DiffPose: Toward More Reliable 3D Pose Estimation (Supplementary)

Jia Gong[1†] Lin Geng Foo[1†] Zhipeng Fan[2§] Qiuhong Ke[3] Hossein Rahmani[4] Jun Liu[1‡]
[1]Singapore University of Technology and Design
[2]New York University [3]Monash University [4]Lancaster University
{jia_gong,lingeng_foo}@mymail.sutd.edu.sg, zf606@nyu.edu, qiuhong.ke@monash.edu,
h.rahmani@lancaster.ac.uk, jun_liu@sutd.edu.sg

1. Additional Details of GMM Forward Diffusion

In Section 4.2 of the main paper, we describe the GMM-based forward diffusion process. Here, we explain it in more detail, particularly about how it can be framed in a step-wise formulation. We first re-state Eq. 7 in the main paper as follows:

$$\hat{h}_k = \mu^G + \sqrt{\alpha_k}(h_0 - \mu^G) + \sqrt{(1-\alpha_k)} \cdot \epsilon^G. \tag{1}$$

where $\mu^G = \sum_{m=1}^{M} \mathbf{1}_m \mu_m$, $\epsilon^G \sim \mathcal{N}(0, \sum_{m=1}^{M}(\mathbf{1}_m \Sigma_m))$, and $\mathbf{1}_m \in \{0,1\}$ is a binary indicator for the m^{th} component such that $\sum_{m=1}^{M} \mathbf{1}_m = 1$, and $Prob(\mathbf{1}_m = 1) = \pi_m$.

We remark that Eq. 1 directly formulates $\hat{h}_k$ as a function of h_0 instead of $\hat{h}_{k-1}$, because this clearly expresses the aim of our GMM-based forward diffusion design, i.e., such that the generated $\hat{h}_1, ..., \hat{h}_K$ can converge to the fitted GMM model ϕ_{GMM}. Yet, we note that the step-wise formulation of $\hat{h}_k$ in terms of $\hat{h}_{k-1}$ can still be defined, if necessary. First, we sample according to probabilities $\{\pi_m\}_{m=1}^{M}$, and select a Gaussian component $\hat{m}$, i.e., $\mathbf{1}_{\hat{m}} = 1$. Next, we first calculate $\tilde{h}_0$, a "centered" version of h_0, using $\tilde{h}_0 = h_0 - \mu^G$, where $\mu^G = \sum_{m=1}^{M}(\mathbf{1}_m \mu_m) = \mu_{\hat{m}}$. Then, we follow the step-wise formulation as follows:

$$\tilde{h}_k = \sqrt{\frac{\alpha_k}{\alpha_{k-1}}}\tilde{h}_{k-1} + \sqrt{(1 - \frac{\alpha_k}{\alpha_{k-1}})}\epsilon^G, \tag{2}$$

where $\epsilon^G \sim \mathcal{N}(0, \sum_{m=1}^{M}(\mathbf{1}_m \Sigma_m))$, which is equivalent to $\epsilon^G \sim \mathcal{N}(0, \Sigma_{\hat{m}})$. After taking k steps of Eq. 2 starting from $\tilde{h}_0$, we can get:

$$\tilde{h}_k = \sqrt{\alpha_k}(\tilde{h}_0) + \sqrt{(1-\alpha_k)} \cdot \epsilon^G. \tag{3}$$

We observe that the result of the stepwise formulation is thus equivalent to Eq. 1, as we can simply "de-center" our $\tilde{h}_0$ and $\tilde{h}_k$ by substituting $\tilde{h}_0 = h_0 - \mu^G$ and $\tilde{h}_k = \hat{h}_k - \mu^G$.

† Equal contribution; § Currently at Meta; ‡ Corresponding author

2. Additional Details of Diffusion Network g

In order to provide information to the model regarding the current step number k, we generate a diffusion step embedding $f_D^k \in \mathbb{R}^{J \times 256}$ using the sinusoidal function. Specifically, at each even $(2j)$ index of f_D^k, we set the element $f_D^k[2j]$ to $sin(k/10000^{2j/256})$, while at each odd $(2j+1)$ index, we set the element $f_D^k[2j+1]$ to $cos(k/10000^{2j/256})$.

3. More Implementation Details

In the forward diffusion process, we generate the decreasing sequence $\alpha_{1:K}$ via the formula: $\alpha_k = \prod_{i=1}^{k}(1 - \beta_i)$, where $\beta_{1:K}$ is a sequence from $1e-4$ to $2e-3$, which is interpolated by the linear function. To optimize the GMM parameters ϕ_{GMM}, we sample 1000 poses from H_K (i.e., $N_{GMM} = 1000$) and then model H_K via a GMM model.

During model pre-training, the Context Encoder ϕ_{ST} is first pre-trained on the training set to predict 3D poses from 2D poses. Then we adopt the Adam optimizer [7] to train our diffusion model g, where the initial learning rate is set to $1e{-}4$ with a decay rate of 0.9 after ten epochs, and the batch size is set to 4096. Our DiffPose is implemented using PyTorch, and can be trained on a single GeForce RTX 3090 GPU within 96 hours.

