A Study on 3D Human Pose Estimation via Diffusion Models

劉　心毅[†]　　渡辺　裕[†]
Xinyi LIU[†]　and　Hiroshi WATANABE[†]

[†] 早稲田大学大学院　基幹理工学研究科　情報理工・情報通信専攻
[†] GRADUATE SCHOOL OF FUNDAMENTAL SCIENCE AND ENGINEERING,
WASEDA UNIVERSITY

Abstract　This paper proposes a conditional diffusion framework for temporal 3D human pose estimation. The method uses Transformer architecture to refine pose sequences by treating temporal consistency as a denoising process. Experiments on Human3.6M dataset demonstrate significant improvements in pose accuracy while maintaining temporal consistency.

1. Introduction

Single-frame 3D human pose estimation has shown good results in recent years. But temporal inconsistencies remain a major problem when processing video sequences. Most current methods handle video frames separately. This creates jittery and unnatural motion in the final pose sequences.

Recent methods like ZeDO [1] achieve excellent results for single frames. But they cannot maintain temporal consistency across sequences. Other approaches like SmoothNet [2] focus on temporal smoothing after pose estimation. But they lack the strong pose knowledge that diffusion models can provide.

This work proposes a conditional diffusion framework to address these issues. The method uses a Transformer-based [3] architecture for temporal 3D human pose refinement. We treat pose sequence correction as a step-by-step denoising process. The framework uses a Transformer denoising model to process sequential pose data. This helps model temporal dependencies across joint positions well. Our method conditions the diffusion process on initial pose predictions. This way, it learns to fix inconsistencies and keeps underlying motion patterns at the same time.

2. Related Works

2.1 3D Human Pose Estimation

Current 3D pose estimation methods can be split into single-frame and video-based approaches. Single-frame methods like ZeDO represents a recent breakthrough that uses diffusion models as optimization tools. It gets top performance without needing paired 2D-3D training data.

Video-based methods try to use temporal information through temporal convolutional networks and transformer architectures. Methods process spatial and temporal information together. But these approaches still have problems with temporal consistency.

2.2 Diffusion Models

Diffusion models learn to generate data by reversing a noise corruption process. Ho et al. introduced DDPM [4], which gets strong results in various tasks. Song et al. proposed DDIM [5] for faster sampling. This addresses the slow sampling problem of standard diffusion models. Recent work has applied diffusion models to 3D pose estimation.

2.3 Temporal Consistency Solutions

Traditional methods use filtering techniques like Gaussian filtering and Savitzky-Golay filtering for pose smoothing. SmoothNet proposes a plug-and-play temporal refinement network. It reduces jitters in outputs from existing pose estimators. The method uses motion-aware networks to learn temporal relationships for each joint.

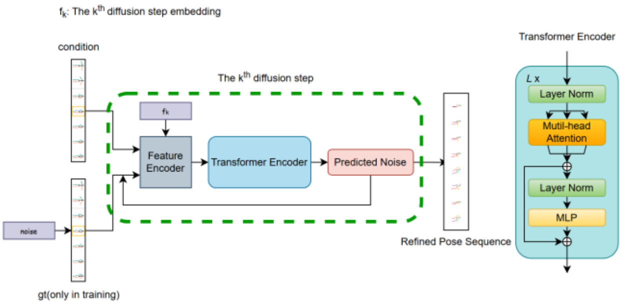

Fig. 1 Overview of the proposed method.

3. Proposed Method

3.1 Temporal Pose Diffusion Framework

Our method extends diffusion models to temporal pose sequences. We set up temporal pose refinement as a conditional sequence-to-sequence denoising task. The key idea is that predicted pose sequences contain useful structural information that can guide the diffusion process toward better solutions.

Input: A predicted pose sequence $X^{pred} = \{x_1, x_2, \dots, x_T\} \in R^{T \times J \times 3}$ from any pose estimator. Here T is sequence length and $J = 17$ is the number of joints.

Output: A refined pose sequence $X^{refined} \in R^{T \times J \times 3}$ with improved temporal consistency and accuracy.

3.2 Forward and Reverse Process

Following the standard DDPM framework, our forward process adds Gaussian noise to ground truth pose sequences step by step. For a ground truth pose sequence, noise versions can be sampled at timestep t using: $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I)$.

