Probabilistic Monocular 3D Human Pose Estimation with Normalizing Flows —Supplemental Material—

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

[1]Leibniz University Hannover, [2]University of British Columbia

wehrbein@tnt.uni-hannover.de

A. Qualitative Evaluation

A.1. Condition Influence

To further show the influence of the heatmap condition and of the loss $\mathcal{L}_{\text{HM}}$ that forces the network to reflect the 2D detector uncertainty in the 3D hypotheses, we present several qualitative results in Fig. 2. Evidently, incorporating the heatmap condition alone already leads to meaningful diversity along the x- and y- directions. Additionally optimizing $\mathcal{L}_{\text{HM}}$ further increases the meaningful diversity of the pose hypotheses such that the uncertainties of the 2D detector as well as the depth ambiguities are modeled best.

A.2. Competitor Comparison

In Fig. 3, we show additional qualitative results comparing our method with the competing methods [1, 2]. As can be seen, our method achieves significantly higher diversity mainly for occluded joints. The competing methods are unable to effectively model occlusions and uncertain detections. They only achieve significant diversity along the ambiguous depth of the joints.

A.3. High Confidence Detections

If the 2D detector has a high degree of confidence for the 2D pose detection in a given image, then low variance in the generated 3D hypotheses along the x- and y- directions is expected. To validate this, we show qualitative results for images from Human3.6M and MPI-INF-3DHP with low 2D detector uncertainty in Fig. 4. The generated hypotheses are shown from two perspectives such that diversity along the image and depth directions can be seen. Evidently, the hypotheses vary only slightly along the image directions and thus are all consistent with the input image. They show meaningful diversity along the ambiguous depth of the joints.

B. Captured 2D Detector Uncertainty

In the following, we want to further verify that fitting a Gaussian to the heatmap can capture the uncertainty of the 2D detector well. Therefore, for each joint in the test split of Human3.6M, we show the mean of the standard deviations of the fitted Gaussian together with the 2D error in Fig. 1. As can be seen, the variances of the Gaussians correlate with the 2D error and thus are a good surrogate for the uncertainty of the 2D detector.

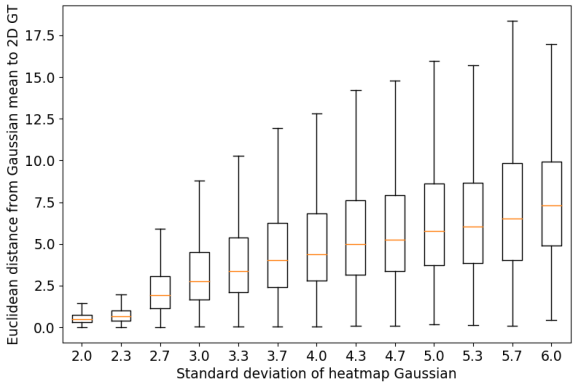

Figure 1. Computed for all joints in the test split of Human3.6M.

Method	Hypo.	MPJPE↓	PMPJPE↓	Method	Hypo.	MPJPE↓	PMPJPE↓
Li [1]	5	52.7	42.6	Ours (K-Means)	5	53.2	38.4
*Li [1]	5	74.9	63.3	*Ours	5	59.2	42.3
*Li [1]	10	70.3	59.7	*Ours	10	55.0	39.6
*Li [1]	200	59.6	50.2	*Ours	200	44.3	32.4

Table 1. Results on Human3.6M under Protocol 1 (MPJPE) and Protocol 2 (PMPJPE). The scores for the rows marked with * are computed by sampling from the models.

C. Performance Lower Number of Samples

To assess the influence of the number of generated hypotheses and make our approach better comparable to Li *et al.* [1], we evaluate on Human3.6M under Protocol 1 (MPJPE) and Protocol 2 (PMPJPE) for lower number of hypotheses in two different settings. However, we want to emphasize that our main goal is to model the full posterior distribution, which requires a larger number of samples. In-

stead of sampling from their model, Li *et al*. [1] take the means of the Gaussian kernels as pose predictions. Thus, for better comparison, we emulate this by running K-Means on our $M = 200$ generated hypotheses. Additionally, we compare the performance when *sampling* from [1] in Table 1 (rows marked with *). We outperform them in almost every setting and metric.

