Supplementary Material for
DiffPose: Multi-hypothesis Human Pose Estimation using Diffusion Models

1. Additional Results

In addition to the experiments, we report results for Human3.6M (H36M), H36MA and MPI-INF-3DHP (3DHP) for a larger number of hypotheses in Tables 5 and 6, respectively. With an increasing number of samples, the performance of our model increases significantly. Fig. 6 shows this effect visually. Although this behavior is partially expected, our closest competitor [6] shows no such improvements and saturates in performance at approximately 500 samples. This underlines that our model produces more diverse samples that better cover the posterior distribution.

Furthermore, we provide results for MPJPE and PA-MPJPE for the MPI-INF-3DHP dataset in Tab. 6 that were not evaluated by previous methods.

1.1. Non-finetuned HRNet

In addition to our main results, which use the same 2D joint detector as Wehrbein *et al*. [6], we report the results using the non-finetuned network weights from [5] in Tables 6 and 7. It is important to note that most of the methods compared on the full Human3.6M dataset and all on the hard subset (H36MA) utilize the 2D annotations of the H36M dataset to pre-train the 2D joint detector, as done in [2, 1, 4, 6, 3]. Using a non-finetuned 2D pose detector, the results on H36M and H36MA in Tab. 7 are expected to decrease, DiffPose is still competitive, especially on the harder samples in H36MA. Probably much due to the more realistic poses, as indicated by the lower symmetry error and higher CPS.

When investigating the generalizability of the network to the MPI-INF-3DHP dataset, there is a clear improvement of the performance in all settings when not using the fine-tuned weights. We hypothesize that the fine-tuning on H36M overfits to the background and poses in the dataset and is therefore not as well generalizable to other datasets. All PCK metrics improve by $10-20\%$ except for the *Studio GS* setting. Table 6 also shows that an increasing number of hypotheses leads to a drastic increase in the performance of DiffPose when using the original weights.

2. Network Architecture - details

The number of parameters of DiffPose are 11.4M in total (6.4M for the denoiser and 5.0M for the embedding transformer) when using the proposed hyper-parameters.

2.1. Condition Embedding

Positional embedding. We use 64 bases per dimension to create the positional embedding, the centers are evenly spread on the interval [-1, 1]. The embeddings of the x- and y-coordinate are concatenated into a single vector and passed into a linear layer with 128 input and output channels. The embedded joint samples are summed into a joint embedding of dimension 128 before a learned positional joint embedding is added to the embedding. Each individual joint embedding is passed to a transformer encoder with 4 layers, each using 4-heads and a feed-forward dimension of 512. The modified joint embeddings are concatenated into a single 16×128-dimensional feature vector and projected using a linear layer into a 16×128-dim conditioning vector.

2.2. Denoiser

The denoiser concatenates the 16×128-dimensional condition vector, with the 48-dimensional positional vector $\boldsymbol{x}$ and the 1-dimensional timestep (in the range [0, #steps)). This results in a vector of $16\cdot128+48+1=2097$-dimensions that is projected into a 1024-dim vector before being processed by two fully-connected ResNet-blocks (w/o any normalization layer).

2.3. Dropout

Dropout of joints was used during training by randomly selecting joints with a probability of 1% and setting the positional embedding (before the concatenation and the projection-layer) for all samples of those joints to zero. It is possible that multiple or no joints are dropped for a given pose.

2.4. Full heatmap conditioning

As an alternative to our embedding transformer a ResNet-18 network was used for directly generating an embedding vector from the full stack of heatmaps for *Diffusion baseline - full heatmap* in the ablation study. This was accomplished by replacing the first layer of a non-pretrained ResNet18[1], with a 16-channel convolutional layer and passing the full stack of joint heatmaps as an input. Changing the output dimensionality to match the embedding vector used for DiffPose. The performance is drastically worse than DiffPose, as shown in Tab. 4. Given the qualitative re-

[1]ResNet-model from torchvision: https://pytorch.org/vision/stable/models.html

Table 5. Quantitative results on the two H36M datasets. The table illustrates how an increasing number of hypotheses affects the performance of our method. As seen, our method continues to improve, far exceeding competing methods when the number of hypotheses increases.

