Utilizing Uncertainty in 2D Pose Detectors for Probabilistic 3D Human Mesh Recovery

Supplementary Material

Tom Wehrbein[1] Marco Rudolph[1] Bodo Rosenhahn[1] Bastian Wandt[2]

[1]Leibniz University Hannover, [2]Linköping University

wehrbein@tnt.uni-hannover.de

A. Implementation Details

Network and training. The normalizing flow (NF) consists of eight RealNVP [3] coupling layers, each parameterized by an MLP with three linear layers of 1024 hidden dimensions and ReLU activations in between. The NF implementation is based on the `FrEIA` package [1] and the soft-clamping parameter is set to $\alpha = 2.0$. Our model is trained for 400K iterations using Adam [7] with weight decay and learning rate set to $1e^{-4}$, and a batch size of 64. Training takes around two days on a single A100 GPU. We use an input image size of 224×224 and apply data augmentation following [2] which includes random crops, scale and different kinds of image blur, compression, and brightness modifications. The loss weights are set to $\lambda_\beta = 5e^{-4}$, $\lambda_{\text{2D}} = 1e^{-2}$, $\lambda_{\text{NLL}} = 1e^{-1}$, $\lambda_{\text{orth}} = 1e^{-1}$, $\lambda_{\text{MMD}} = 5e^{-2}$, $\lambda_{\text{mask}} = 1e^{-1}$.

When using crop or scale data augmentation during training, it would be intuitive to apply it to the 2D pose condition as well by masking (*i.e.* setting to zero) the corresponding keypoints. However, we found it is beneficial to always use the highest-likelihood 2D pose of the original crop as condition. This leads to better generalization, since the model learns to focus more on the 2D pose instead of solely on the image feature.

Since annotations for BEDLAM [2] were initially only released in SMPL-X [11] format, we follow BEDLAM-CLIFF [2] and predict the first 22 body pose parameters of SMPL-X. Hence, our normalizing flow models a distribution of 132 dimensions. We use 11 shape components in the gender-neutral shape space. The SMPL-X labels for the training set of 3DPW [16] are provided by [2]. All evaluation is performed using the SMPL [10] body, by converting predicted SMPL-X meshes to SMPL using a vertex mapping $V \in \mathbb{R}^{10475 \times 6890}$ [11].

Competitors. Since ScoreHypo [17] does not evaluate on EMDB [6], we use their released inference code to calculate the distribution accuracy metrics on EMDB in Table 1 of the main paper. They employ VirtualPose [14] to estimate the root joint depth which is required to transform their predicted 2.5D pose representations to metric 3D space. However, we find that in rare cases VirtualPose fails to predict reasonable depth for the target person or even fails to detect the person at all, resulting in degenerated ScoreHypo outputs. We use the predictions of neighboring frames to fill in missing estimates. Due to the failure cases of VirtualPose, other methods to recover metric scale such as the bone-length optimization method from Pavlakos *et al.* [12] might lead to slightly better results on EMDB. The distribution accuracy metrics for 3DPW [16] are provided by ScoreHypo and we outperform them by a large margin.

To generate the qualitative results for ProHMR [8] in Fig. 1 and Fig. 2 of the main paper, we use our retrained baseline model ProHMR†. This baseline is trained on the same three datasets using the same image backbone as our proposed model, and is more accurate than the officially released checkpoint.

B. Additional Quantitative Results

Number of hypotheses. Fig. S1 shows the Per Vertex Error (PVE) for an increasing amount of hypotheses on 3DPW. The PVE continues to improve significantly when generating more than 100 hypotheses, reaching a PVE of $47.8\,mm$ for 1000 samples compared to $54.4\,mm$ for 100.

