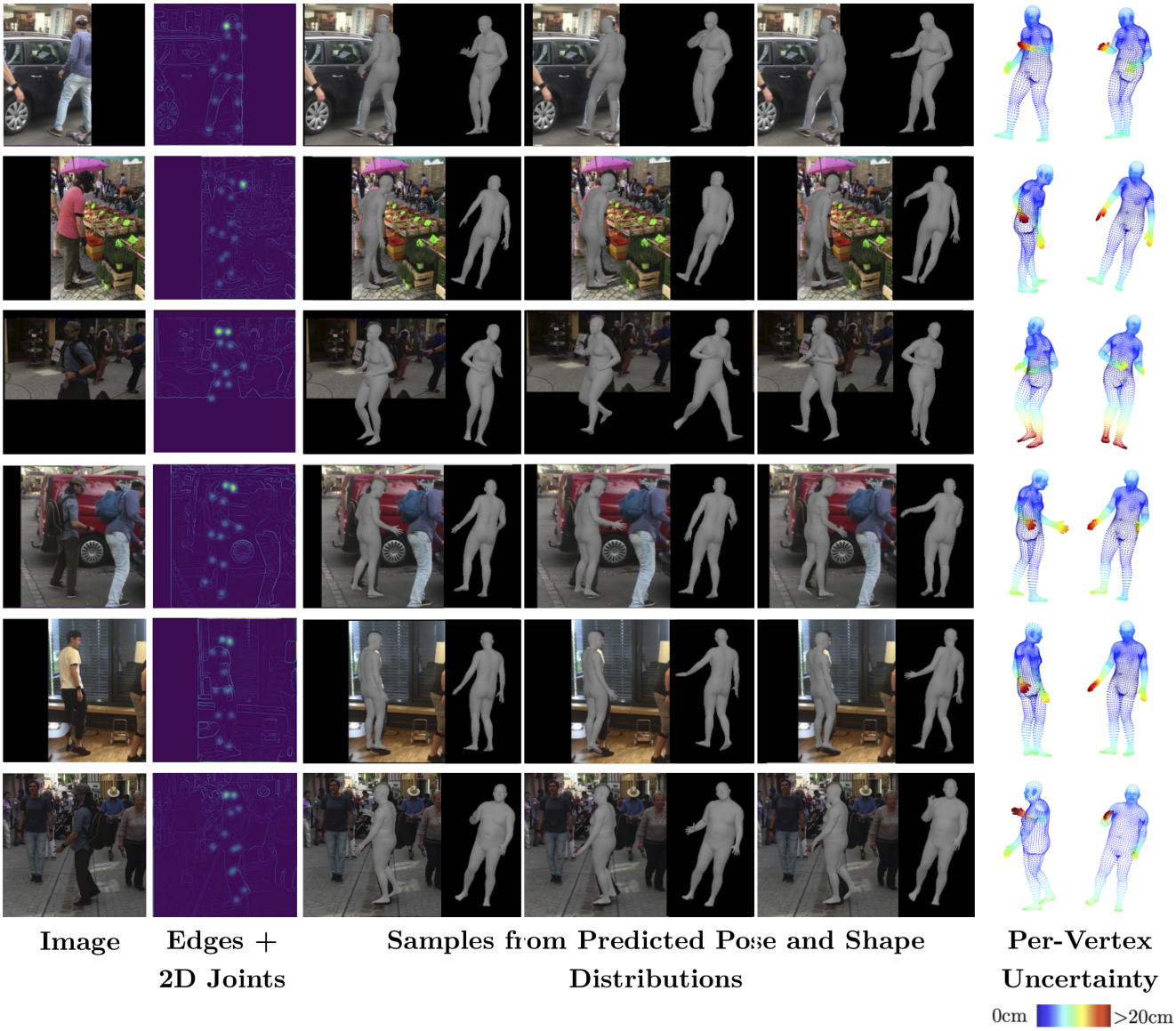

Figure 2. 3D reconstruction samples and per-vertex uncertainties corresponding to shape and relative 3D joint rotation distributions predicted from 3DPW images[13]. The selected images exhibit self-occlusion and out-of-frame body parts, which result in greater 3D location uncertainty for vertices belonging to ambiguous parts.

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 4 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 [10, 11]. This is demonstrated in the non-occluded images in Figure 4 (left), by the left arm samples in rows 1 and 5 and the right arm in row 4.

Figure 3. 3D reconstruction samples and per-vertex uncertainties corresponding to shape and relative 3D joint rotation distributions predicted from SSP-3D images[9]. The images are artificially occluded, resulting in greater 3D location uncertainty for vertices belonging to ambiguous parts.