D. Inference Time

For inference time measurements, we run the code with PyTorch 1.7.1 on a NVIDIA GeForce RTX 3090 (CUDA 11.4). The majority of the inference time comes from the 2D detector (32 ms) and the Gaussian fitting process (70 ms). Due to batch processing, generating multiple hypotheses brings nearly no overhead, with an inference time of 4.6 ms for a single and 5.1 ms for 1000 samples.

References

[1] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 4

[2] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 4

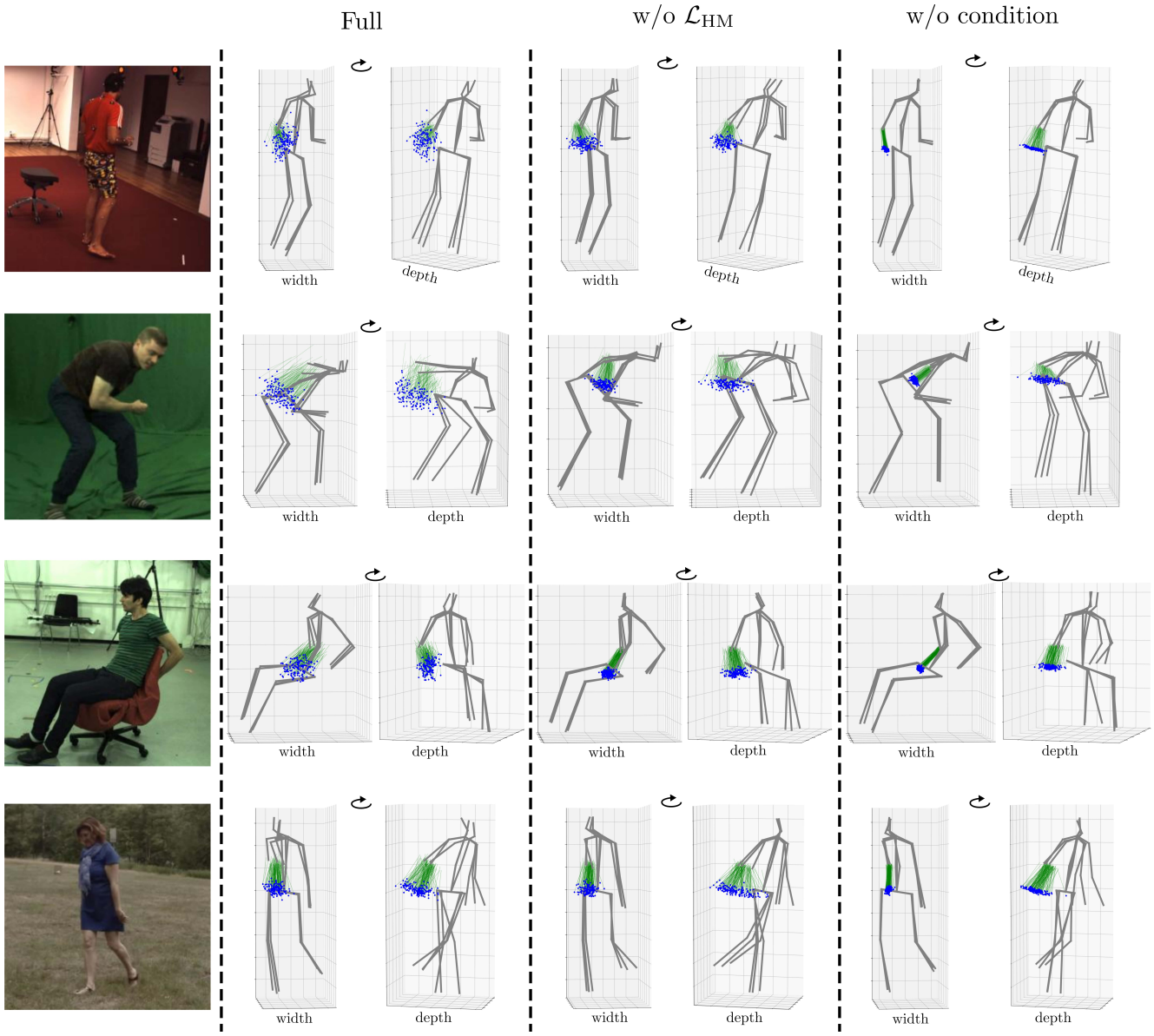

Figure 2. Qualitative results of our full model, model without $\mathcal{L}_{\text{HM}}$, and the model without condition. For visualization purposes, more than three hypotheses are shown only for the most ambiguous joint.