Number of heatmap samples. We utilize heatmaps of the 2D pose detector ViTPose [18] to directly supervise the learned distributions of our model using the sample-based loss $\mathcal{L}_{\text{MMD}}$. The loss computes the Maximum Mean Discrepancy between samples drawn from heatmaps and 2D reprojections of random NF hypotheses. To analyze the in-

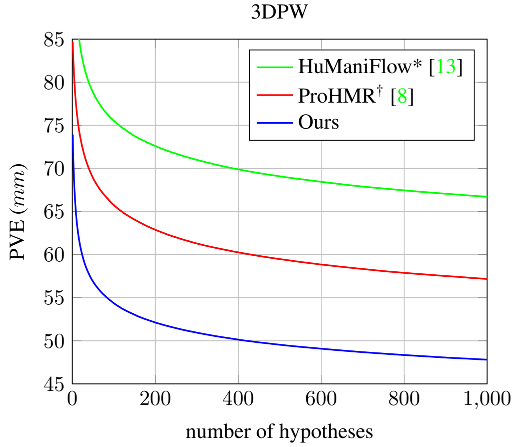

Figure S1. Evaluation results on 3DPW for an increasing number of generated 3D human mesh hypotheses.

	EMDB (24)		
Models	MPJPE ↓	PA-MPJPE ↓	PVE ↓
ProHMR† [8]	76.7	47.1	87.3
+ bbox info [9]	73.1	46.5	82.3
+ 2D pose condition	69.0	44.0	77.9
+ RealNVP	68.5	43.1	77.5
+ $\mathcal{L}_{\text{MMD}}$	63.9	40.7	72.4
+ $\mathcal{L}_{\text{mask}}$ (Ours full)	63.6	40.9	72.0

Table S1. Ablation study analyzing our proposed design choices and loss functions. Components are added successively, and the minimum errors out of 100 hypotheses are reported.

fluence of the number of samples used for $\mathcal{L}_{\text{MMD}}$, we show the performance for different configurations in Fig. S2. The performance first improves with an increasing number of samples, and then remains stable over a wide range. When using only very few samples for computing $\mathcal{L}_{\text{MMD}}$, the model cannot successfully learn to reproduce the distributions encoded in the heatmaps and often predicts distributions with very low diversity. Intuitively, a sufficient number of samples is required to represent the heatmap distributions, while the computational complexity grows with increasing number of samples. As a good trade-off, we use 25 samples in all other experiments.

Ablation study on EMDB. We conduct the ablation study of the main paper on EMDB and present the results in Table S1. Our proposed design choices and loss functions all contribute to the accuracy of the predicted distributions. Notably, despite being added last in the ablation study, the use of $\mathcal{L}_{\text{MMD}}$ results in large improvements.

Detailed $\mathcal{L}_{\text{MMD}}$ ablation study. A main contribution of this work is to directly supervise the learned distributions

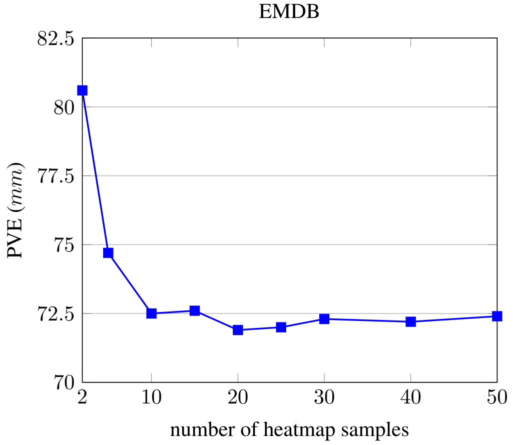

Figure S2. Evaluation results on EMDB for an increasing number of joint samples drawn from the heatmaps for calculating $\mathcal{L}_{\text{MMD}}$. The minimum Per Vertex Error (PVE) of 100 hypotheses is evaluated. Each square denotes a model trained with the specified number of heatmap samples.