4. Experiment Results on Human3.6M under P-MPJPE (Protocol 2)

Tab. 1 and Tab. 2 present the video-based and frame-based results of our DiffPose on Human3.6M under P-MPJPE, where the input 2D poses are detected by CPN [1]. As shown in Tab. 1, our DiffPose can significantly outperform the state-of-the-art methods [8, 21] on all actions with a large margin. Moreover, from Tab. 2, we observe that our method can achieve promising performance on the challenging frame-based setting.

Table 1. Video-based results on Human3.6M with detected 2D poses in millimeters under P-MPJPE.

P-MPJPE	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Lin [9]	32.5	35.3	34.3	36.2	37.8	43.0	33.0	32.2	45.7	51.8	38.4	32.8	37.5	25.8	28.9	36.8
Pavllo [15]	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Liu [12]	32.3	35.2	33.3	35.8	35.9	41.5	33.2	32.7	44.6	50.9	37.0	32.4	37.0	25.2	27.2	35.6
Zheng *et al.* [24]	32.5	34.8	32.6	34.6	35.3	39.5	32.1	32.0	42.8	48.5	34.8	32.4	35.3	24.5	26.0	34.6
Li [8]	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.2	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
Zhang [21]	28.0	30.9	28.6	30.7	30.4	**34.6**	28.6	28.1	37.1	47.3	30.5	29.7	30.5	21.6	20.0	30.6
ours	**26.3**	**29.0**	**26.1**	**27.8**	**28.4**	**34.6**	**26.9**	**26.5**	**36.8**	**39.2**	**29.4**	**26.8**	**28.4**	**18.6**	**19.2**	**28.7**

Table 2. Frame-based results on Human3.6M with detected 2D poses in millimeters under P-MPJPE.

P-MPJPE	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Sun [17]	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	48.3
Martinez [13]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Pavlakos [14]	34.7	39.8	41.8	**38.6**	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Liu [11]	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	31.7	36.2	41.2
ours	**33.9**	**38.2**	**36.0**	39.2	**40.2**	**46.5**	**35.8**	**34.8**	**48.0**	**52.5**	**41.2**	**36.5**	**40.9**	**30.3**	**33.8**	**39.2**

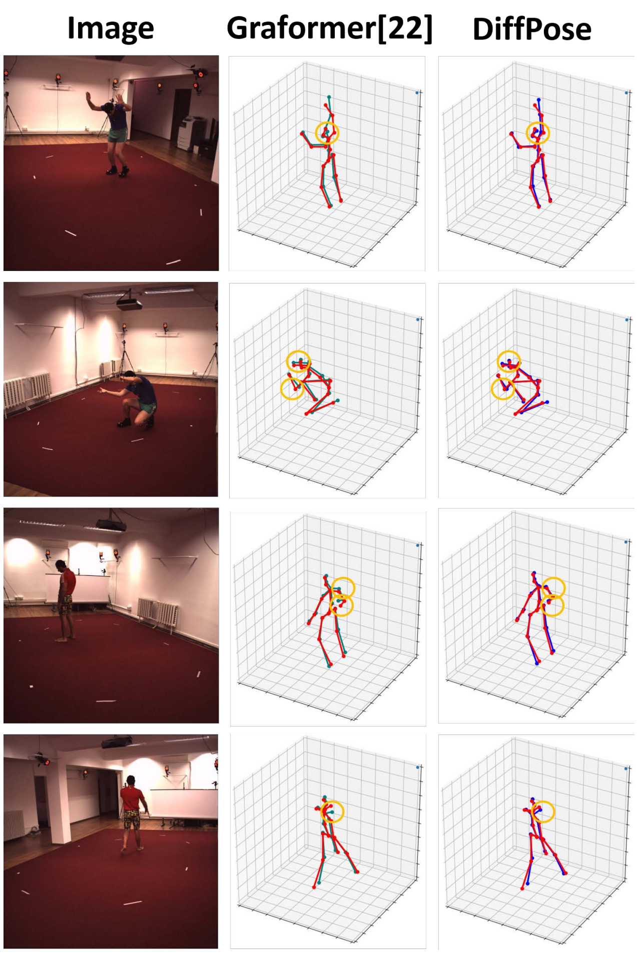

Figure 1. Qualitative comparison between Graformer [22] and our method. Red colored 3D pose corresponds to the ground truth.

5. Additional Results

In this section, we further investigate the performance of our method on the frame-based scenario, by conducting experiments on Human3.6M [6].

3D Pose visualization. First, we qualitatively compare our method with state-of-the-art method [22] in this setting, and present results in Fig. 1. We observe that our method can predict more reliable and accurate poses, especially for novel human gestures (e.g., the first and second rows in Fig. 1) and occluded body parts (e.g., the third and fourth rows in Fig. 1).

Forward diffusion process visualization. Extending from our results in Tab. 5 of the main paper, here we qualitatively compare our GMM-based forward diffusion process with the standard diffusion process (as described in Sec. 3 of our main paper). As shown in Fig. 2, the standard diffusion process recurrently adds noise to the source sample and tends to spread the joints' positions to the whole space. However, our GMM-based diffusion process can add noise according to pose-specific information (obtained from heatmaps) and the data distribution, which generates noise in a more constrained manner. Thus, during training, the GMM-based diffusion process allows us to initialize a $\hat{H}_K$ that captures the uncertainty of the 3D pose, which boosts the performance of DiffPose.