	H36M		H36MA			
Method	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	PCK ↑	CPS ↑
Li *et al*. [1] (M=5)	52.7	42.6	81.1	66.0	85.7	119.9
Li *et al*. [2] (M=10)	73.9	44.3	-	-	-	-
Sharma *et al*. [4] (M=10)	46.8	37.3	78.3	61.1	88.5	136.4
Wehrbein *et al*. [6] (M=200)	44.3	32.4	71.0	54.2	93.4	171.0
DiffPose (M=200)	$42.9^{\pm 0.27}$	$30.8^{\pm 0.05}$	$63.1^{\pm 0.43}$	$46.7^{\pm 0.14}$	$94.9^{\pm 0.01}$	$195.5^{\pm 3.5}$
DiffPose (M=2000)	$37.9^{\pm 0.17}$	$27.5^{\pm 0.02}$	$56.6^{\pm 0.18}$	$41.6^{\pm 0.07}$	$96.5^{\pm 0.004}$	$212.5^{\pm 2.0}$
DiffPose (M=10000)	$35.3^{\pm 0.14}$	$25.7^{\pm 0.06}$	$52.9^{\pm 0.08}$	$38.9^{\pm 0.05}$	$97.2^{\pm 0.005}$	$221.2^{\pm 1.8}$

Table 6. Quantitative results on MPI-INF-3DHP. This table contains the mean and variance over 5 runs, evaluated using different amount of hypotheses. As can be seen, our method continuously improves with increasing amounts of hypotheses. We also report the results using the same 2D detector but without finetuning on H36M.

Num hypo.	MPJPE	PA-MPJPE	Studio GS ↑	Studio no GS ↑	Outdoor ↑	All PCK ↑
Li *et al*. [2] (M=10)	—	—	86.9	86.6	79.3	85.0
Li *et al*. [1] (M=5)	—	—	70.1	68.2	66.6	67.9
Wehrbein *et al*. [6] (M=200)	—	—	86.6	82.8	82.5	84.3
DiffPose						
M=200	108.3 ± 5.6	66.4 ± 0.4	$87.4^{\pm 0.37}$	$82.7^{\pm 0.16}$	$83.6^{\pm 0.26}$	$84.7^{\pm 0.13}$
M=2000	99.0 ± 4.8	60.8 ± 0.3	$90.2^{\pm 0.39}$	$85.8^{\pm 0.08}$	$86.3^{\pm 0.33}$	$87.6^{\pm 0.08}$
M=10000	93.9 ± 5.4	57.8 ± 0.2	$91.7^{\pm 0.25}$	$87.3^{\pm 0.09}$	$87.5^{\pm 0.4}$	$89.1^{\pm 0.06}$
DiffPose (Not finetuned on H36M):						
M=200	94.7 ± 2.4	64.7 ± 1.4	$87.9^{\pm 0.16}$	$84.9^{\pm 0.21}$	$86.5^{\pm 0.36}$	$86.5^{\pm 0.07}$
M=2000	84.9 ± 1.8	58.6 ± 0.8	$91.3^{\pm 0.12}$	$88.8^{\pm 0.19}$	$89.9^{\pm 0.27}$	$90.0^{\pm 0.10}$
M=10000	80.0 ± 1.8	55.1 ± 2.2	$92.8^{\pm 0.06}$	$90.7^{\pm 0.15}$	$91.4^{\pm 0.20}$	$91.7^{\pm 0.08}$

sults in Fig. 7, we hypothesize this is due to a mode-collapse where only a single pose is predicted for all datasets.