by minimizing the distance to distributions encoded in heatmaps of a 2D pose detector [18] using the sample-based distance measure $\mathcal{L}_{\text{MMD}}$. To further analyze the influence of $\mathcal{L}_{\text{MMD}}$, we perform additional experiments on 3DPW and EMDB. The goal is to evaluate different ways of supervising random hypotheses generated by the normalizing flow during training. The mask loss $\mathcal{L}_{\text{mask}}$ is not applied in this study. As a baseline, we first train a model without $\mathcal{L}_{\text{MMD}}$ and where the 2D reprojection loss $\mathcal{L}_{\text{2D}}$ is only computed for the approximated mode prediction. Random NF hypotheses are thus not supervised for this model. Based on this baseline, we train models where either all joints ($\mathcal{L}_{\text{2D-all}}$) or only visible joints ($\mathcal{L}_{\text{2D-vis}}$) of random hypotheses are penalized by minimizing the distance to the ground-truth 2D joints using an l_1 loss. This is done by ProHMR [8] and HuManiFlow [13], respectively. Furthermore, we train a model that receives the embedding proposed in DiffPose [5] as additional condition, which is computed based on samples drawn from the heatmaps. 2D reprojections of random hypotheses are not penalized during training of this model. The distribution accuracy metrics for 100 hypotheses per image are presented in Table S2. Supervising all 2D joints of random hypotheses by minimizing the distance to the ground-truth position has overall no positive impact on the distribution accuracy. On the contrary, it heavily restricts the learned distributions, leading to low sample diversity. When only supervising visible joints using a loss weight of $\lambda = 5e^{-3}$, the metrics slightly improve. While it is intuitive to enforce all visible joints to be at the 2D location of the ground-truth, we find that this also leads to significantly less diversity generated for invisible joints, which has negative influence on the distribution

Supervising	3DPW (14)			EMDB (24)		
random hypotheses	MPJPE $\downarrow$	PA-MPJPE $\downarrow$	PVE $\downarrow$	MPJPE $\downarrow$	PA-MPJPE $\downarrow$	PVE $\downarrow$
no supervision	48.9	32.1	57.4	67.8	43.2	76.8
$\mathcal{L}_{\text{2D-all}}$, $\lambda = 1e^{-3}$	48.1	31.8	56.9	68.5	43.1	77.5
$\mathcal{L}_{\text{2D-all}}$, $\lambda = 5e^{-3}$	51.3	33.3	60.5	71.6	45.4	81.0
$\mathcal{L}_{\text{2D-all}}$, $\lambda = 1e^{-2}$	53.2	34.2	62.9	71.3	46.2	80.8
$\mathcal{L}_{\text{2D-vis}}$, $\lambda = 1e^{-3}$	48.4	31.9	57.3	68.7	43.0	77.4
$\mathcal{L}_{\text{2D-vis}}$, $\lambda = 5e^{-3}$	47.6	31.4	56.3	67.9	42.3	76.6
$\mathcal{L}_{\text{2D-vis}}$, $\lambda = 1e^{-2}$	50.6	32.7	59.5	71.3	45.4	80.8
DiffPose condition [5]	48.1	32.1	57.0	68.7	42.9	77.4
$\mathcal{L}_{\text{MMD}}$ (Ours)	**46.5**	**29.7**	**54.8**	**63.9**	**40.7**	**72.4**

Table S2. Evaluation results for the ablation study on how to best supervise random hypotheses during training. The minimum errors out of 100 hypotheses are reported. Random samples are either not supervised, supervised by minimizing an l_1 loss to the ground-truth 2D positions for either all ($\mathcal{L}_{\text{2D-all}}$) or only visible ($\mathcal{L}_{\text{2D-vis}}$) 2D joints, or by using our proposed $\mathcal{L}_{\text{MMD}}$ loss.

accuracy. Moreover, performance of the models heavily depends on the 2D loss weight. Using DiffPose embeddings as additional condition does not lead to improvements in our setting. Note that in contrast to our setting, the original DiffPose model does not use image features as condition and thus has more incentives to process and exploit the information encoded in the embeddings. Finally, jointwise minimizing the Maximum Mean Discrepancy between 2D reprojections of random hypotheses and samples drawn from heatmaps consistently leads to learned distributions with the highest accuracy. With $\mathcal{L}_{\text{MMD}}$, the learned distributions are explicitly optimized to have high diversity for ambiguous and low diversity for unambiguous joints.