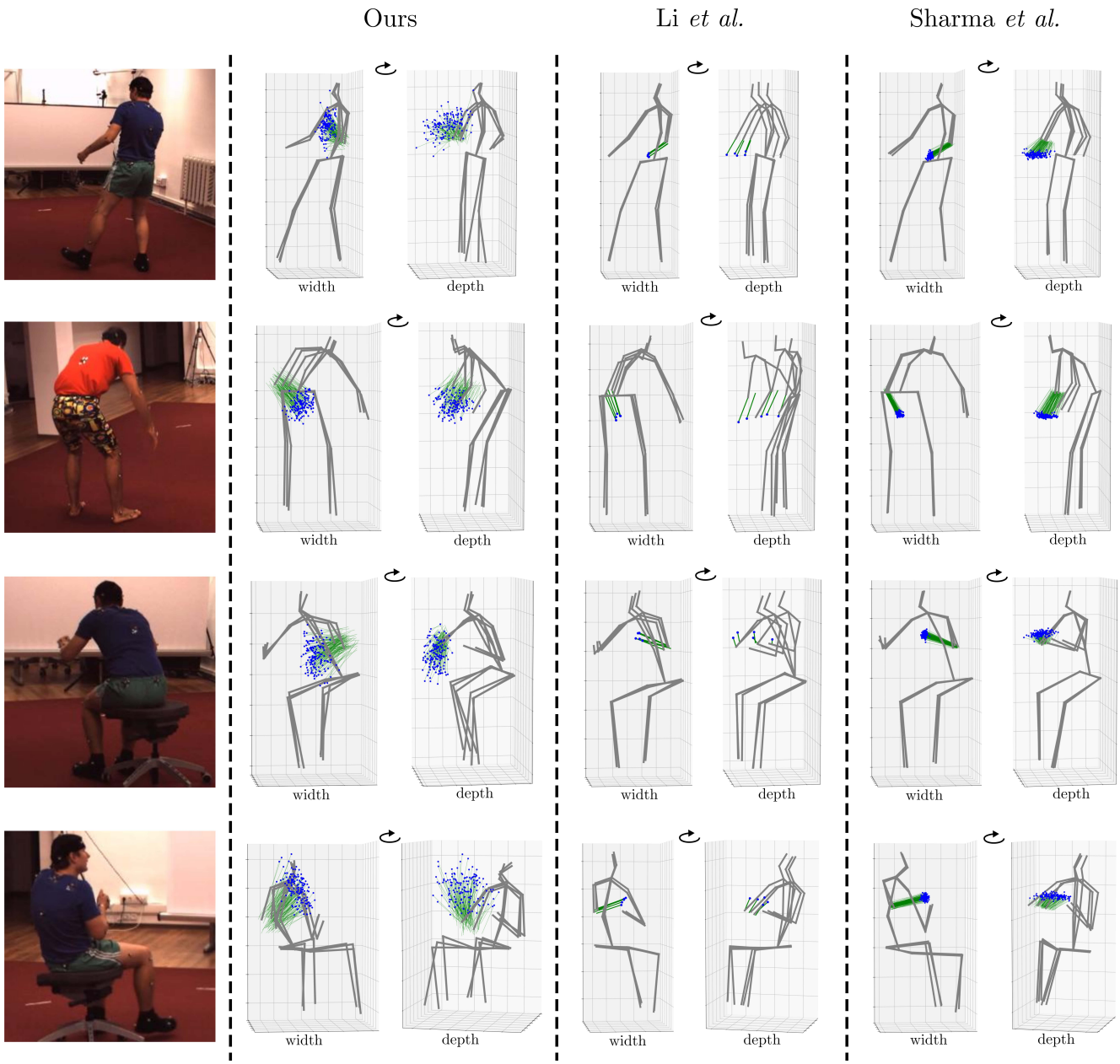

Figure 3. Comparison with competing methods [1, 2]. For visualization purposes, more than three hypotheses are shown only for the most ambiguous joint. The model from Li *et al*. [1] can only generate five pose hypotheses.

