CondiMen: Conditional Multi-Person Mesh Recovery Supplementary material

Romain Brégier[1] Fabien Baradel[1] Thomas Lucas[1] Salma Galaaoui[1,2]
Matthieu Armando[1] Philippe Weinzaepfel[1] Grégory Rogez[1]
[1] NAVER LABS Europe
[2]LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France

This document provides additional material regarding *CondiMen: Conditional Multi-Person Mesh Recovery*. In Sec. 1, 2 and 3 we report results of additional experiments aiming at better characterizing properties of CondiMen. Sec. 4 contains experimental results considering different Bayesian network connectivity, complementing results presented in the main paper. Lastly, in Sec. 5 we describe some implementation details used in our experiments.

1. Attributes dependency modeling

In addition to the numerical results reported in the main paper, Fig. 1 provides qualitative results of counterfactual experiments that illustrate the ability of our approach to model dependencies between attributes in the mesh recovery problem. In Fig. 1a, we vary the principal component of body shape parameters (as external inputs) while keeping camera intrinsics constant, and observe the effect on predicted distances to the camera. Similarly, Fig. 1b illustrates the effect of setting different focal lengths as inputs, demonstrating how this variation influences other variables, particularly the distance to the camera.

2. Uncertainty modeling

Empirically, we observe a correlation between the conditional likelihoods of our predictions – *i.e*. the value of conditional probability densities predicted by our Bayesian network – and actual prediction errors, as illustrated in Fig. 2 for various test sets. This suggests that the proposed model is able to capture the uncertainty of its predictions to some extent, which could be useful in downstream applications.

3. Failure cases

Overall, CondiMen produces plausible predictions. However, it also inherits common limitations of existing mesh recovery methods. Notable failure cases (not specific to our method) include unusual poses that deviate significantly from the training data (Fig. 3 top row). Additionally, scenes with mutually occluding persons introduce ambiguity in the detection task (Fig. 3 bottom row).