3. Qualitative examples

Fig. 8 shows more qualitative examples. Joints that are easy to detect result in a very clear heatmap and, accordingly, a 3D reconstruction with a low diversity (row 1 and 2). When the 2D joint detector predicts a heatmap with high uncertainty, the method of Wehrbein *et al*. [6] struggles to fully cover it. This leads to

1. massively diverse poses, which are highly implausible (cf. symmetry error in Tab. 2 in the main paper), as shown in row 5, and

2. over-confident predictions of a set of close poses that might be far away from the ground truth, as shown in rows 4, 6, and 8.

Our method compensates for these effects by either predicting some poses that correspond to lower confidence areas or selecting joint positions that are anatomically plausible. An example of anatomical plausibility is shown in row 4 where the position of the elbow is clear but the wrist is wrongly detected. Our model predicts a set of poses where the wrist joint follows an arc, which can be interpreted as a rotation around the elbow joint.

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, 3

[2] Chen Li and Gim Hee Lee. Weakly supervised generative network for multiple 3d human pose hypotheses. *British Machine Vision Conference (BMVC)*, 2020. 1, 2, 3

[3] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 1

[4] 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, 2, 3

[5] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[6] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11199–11208, 2021. 1, 2, 3

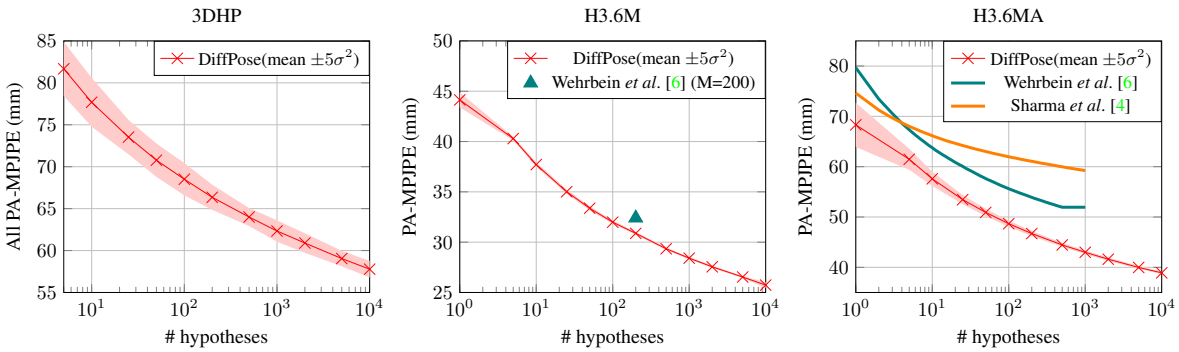

Figure 6. Plots for the three different datasets that illustrate how performance changes as the number of hypotheses generated increases. The results of DiffPose is the average of five different models trained with the same architecture and parameter setting but different random seeds, in addition to the mean we also show the variance as a light-red band of $\pm 5\sigma^2$. Note that the variance is scaled to increase visibility.

Table 7. Results of finetuned and non-finetuned 2D joint detector on the various datasets. Note that all compared methods in this table train or finetune the 2D joint detector on the Human3.6m dataset.

		H36M		H36MA				3DHP			
	Method	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	CPS ↑	Sym ↓	Studio GS ↑	Studio no GS ↑	Outdoor ↑	All PCK ↑
Finetuned 2D detector on H36M	Li *et al.* [1]	52.7	42.6	81.1	66.0	119.9	-	70.1	68.2	66.6	67.9
	Li *et al.* [2]	73.9	44.3	-	-	-	-	86.9	**86.6**	79.3	85.0
	Sharma *et al.* [4]	46.8	37.3	78.3	61.1	136.4	23.9	-	-	-	-
	Wehrbein *et al.* [6]	44.3	32.4	71.0	54.2	171.0	27.4	86.6	82.8	82.5	84.3
	DiffPose	**42.9**$^{\pm 0.27}$	**30.8**$^{\pm 0.05}$	**63.1**$^{\pm 0.43}$	**46.7**$^{\pm 0.14}$	**195.5**$^{\pm 3.5}$	14.9$^{\pm 0.02}$	87.4$^{\pm 0.37}$	82.7$^{\pm 0.16}$	83.6$^{\pm 0.26}$	84.7$^{\pm 0.13}$
	DiffPose	48.1$^{\pm 0.17}$	34.9$^{\pm 0.02}$	75.5$^{\pm 0.75}$	56.6$^{\pm 0.17}$	168.6$^{\pm 2.2}$	**14.3**$^{\pm 0.09}$	**87.9**$^{\pm 0.16}$	84.9$^{\pm 0.21}$	**86.5**$^{\pm 0.36}$	**86.5**$^{\pm 0.07}$