C. Additional Qualitative Results

In the following, we will present additional qualitative results. Predicted camera parameters are used for rendering the 3D human mesh hypotheses and a side-view of each human mesh is created by a rotation of 90° or 270° around the y-axis in camera space.

Failure cases. A few examples of undesirable behavior of our model are depicted in Fig. S3. While optimizing $\mathcal{L}_{\text{mask}}$ significantly decreases the number of incorrect hypotheses, the model still sometimes generates hypotheses where joints are visible that should be invisible. This typically happens for highly ambiguous joints for which the model predicts distributions with very high diversity. We find that using a larger loss weight λ_{mask} for the mask loss can further decrease the number of incorrect hypotheses. However, this comes at a cost of reducing the diversity of the learned distributions too much, resulting in worse accuracy metrics. Finding a way to further reduce the number of incorrect hypotheses while maintaining meaningful diversity could be promising future work. Another typical failure case occurs when the model is presented with highly unusual poses not seen during training. For such examples, high diversity is generated even for unambiguous joints. However, in contrast to deterministic regressors, our model provides information about the prediction uncertainty, either by computing the variance of the hypotheses or by directly calculating their likelihoods. This is useful for downstream tasks that need to know whether the reconstructions results are accurate or not.

Depth ambiguity. Even if all body parts of the person are clearly visible in the image, the depth often cannot be uniquely reconstructed. We show two of such examples in Fig. S4. The predicted hypotheses vary only slightly along the image directions, but have high variance for the depth.

Uncertainty in heatmaps of ViTPose [18]. We visualize heatmap predictions of ViTPose for occluded joints together with 3D mesh hypotheses generated by our model in Fig. S5. The predicted heatmaps encode meaningful joint uncertainty information that our model successfully utilizes during training.

Comparison with competitors. We qualitatively compare the performance of our model with ProHMR† and HuManiFlow by visualizing the reprojections of 100 hypotheses for highly ambiguous joints in Fig. S6. Our model generates more plausible and more meaningfully diverse 3D human mesh hypotheses than the competitors.

D. Limitations and Future Work

Following previous work [13], we define a joint to be invisible if the corresponding heatmap predicted by a 2D pose detector has a maximum value below a certain threshold. While this works well for most cases, we observe that the 2D detector sometimes tends to be overconfident. This calibration gap in 2D human pose estimation frameworks was also recently observed and analyzed by Gu *et al*. [4]. Future work could examine using explicitly predicted joint visibility scores [4,15] instead of maximum heatmap values to decide if a joint is invisible for training and evaluation.

Since we use the distributions encoded in the heatmaps of a 2D pose detector as supervision signal, the performance of our model is influenced by the accuracy of these encoded distributions. Thus, another interesting future research direction would be to improve the distribution modeling capabilities of 2D pose estimators.