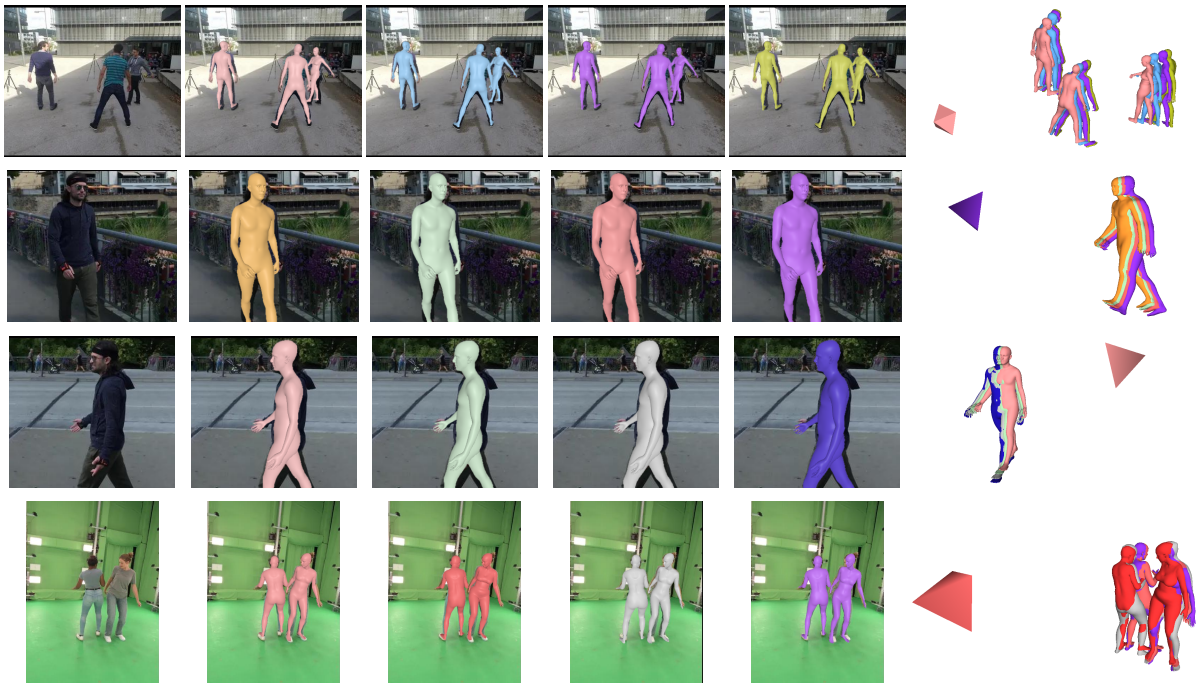
4. Bayesian network connectivity

Beyond *CondiMen* and the *Naive Bayes* baseline presented in the main paper, we also experimented with two additional variants to study the impact of Bayesian network connectivity on numerical performances. Fig. 4 shows the full connectivity of the different Bayesian networks considered in this study, from which the graphical model Fig. 1 of the main paper is extracted. *Variant1* features a denser set of conditional dependency connections compared to *CondiMen*, and in *Variant2* the dependency order between body shape and encoded depth variables is furthermore permuted. We report results of quantitative evaluations in Table 1. The dependency order in *Variant2* prevents from properly exploiting external camera intrinsics and body shape inputs, leading to much larger absolute position errors in this setting than with *CondiMen* and *Variant1*, but still outperforming the *Naive Bayes* baseline (*e.g*. on *Human3.6M* in *Single-View intr-shape* setup, $PE = 676.9mm$ for *Variant2* *vs*. $284.6mm$ for *CondiMen*, $366.2mm$ for *Variant1*, and $898.0mm$ for *Naive Bayes*). Overall, *CondiMen* achieves better numerical performances than *Variant1*. The restricted connectivity of *CondiMen* imposes stronger inductive biases than the connectivity of *Variant1*, and we posit it helps learning meaningful correlations between human attributes. This observation is arguably dependent of our training strategy, and we expect that further increasing the amount and variability of training data would diminish the benefits of imposing such inductive priors.

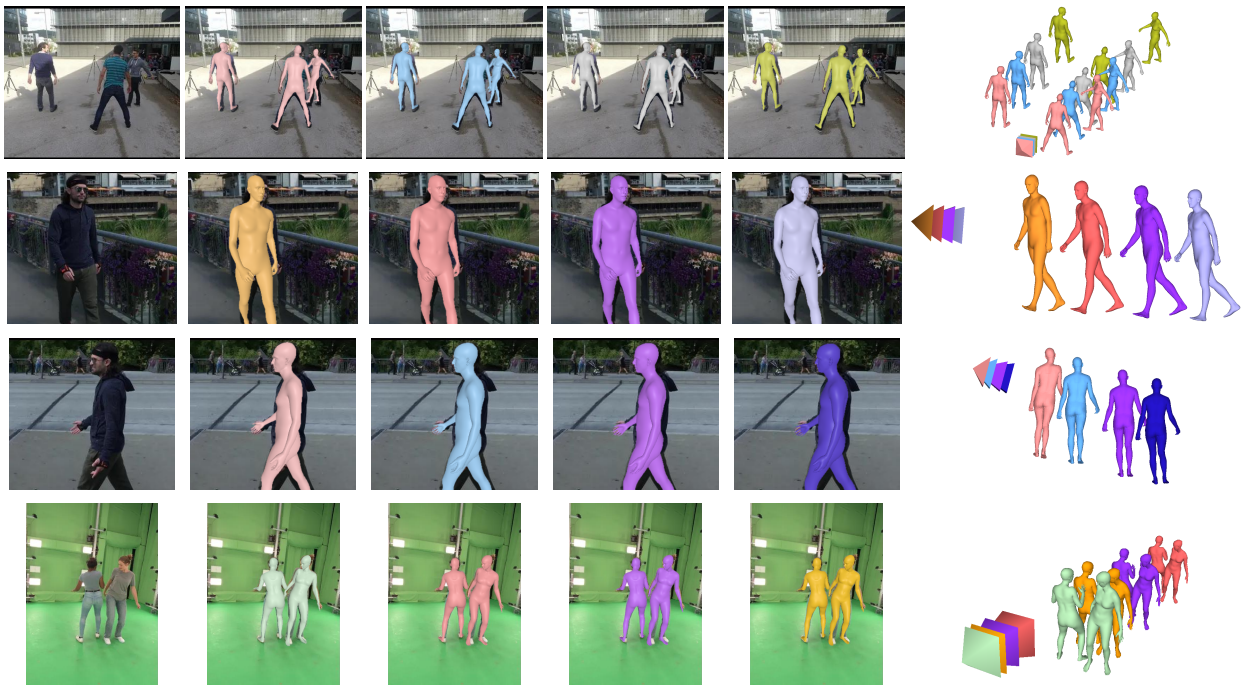
5. Implementation details

5.1. Additional synthetic data

Fig. 5 illustrates the additional synthetic data generated to train our method. The images were rendered using Blender [1]. We created a collection of 3D scenes, each