The reverse process conditions every denoising step on the predicted pose sequence by $\mathrm{p}_\theta(\mathrm{x}_{t-1} \mid \mathrm{x}_t, \mathrm{c}) = \mathcal{N}(\mathrm{x}_{t-1}; \mu_\theta(\mathrm{x}_t, \mathrm{t}, \mathrm{c}), \sigma_t^2 \mathrm{I})$, where c represents the conditioning information.

3.3 Temporal Denoising Architecture

The denoising network puts together three information sources: current noisy sequence state, diffusion timestep, and condition poses. We use learnable positional embeddings for temporal and joint positions. All encoded features are combined through element-wise addition.

A Transformer encoder processes the fused features with 6 layers, 8 attention heads, and dimension 256. We reshape the feature tensor to create T × 17 tokens per sequence. Each token represents a specific joint at a specific time frame. The self-attention mechanism captures both joint dependencies within frames and temporal relationships across frames.

4. Experiment

4.1 Dataset and Metrics

We use the Human3.6M [6], [7] dataset for evaluation. This dataset contains 3.6 million video frames with accurate 3D pose annotations. We use subjects S1, S5, S6, S7, S8 for training and subjects S9, S11 for testing.

We use pose predictions from ZeDO as input to our temporal refinement approach. ZeDO represents current best practices in single-frame pose estimation. We check results using three metrics: MPJPE measures absolute pose accuracy, P-MPJPE checks pose structure, and MPJAE measures temporal smoothness.

4.2 Results

Table 1. Quantitative comparison on Human3.6M. Mm for MPJPE/P-MPJPE and mm/frames² for MPJAE.

	MPJPE	**P-MPJPE**	**MPJAE**
ZeDO	54.77	37.48	2.52
SmoothNet	53.91	37.45	**0.98**
Ours	**38.91**	**27.18**	2.58

Our method gets big improvements over baseline methods. Compared to ZeDO, our approach reduces MPJPE by 15.86mm (from 54.77mm to 38.91mm) and P-MPJPE by 10.30mm (from 37.48mm to 27.18mm). The method also works better than SmoothNet across MPJPE metrics.

We test different setups to understand key components. Longer sequence lengths give consistent accuracy improvements. Using 32 frames instead of 16 frames shows better results. DDIM sampling offers big speed advantages with 22× speedup but creates temporal quality trade-offs.

Table 2. Ablation study results on Human3.6M.

	MPJPE	**P-MPJPE**	**MPJAE**	**Inference FPS**
16 frame + DDPM	38.91	27.18	2.58	13.9
16 frame + DDIM	39.09	27.23	10.45	306.2
32 frame + DDPM	38.21	27.02	2.38	7.0
32 frame + DDIM	37.77	26.86	7.81	151.5

5. Conclusion

This work presents a conditional diffusion framework that addresses temporal inconsistencies in 3D human pose estimation. The method conditions diffusion models on predicted poses and uses Transformer-based denoising architecture. Experimental results on Human3.6M show large improvements over baseline methods. The framework keeps plug-and-play compatibility for practical deployment and opens new possibilities for temporal motion analysis.

References

[1] Z. Jiang, Z. Zhou, L. Li, W. Chai, C.-Y. Yang, and J.-N. Hwang, “Back to optimization: Diffusion-based zero-shot 3d human pose estimation,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6142–6152, 2024.

[2] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu, “Smoothnet: A plug-and-play network for refining human poses in videos,” in Proceedings of European Conference on Computer Vision, pp. 625–642, 2022.

[3] A. Vaswani et al., “Attention is all you need,” Neural Information Processing Systems, vol. 30, 2017.

[4] J. Ho, A. Jain, and Pieter Abbeel, “Denoising diffusion probabilistic models,” Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020.

[5] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” arXiv preprint arXiv:2010.02502, 2020.

[6] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.

[7] C. S. Catalin Ionescu Fuxin Li, “Latent Structured Models for Human Pose Estimation,” in Proceedings of the International Conference on Computer Vision, 2011.

† 早稲田大学大学院　基幹理工学研究科 情報理工・情報通信専攻

〒169-0072　東京都新宿区大久保　3-14-9　早大シルマンホール　401 号室

TEL.090-8120-7984　E-mail: liuxy@toki.waseda.jp