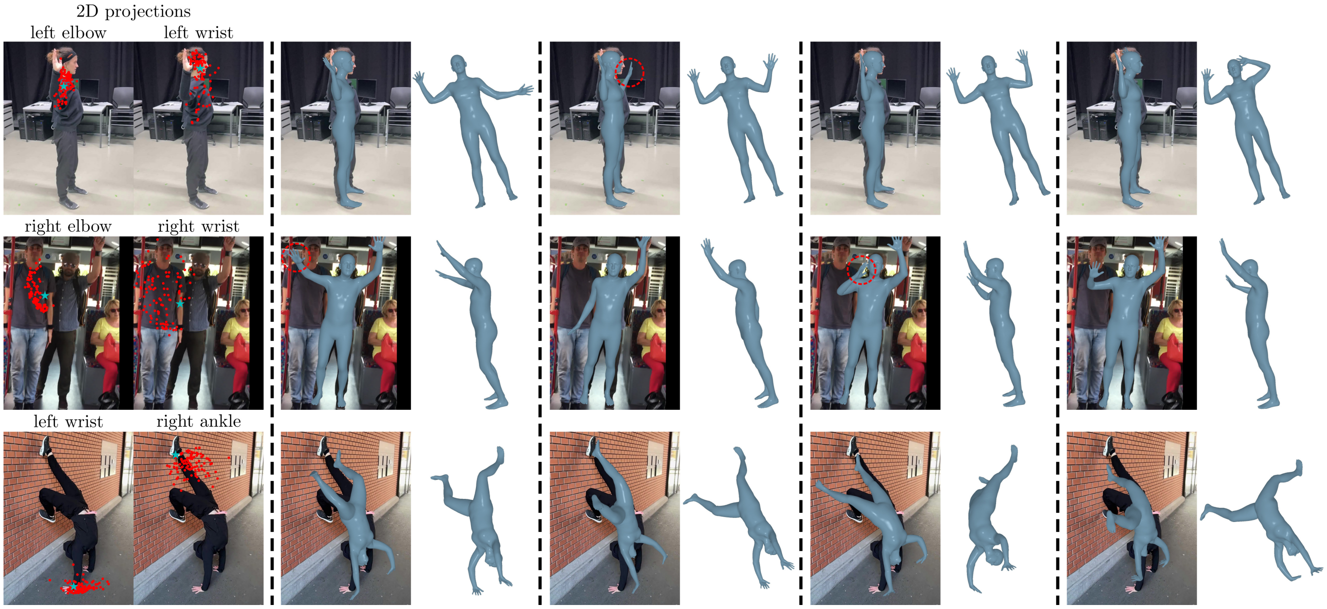

Figure S3. Typical failure cases of our approach. For highly ambiguous joints, our model predicts distributions with very high diversity, sometimes containing a few incorrect samples highlighted with a red circle (rows 1 and 2). The model fails to predict meaningful distributions for very unusual poses not seen during training (row 3).

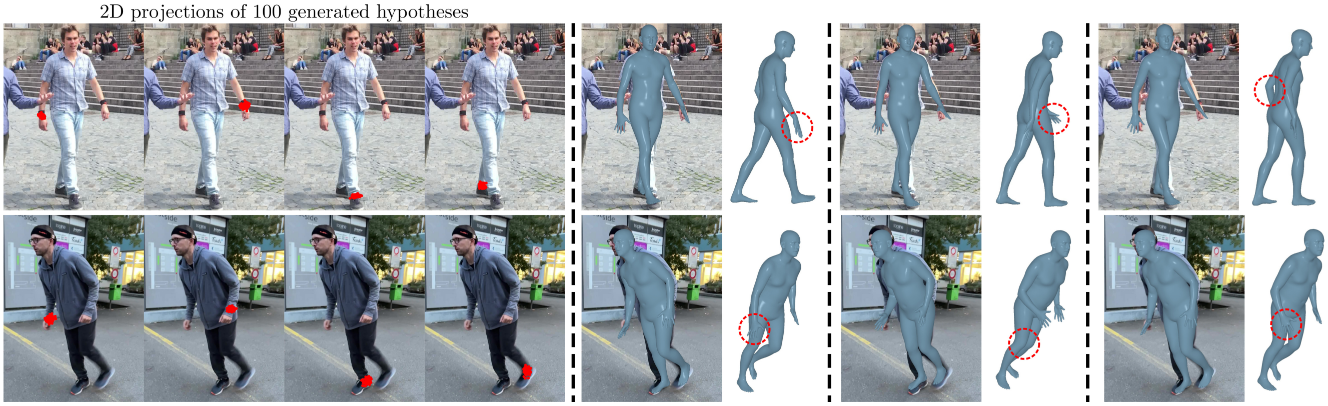

Figure S4. Examples demonstrating depth ambiguity for monocular 3D human mesh estimation. Although all hypotheses vary only slightly along the image directions, significant diversity for the depth is generated. Reprojections of 100 hypotheses for the right wrist, left wrist, right ankle, and left ankle are shown.