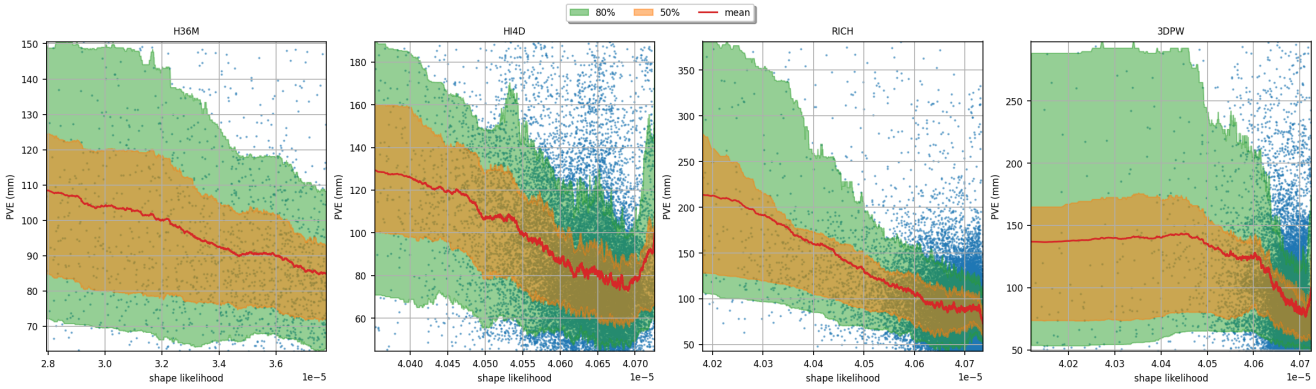


(a) Predictions assuming different body shapes.

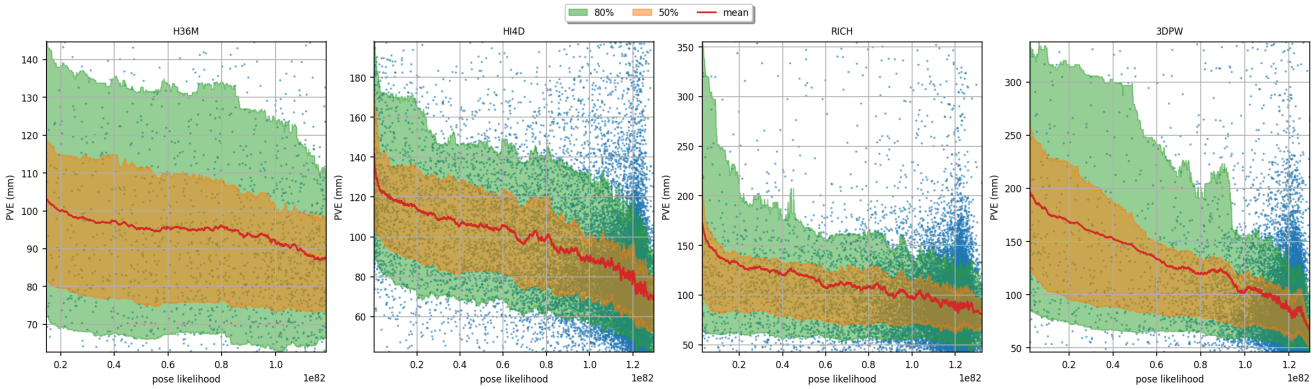


(b) Predictions assuming different focal lengths.