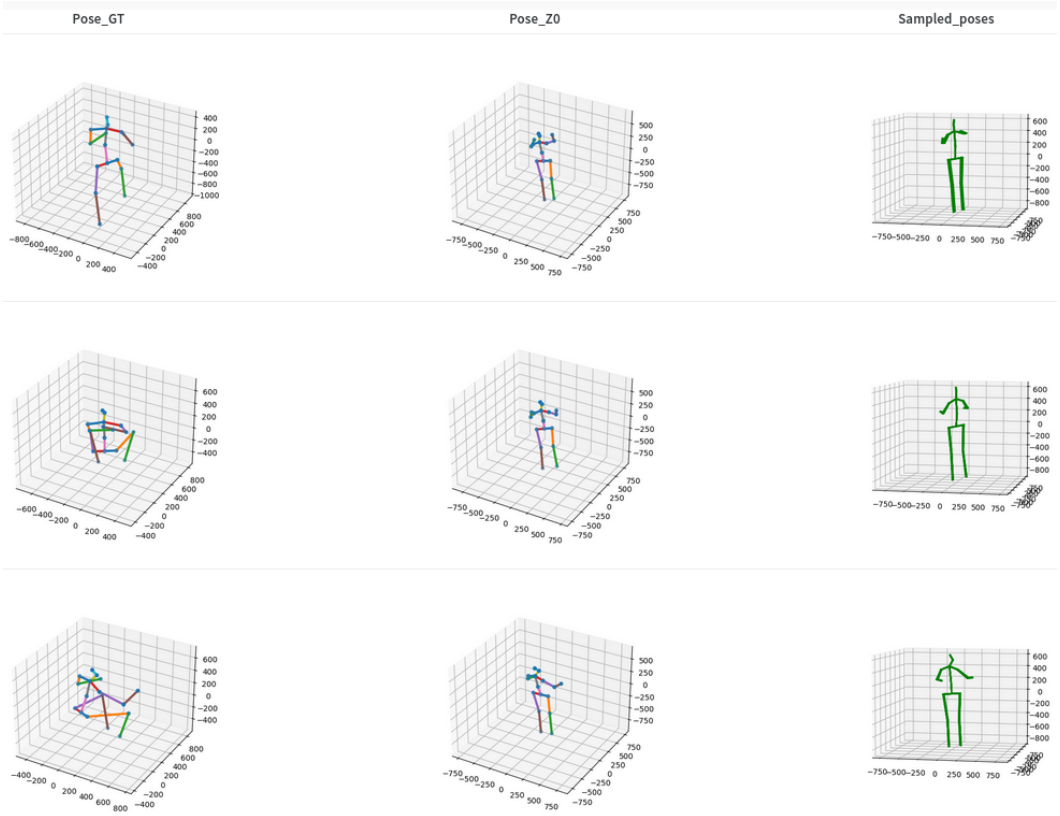

Figure 7. Qualitative examples for ResNet18 embedded heatmaps on H36M. All poses degenerates to a single prediction. Left column is ground-truth poses while right column shows some of the sampled poses.

