Back to Optimization: Diffusion-based Zero-Shot 3D Human Pose Estimation Supplementary Material

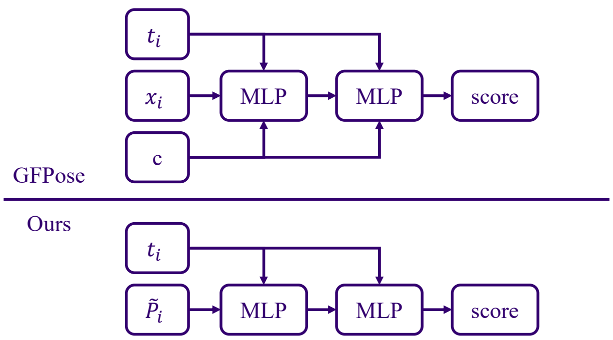

Figure 1. The architecture of GFPose and our diffusion model. Compared with GFPose, there is no pose condition c as input, and the noise x_i is replaced by optimized pose $\widetilde{P}_i$.

1. Architecture Difference with GFPose

As shown in Fig 1, compared with GFPose, there is no pose condition c as input, and the noise x_i is replaced by optimized pose $\widetilde{P}_i$. Our model is not the same as the GFPose. We utilize Score Matching Network to build our human pose generation model.

2. Initial Pose Optimizer

In the initial pose optimizer, our optimization target is

$$\operatorname*{arg\,min}_{R_o, T_o} \quad \left\| K(R_o P_{init} + T_o) - p_{2d} \right\|_2 \tag{1}$$

$$\text{s.t.} \quad T_{min} \leq T_o \leq T_{max}. \tag{2}$$

To solve this optimization problem, we use the Adam optimizer, with the learning rate as 0.1 and optimization iterations as 500. Instead of optimizing the 3×3 rotation matrix, we optimize R_o based on quaternion to ensure the generated $R_o \in \mathrm{SO}(3)$.

3. Optimize Translation

As described in Sec 3, there is a closed-form solution of translation optimization. The optimization target is

$$\operatorname*{arg\,min}_{T_i} \quad \left\| C_{2d}\Big(K(P_i + T_i) - p_{2d}\Big) \right\|_2. \tag{3}$$

The target can be solved by formalizing to

$$\begin{aligned} &\operatorname*{arg\,min}_{T_i} \quad \left\| C_{2d}\Big(K(P_i + T_i) - p_{2d}\Big) \right\|_2 \\ = &\operatorname*{arg\,min}_{T_i} \quad \left\| AT_i - b \right\|_2 \\ &\text{where,} \end{aligned}$$

$$A = \begin{bmatrix} -C_{2d,0} & 0 & C_{2d,0} r_{(0,0)} \\ 0 & -C_{2d,0} & C_{2d,0} r_{(0,1)} \\ & \vdots & \\ -C_{2d,J} & 0 & C_{2d,J} r_{(J,0)} \\ 0 & -C_{2d,J} & C_{2d,J} r_{(J,1)} \end{bmatrix}$$

$$b = \begin{bmatrix} C_{2d,0}(P_{i(0,0)} - P_{i(0,2)} r_{(0,0)}) \\ C_{2d,0}(P_{i(0,1)} - P_{i(0,2)} r_{(0,1)}) \\ \vdots \\ C_{2d,J}(P_{i(J,0)} - P_{i(J,2)} r_{(J,0)}) \\ C_{2d,J}(P_{i(J,1)} - P_{i(J,2)} r_{(J,1)}) \end{bmatrix}$$

$$r = \frac{K^{-1} p_{2d}}{\|K^{-1} p_{2d}\|}$$

The optimization target can be solved as

$$T_i = (A^T A)^{-1} A^T b \tag{4}$$

4. 3D Pose Refinement Results

ZeDO not only has the capacity of denoising pre-defined pose priors but also refines outputs produced by existing 2D-3D lifting networks. In order to validate its effectiveness, we conduct comparative experiments pitting single frame VideoPose3D [9] against our model, aiming to prove that our model could further enhance performance. As demonstrated in 1, we run our mixed-dataset-trained model by taking the keypoint outputs from VideoPose3D as initialization. As a result, we attain lower MPJPE performance on all the datasets, which proves ZeDO's outstanding refinement ability.