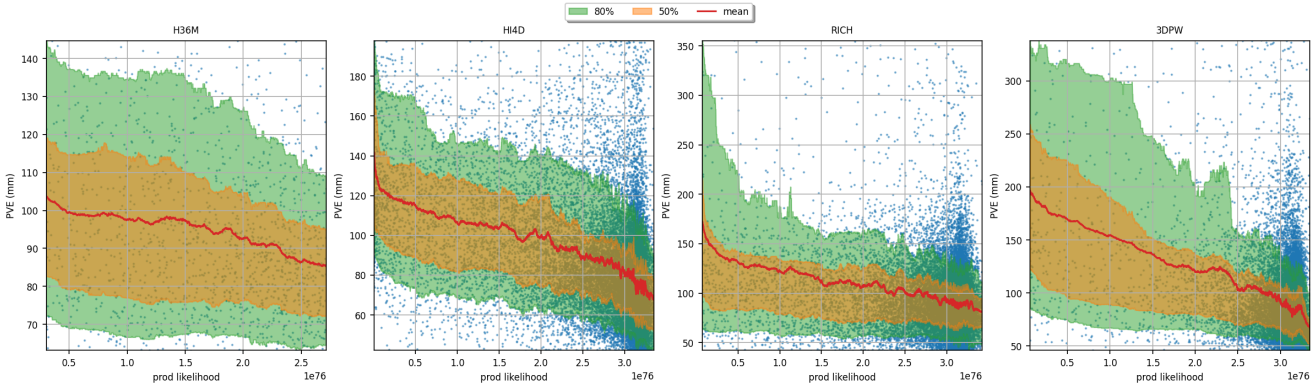
Figure 1. **Counterfactual experiments using different external inputs.** Input image (left) and predictions using different external inputs visualized from camera (middle) and side view (right).



(a) Shape likelihood.



(b) Pose likelihood.



(c) Joint likelihood $p(\hat{t}, \hat{\theta}, \hat{\beta}|\mathcal{I})$.

Figure 2. **Relationship between prediction error and predicted likelihood across datasets.** The predicted likelihood values are correlated with the test error, providing a proxy for prediction confidence. Trend curves (shown in red, yellow, and green) are calculated using a sliding window of 400 samples.

comprising a reconstructed indoor environment, an environment map for background and outdoor lighting, human characters, additional indoor light sources and cameras for rendering. These components were procedurally selected and combined to enhance the realism of the scenes. Specifically, we used scene meshes from Matterport3D [5], Gibson [7] and Habitat [6], along with environment maps from PolyHaven [3]. The characters were generated using HumGen3D [2], a human generator plug-in for Blender. Our synthetic data features a body shape distribution with a thicker tail than BEDLAM for increased diversity, as illustrated in Fig. 6.

Figure 3. **Limitations**. Like other existing approaches, our method struggles with unusual poses far from the training data (top row). Images depicting multiple person with reference keypoints (head) reprojecting at similar 2D locations can lead to missed detections and ambiguous predictions (bottom row).

5.2. Matching

Matching additional inputs To perform experiments with external body shape (resp. distance) inputs, we associate to each prediction the shape (resp. distance) of the closest ground truth annotation, according to the 2D distance between their reference keypoints.

Evaluation For evaluation, we associate each ground truth mesh with the closest prediction, according to their PA-PJE distance.

References

[1] Blender. https://www.blender.org/. 1

[2] Humgen3d. https://www.humgen3d.com/. 3

[3] Poly haven. https://polyhaven.com/. 3

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 6

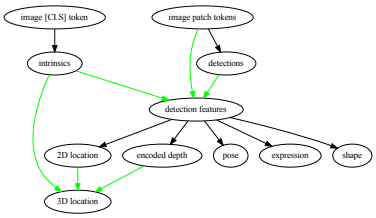
[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *3DV*, 2017. 3

[6] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019. 3

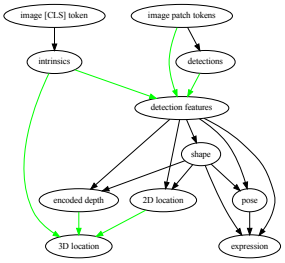
[7] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *CVPR*, 2018. 3

Table 1. **Quantitative evaluation** for different Bayesian network connectivity settings, with different additional inputs: camera intrinsics (*intr*), distance to the camera (*dist*), or known body shape (*shape*).