Reverse diffusion process visualization. We visualize the poses reconstructed by our diffusion model with/without the context information f_{ST}. Note that the model without f_{ST} means that no context decoder is used. From the last column of Fig. 3, we observe that both methods can reconstruct realistic human poses while the model with f_{ST} can predict more accurate poses. Moreover, compared to the unconditioned reverse diffusion process (i.e., the model without f_{ST}), the model conditioned by f_{ST} can converge to the desired pose faster.

6. Future Work

In this work, we explore a novel diffusion-based framework to tackle monocular 3D pose estimation. Future work includes more investigations into the architecture of the diffusion network, as well as extending to the online setting [2, 5, 19], the few-shot setting [18, 23] and other pose-based tasks [3, 4, 10, 16, 20].

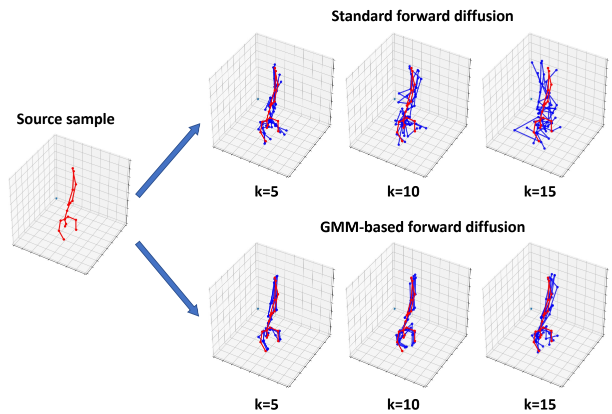

Figure 2. Qualitative comparison between standard diffusion forward process and our GMM-based forward diffusion process.

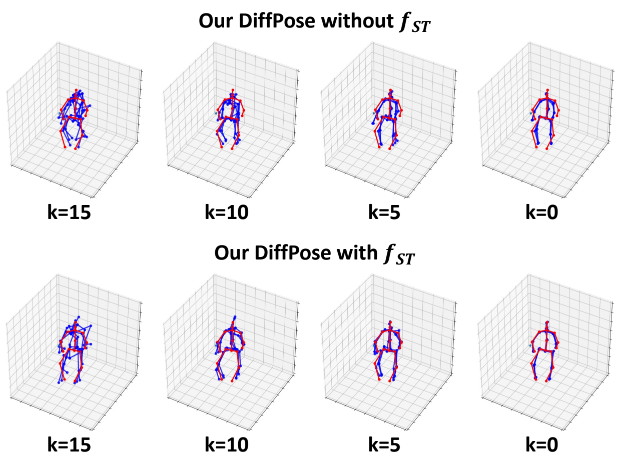

Figure 3. Qualitative comparison between our reverse diffusion process conditioned on context information f_{ST} (bottom), against a reverse diffusion process without using f_{ST} (top).

References

[1] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 1

[2] Lin Geng Foo, Jia Gong, Zhipeng Fan, and Jun Liu. System-status-aware adaptive network for online streaming video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[3] Lin Geng Foo, Tianjiao Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. Era: Expert retrieval and assembly for early action prediction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 670–688. Springer, 2022. 2

[4] Lin Geng Foo, Tianjiao Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. Unified pose sequence modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[5] Amirhossein Habibian, Davide Abati, Taco S Cohen, and Babak Ehteshami Bejnordi. Skip-convolutions for efficient video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2695–2704, 2021. 2

[6] 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*, 36(7):1325–1339, 2013. 2

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[8] 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, 2

[9] Jiahao Lin and Gim Hee Lee. Trajectory space factorization for deep video-based 3d human pose estimation. In *BMVC*, 2019. 2

[10] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016. 2

[11] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *European Conference on Computer Vision*, pages 318–334. Springer, 2020. 2

[12] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5073, 2020. 2

[13] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE ICCV*, pages 2640–2649, 2017. 2

[14] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *IEEE CVPR*, pages 7025–7034, 2017. 2

[15] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 2

[16] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019. 2

[17] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *IEEE ICCV*, pages 2602–2611, 2017. 2

[18] Jonathan Taylor, Jamie Shotton, Toby Sharp, and Andrew Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 103–110. IEEE, 2012. 2

[19] Zuxuan Wu, Caiming Xiong, Yu-Gang Jiang, and Larry S Davis. Liteeval: A coarse-to-fine framework for resource efficient video recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[20] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 2

[21] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13232–13242, 2022. 1, 2

[22] Weixi Zhao, Weiqiang Wang, and Yunjie Tian. Graformer: Graph-oriented transformer for 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20438–20447, 2022. 2

[23] Yunqing Zhao and Ngai-Man Cheung. Fs-ban: Born-again networks for domain generalization few-shot classification. *IEEE Transactions on Image Processing*, 2023. 2

[24] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021. 2