Dataset	Methods	MPJPE ↓	PA-MPJPE ↓
3DPW [11]	VPose3D(f=1) [9]	75.9	48.8
	+ ZeDO	**70.2**(-5.7)	**39.3**(-9.5)
H36M [4]	VPose3D(f=1)	39.2	30.4
	+ ZeDO	**38.7**(-0.5)	**27.8**(-2.6)
3DHP [8]	VPose3D(f=1)	89.1	60.5
	+ ZeDO	**78.2**(-10.9)	**51.9**(-8.6)

Table 1. Refinement quantitative results on all three datasets. Our method could further reinforce the performance of the traditional 2D-3D lifting model VideoPose3D [9], in which $f = 1$ represents the single frame scenario. All experiments are $S = 1$. GT 2D poses are used.

5. Results on Ski-Pose Dataset

Ski-Pose [10] is a dataset focusing on ski data, which provides labels for the skiers' 3D poses in each frame and their projected 2D pose in all 20k images. We tested our model as the cross-dataset evaluation on Ski-Pose dataset. As shown in Table 2, we achieve SOTA as PA-MPJPE 81.0mm with the single hypothesis.

Methods	CE	PA-MPJPE ↓	MPJPE ↓
Rhodin *et al*. [10]		85.0	-
Wandt *et al*. [13]		89.6	128.1
Pavllo *et al*. [9]	✓	88.1	106.0
Gong *et al*. [3]	✓	83.5	105.4
Gholami [2]	✓	83.0	99.4
ZeDO ($S = 1$)	✓	81.0	106.3
ZeDO ($S = 50$)	✓	**56.8**	**74.2**

Table 2. 3D HPE quantitative results on Ski-Pose dataset. S indicates the number of hypotheses. All results are reported in millimeters (mm). The pose generation model is trained on Human3.6M. GT 2D poses are used.

6. In Comparison to Unsupervised Methods

We also compared our results with other unsupervised methods on the Human3.6m and 3DPW datasets, as shown in Table 3 and 4. Here, we only applied backbones trained on the Human3.6m dataset for evaluation. Apparently, our method outperforms all of the previous SOTA methods.

7. Model Hyperparameter

Crucial training and inference hyperparameters are displayed in Table 5.

Supervision	Methods	PA-MPJPE ↓	N-MPJPE ↓
GT			
Unsupervised	Chen [1]	58.0	-
	[1]reimplemented by [14]	46.0	-
	Yu [14](temporal)	42.0	85.3
	ElePose [12]	36.7	64.0
	ZeDO ($S = 1$)	**35.8**	**46.9**
DT			
Unsupervised	Kundu [6]	62.4	-
	Kundu [7]	63.8	-
	Chen [14]	68.0	-
	Yu [14]	52.3	92.4
	ElePose [12]	50.2	74.4
	ZeDO ($S = 1$)	**49.0**	**63.6**

Table 3. Quantitative results in comparison with unsupervised methods on Human3.6m dataset. The top table illustrates the results using GT 2D keypoints, and the bottom shows the results of detected 2D inputs. Our model attains top one performance among all unsupervised methods.

Supervision	Methods	PA-MPJPE ↓	N-MPJPE ↓
Unsupervised	ElePose [12]	64.1	93.0
	ZeDO ($S = 1$,$J = 17$)	**40.3**	**60.8**

Table 4. Quantitative results in comparison with unsupervised methods on Dataset 3DPW. GT 2D poses are used. The number of joints is 17.

Hyperparameter	
Batch Size	1024
Training Epoch	2000
Training Optimizer	Adam [5]
Training Learning rate	2e-4
Training Warmup Iterations	5000
Training β_1	0.9
Training β_2	0.999
Inference timestamp t	(0, 0.1]
Inference Iteration Steps	1000
Inference Optimizer	Adam
Optimization Ratotaion Axis	Z
T_{min}	1.6m
T_{max}	16m

Table 5. Important hyperparameters of training and inference on the 3DPW dataset.