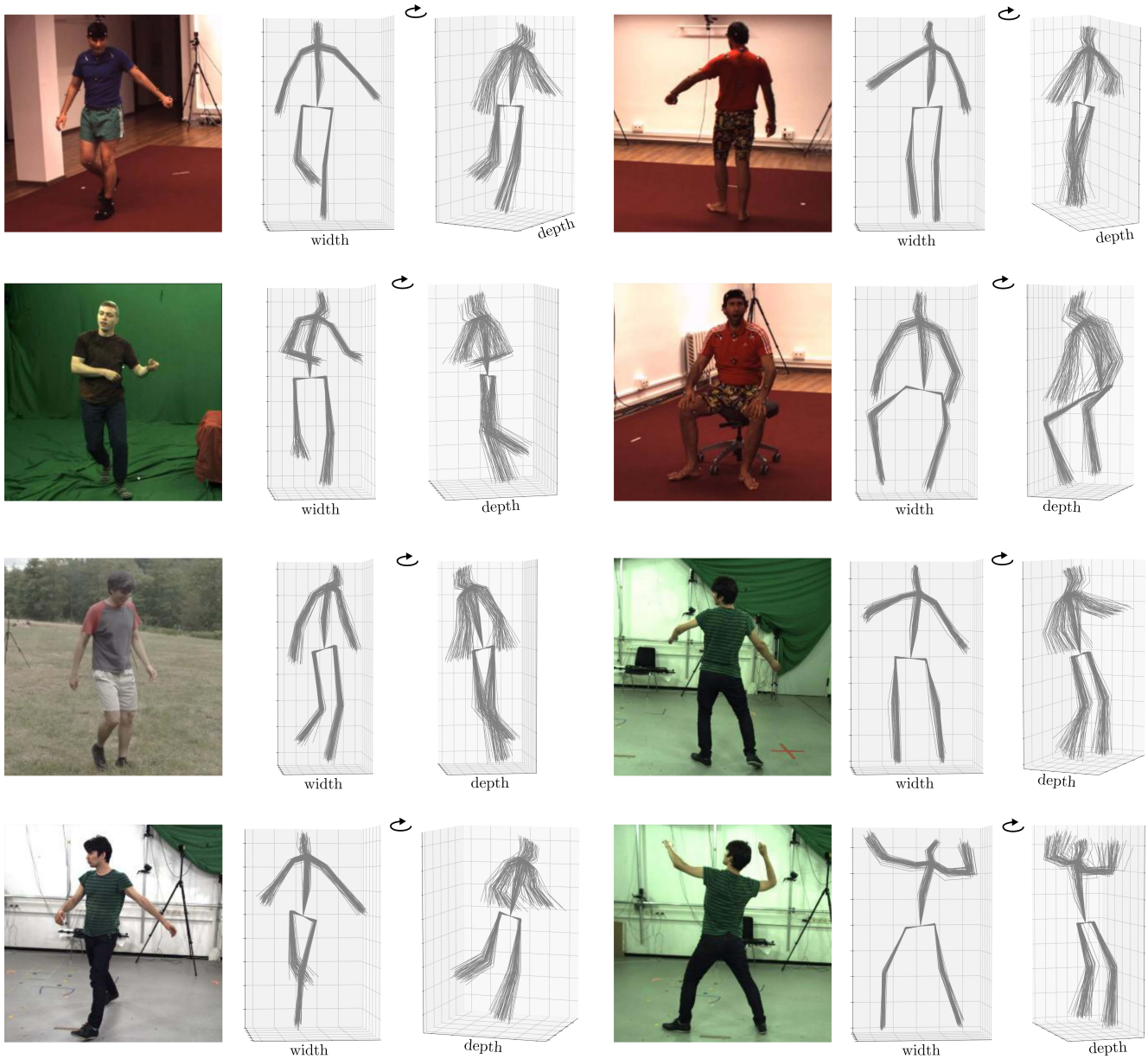

Figure 4. Qualitative results for images from Human3.6M and MPI-INF-3DHP with low 2D detector uncertainty. For each image, 50 pose hypotheses are generated and shown from two perspectives.