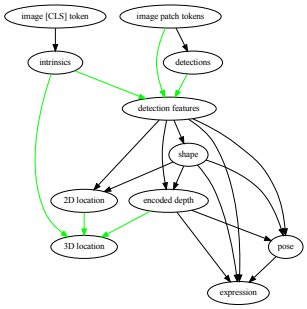
	Experiments		**Human3.6M**			**HI4D**			**RICH**			**3DPW**			**MuPOTS**	
	Ext. input	Connectivity	PVE↓	PA-PVE↓	PE↓	PVE↓	PA-PVE↓	PE↓	PVE↓	PA- PVE↓	PE↓	PJE↓	PA-PJE↓	PE↓	PCK-Matched↑	PCK-All↑
Single-View	none	Naive-Bayes	113.6	55.3	**808.5**	92.8	**47.7**	168.7	119.6	**46.4**	775.3	73.9	47.9	**662.6**	84.8	73.4
		CondiMen	**98.9**	**54.3**	1145.2	**81.4**	48.0	**144.7**	**106.5**	46.6	785.1	**69.2**	**46.4**	728.9	**85.2**	74.5
		Variant1	103.9	58.4	1221.0	91.2	49.0	268.8	130.0	49.8	710.3	71.2	47.4	837.5	85.1	74.5
		Variant2	100.6	55.1	1227.8	101.5	**47.7**	182.1	127.5	46.6	**612.2**	70.2	47.0	908.2	83.9	**74.6**
	intr	Naive-Bayes	104.0	54.1	904.3	93.3	47.7	391.3	118.0	**46.5**	1060.5	73.8	47.7	428.4	84.0	72.7
		CondiMen	**94.3**	**53.9**	648.2	**83.0**	48.0	305.3	**106.6**	46.7	**972.3**	**69.5**	**46.4**	337.2	84.7	74.0
		Variant1	100.3	57.5	**646.8**	90.1	48.6	**286.7**	129.3	49.8	1029.9	71.0	47.2	**298.0**	**85.4**	**74.8**
		Variant2	95.4	54.2	682.8	101.0	**47.5**	394.5	126.1	46.5	1049.6	70.0	46.9	365.6	83.7	74.5
	intr-dist	Naive-Bayes	104.0	54.0	98.1	93.3	47.7	83.3	118.0	**46.4**	115.1	81.2	52.6	112.8	–	–
		CondiMen	**94.3**	**53.9**	89.4	**83.0**	48.0	**69.7**	**106.6**	46.7	**99.1**	76.4	51.2	104.1	–	–
		Variant1	100.8	57.4	90.1	90.1	48.6	77.6	129.7	50.0	124.4	77.5	51.5	**101.3**	–	–
		Variant2	95.4	54.2	**88.5**	100.8	**47.4**	90.7	126.1	46.6	114.8	**75.8**	**51.0**	106.3	–	–
	intr-shape	Naive-Bayes	73.6	54.1	898.0	70.9	47.3	385.0	84.2	47.9	1055.2	73.5	50.0	437.8	–	–
		CondiMen	**70.3**	**53.7**	**284.6**	**62.8**	47.6	132.2	**82.3**	**48.1**	**417.3**	**69.9**	**48.7**	354.2	–	–
		Variant1	80.2	57.4	366.2	62.9	48.2	**113.9**	87.5	51.3	508.7	72.4	49.6	**267.9**	–	–
		Variant2	72.8	54.5	676.9	65.4	**47.1**	382.0	**82.3**	48.1	1038.8	70.4	49.3	375.5	–	–
	intr-shape-dist	Naive-Bayes	73.6	54.1	56.4	70.8	47.3	59.0	84.2	**47.8**	76.8	75.9	51.0	101.6	–	–
		CondiMen	**70.3**	**53.7**	**56.1**	**62.8**	47.6	**46.2**	**82.3**	48.1	73.5	**72.0**	**49.6**	99.8	–	–
		Variant1	80.5	57.4	65.2	**62.8**	48.2	46.4	88.0	51.4	77.9	73.9	50.4	**97.5**	–	–