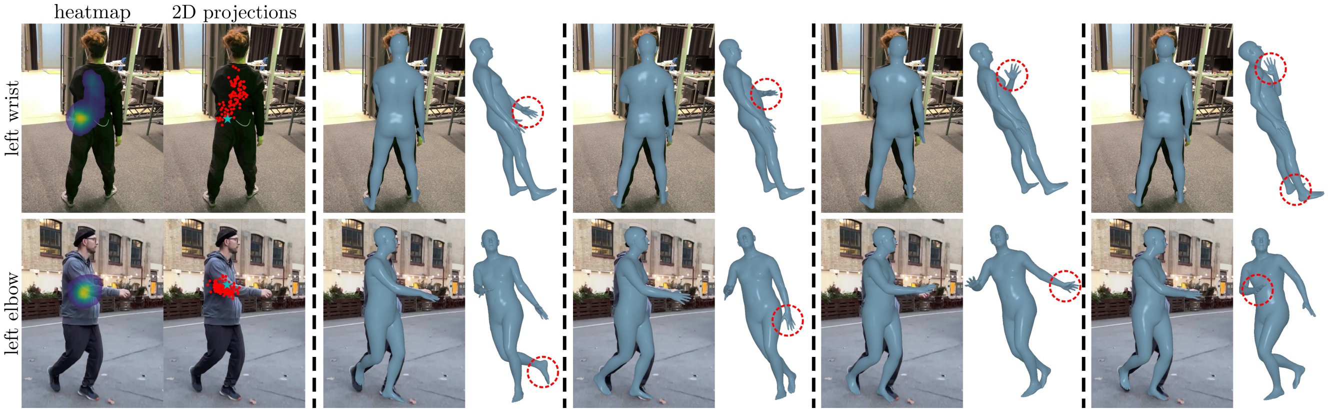

Figure S5. Predicted heatmaps of ViTPose [18] are shown together with 3D human mesh hypotheses generated by our model.