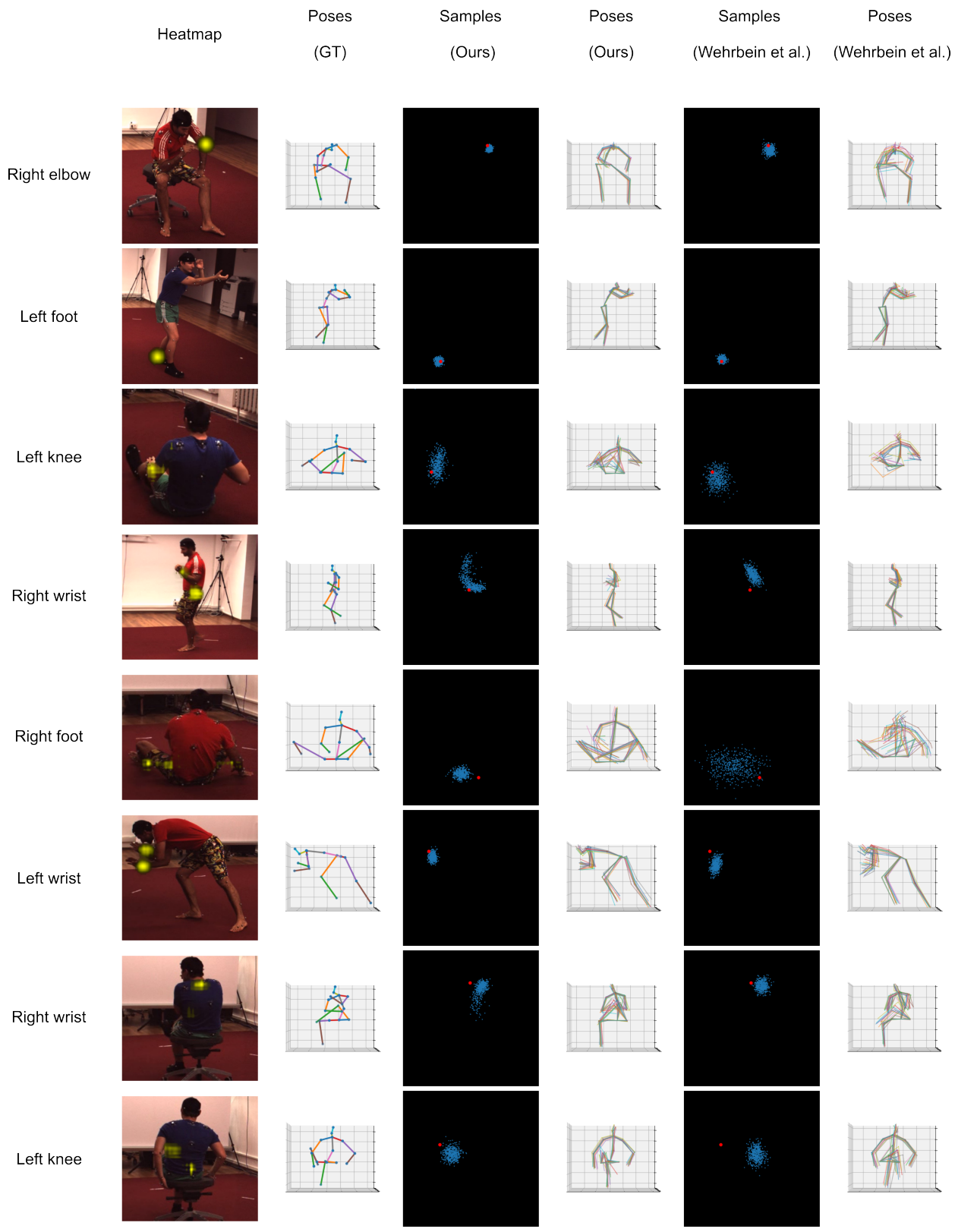

Figure 8. Qualitative examples for H36MA. For better visibility, we only show samples for interesting joints. Our predictions cover the information in the heatmaps well and include the ground truth 3D joint (red dot).