		Variant2	73.1	54.5	57.9	65.5	**47.1**	51.6	82.9	48.2	**70.2**	**72.0**	50.2	**97.5**	–	–
Multi-View	none	Naive-Bayes	104.8	43.8	**828.8**	85.0	**35.3**	162.2	105.8	37.1	770.5	–	–	–	–	–
		CondiMen	90.6	**42.6**	1148.3	**75.2**	35.5	**142.9**	**93.0**	**36.2**	724.5	–	–	–	–	–
		Variant1	96.0	45.3	1214.8	83.4	**35.3**	275.6	113.4	36.8	622.8	–	–	–	–	–
		Variant2	**88.4**	42.7	1240.7	94.7	**35.3**	181.8	115.4	36.5	**613.0**	–	–	–	–	–
	intr	Naive-Bayes	98.9	43.6	895.9	85.6	35.4	381.8	104.3	37.1	1057.1	–	–	–	–	–
		CondiMen	**88.8**	42.6	679.1	**77.0**	35.5	308.9	**92.8**	**36.2**	**903.4**	–	–	–	–	–
		Variant1	94.8	45.2	668.9	82.8	**35.3**	**281.0**	112.8	36.8	930.7	–	–	–	–	–
		Variant2	84.8	**42.4**	678.8	94.4	**35.3**	385.4	114.1	36.5	1050.0	–	–	–	–	–
	intr-dist	Naive-Bayes	98.9	43.6	96.8	85.5	35.4	78.7	104.3	37.1	103.4	–	–	–	–	–
		CondiMen	88.8	42.6	90.3	77.0	35.5	**68.8**	**92.8**	**36.2**	**89.8**	–	–	–	–	–
		Variant1	95.4	45.4	90.4	**82.9**	**35.3**	75.7	113.5	37.0	115.2	–	–	–	–	–
		Variant2	**84.9**	**42.5**	**81.6**	94.3	**35.3**	88.2	114.2	36.7	107.8	–	–	–	–	–
	intr-shape	Naive-Bayes	67.3	43.9	890.2	63.1	**34.9**	376.3	80.3	39.2	1053.8	–	–	–	–	–
		CondiMen	**62.8**	42.9	**275.0**	57.1	35.2	136.9	**77.2**	**38.6**	**439.0**	–	–	–	–	–
		Variant1	71.8	45.4	350.7	**55.8**	35.0	**116.0**	82.3	39.3	518.2	–	–	–	–	–
		Variant2	65.2	**42.8**	673.7	58.3	35.1	373.6	78.9	38.9	1040.2	–	–	–	–	–
	intr-shape-dist	Naive-Bayes	67.3	43.9	55.8	63.0	**34.9**	54.3	80.3	39.2	75.5	–	–	–	–	–
		CondiMen	**62.8**	**42.9**	**54.2**	57.1	35.2	45.7	**77.2**	**38.6**	74.1	–	–	–	–	–
		Variant1	72.2	45.5	62.5	**55.8**	35.1	**45.0**	83.0	39.5	80.5	–	–	–	–	–
		Variant2	65.5	**42.9**	56.5	58.7	35.1	48.4	79.7	39.1	**72.2**	–	–	–	–	–



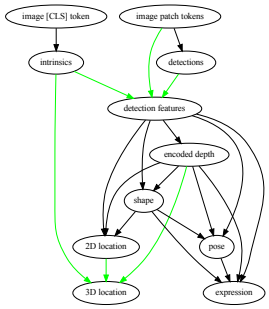
(a) *Naive Bayes*.



(b) *CondiMen* (default).



(c) *Variant1*.



(d) *Variant2*.

Figure 4. **Connectivity of different Bayesian networks considered for this study.** Deterministic dependencies between variables are represented in green.

Figure 5. **Examples of synthetic renderings used in our training.**

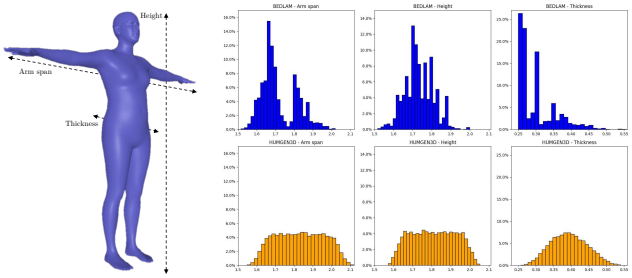


Figure 6. **Body shape statistics** on BEDLAM [4] and our synthetic data.