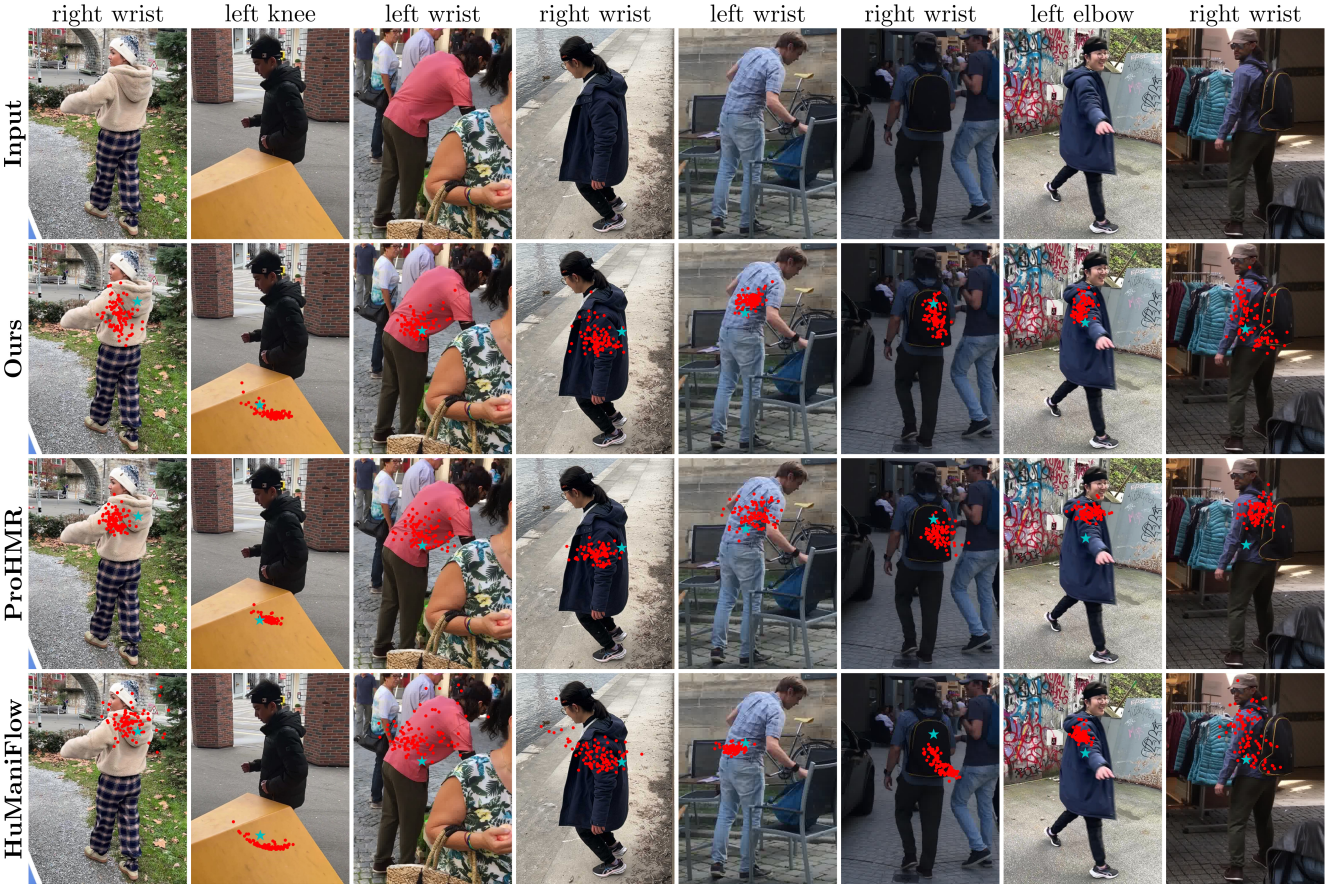

Figure S6. Qualitative comparison with the competing methods ProHMR [8] and HuManiFlow [13]. The 2D reprojections of 100 hypotheses for highly ambiguous joints are shown.

References

[1] Lynton Ardizzone, Till Bungert, Felix Draxler, Ullrich Köthe, Jakob Kruse, Robert Schmier, and Peter Sorrenson. Framework for Easily Invertible Architectures (FrEIA), https://github.com/vislearn/FrEIA, August 2024. 1

[2] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 1

[3] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017. 1

[4] Kerui Gu, Rongyu Chen, Xuanlong Yu, and Angela Yao. On the calibration of human pose estimation. In *ICML*, 2024. 4

[5] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *ICCV*, 2023. 2, 3

[6] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *ICCV*, 2023. 1

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[8] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 1, 2, 6

[9] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 2

[10] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*. ACM, 2015. 1

[11] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1

[12] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017. 1

[13] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Humaniflow: Ancestor-conditioned normalising flows on so(3) manifolds for human pose and shape distribution estimation. In *CVPR*, 2023. 2, 3, 6

[14] Jiajun Su, Chunyu Wang, Xiaoxuan Ma, Wenjun Zeng, and Yizhou Wang. Virtualpose: Learning generalizable 3d human pose models from virtual data. In *ECCV*, 2022. 1

[15] Pengzhan Sun, Kerui Gu, Yunsong Wang, Linlin Yang, and Angela Yao. Rethinking visibility in human pose estimation: Occluded pose reasoning via transformers. In *WACV*, 2024. 4

[16] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018. 1

[17] Yuan Xu, Xiaoxuan Ma, Jiajun Su, Wentao Zhu, Yu Qiao, and Yizhou Wang. Scorehypo: Probabilistic human mesh estimation with hypothesis scoring. In *CVPR*, 2024. 1

[18] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 1, 2, 3, 6