References

[1] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *2019 IEEE/CVF Conference on Computer*

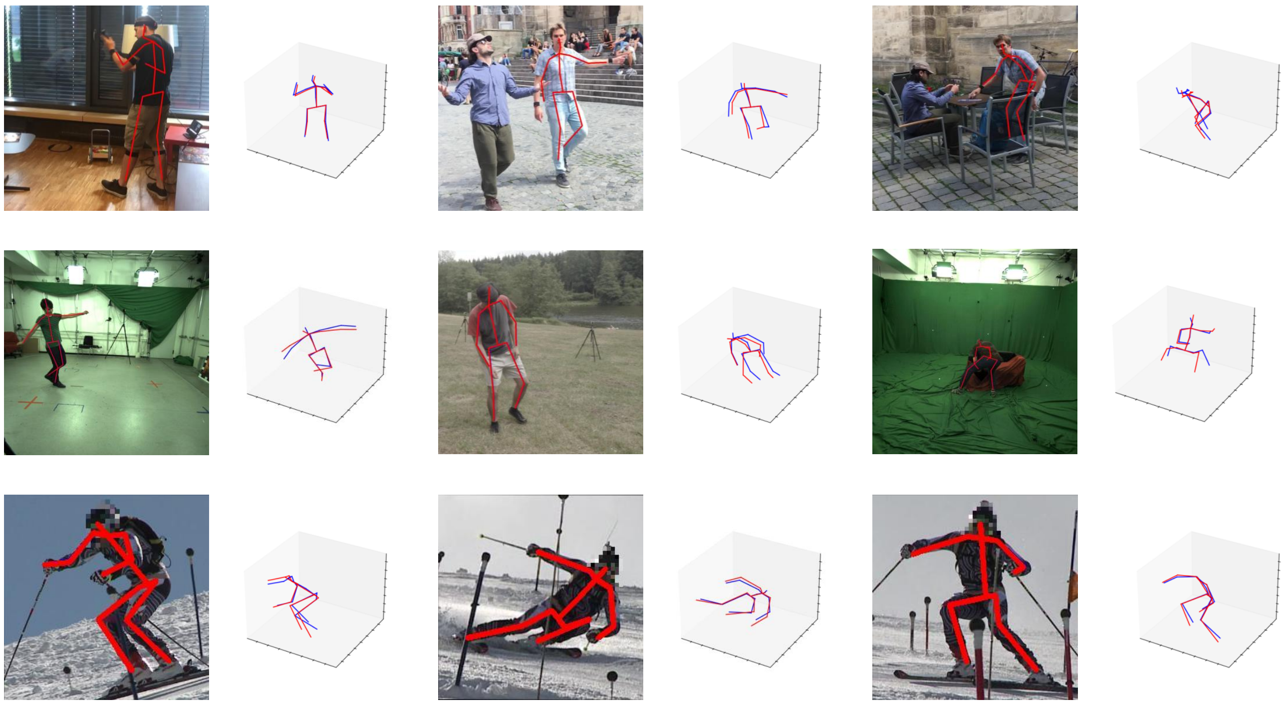

Figure 2. 3D HPE qualitative results on 3DPW, MPI-INF-3DHP and Ski-Pose datasets. First row: 3DPW. Second row: 3DHP. Third row: Ski-Pose.

Vision and Pattern Recognition (CVPR), pages 5707–5717, 2019. 2

[2] Mohsen Gholami, Bastian Wandt, Helge Rhodin, Rabab Ward, and Z Jane Wang. Adaptpose: Cross-dataset adaptation for 3d human pose estimation by learnable motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2022. 2

[3] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8575–8584, 2021. 2

[4] 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

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[6] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6152–6162, 2020. 2

[7] Jogendra Nath Kundu, Siddharth Seth, Mayur Rahul, M. Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. *ArXiv*, abs/2006.14107, 2020. 2

[8] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 2

[9] 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. 1, 2

[10] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8437–8446, 2018. 2

[11] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 2

[12] Bastian Wandt, James J Little, and Helge Rhodin. Elepose: Unsupervised 3d human pose estimation by predicting cam-

era elevation and learning normalizing flows on 2d poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6635–6645, 2022. 2

[13] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304, 2021. 2

[14] Zhenbo Yu, Bingbing Ni, Jingwei Xu, Junjie Wang, Chenglong Zhao, and Wenjun Zhang. Towards alleviating the modeling ambiguity of unsupervised monocular 3d human pose estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8631–8631, 2021. 2