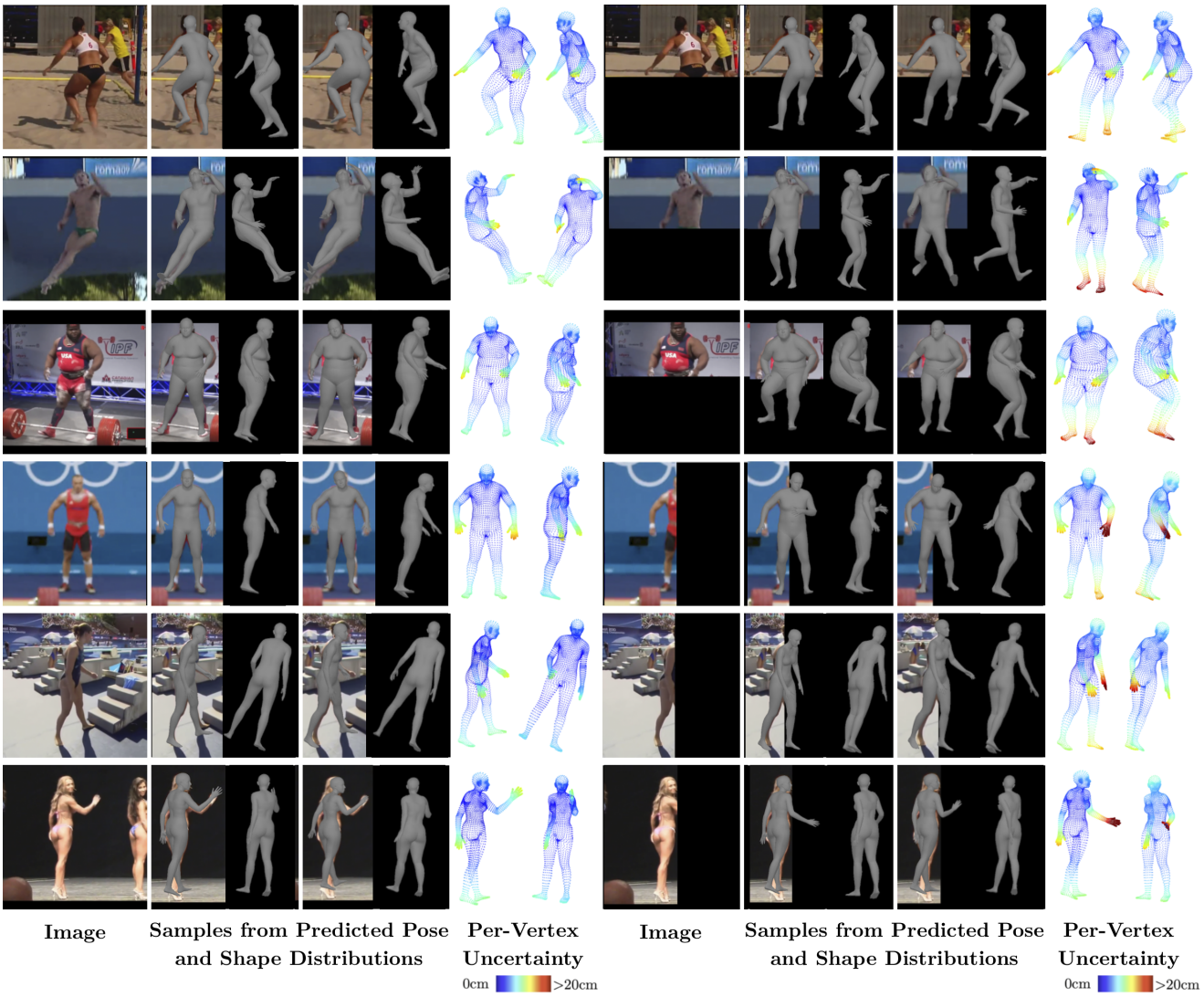

Figure 4. Comparison between 3D samples and per-vertex uncertainties obtained using artificially occluded versus non-occluded input images from SSP-3D [9]. Ambiguous parts have greater prediction uncertainty.

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. 1

[2] John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 8(6):679–698, 1986. 1, 2

[3] 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. 1

[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. 1

[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. 1

[6] 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. 1

[7] 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

[8] 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. 1

[9] 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, 5

[10] 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. 3

[11] 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. 3

[12] 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. 1

[13] 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, 3

[14] 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. 1