PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation – Supplementary Material

Qitao Zhao[1]* Ce Zheng[2] Mengyuan Liu[3] Pichao Wang[4] Chen Chen[2]

[1]Shandong University [2]Center for Research in Computer Vision, University of Central Florida [4]Amazon Prime Video

[3]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School

qitaozhao@mail.sdu.edu.cn cezheng@knights.ucf.edu nkliuyifang@gmail.com

pichaowang@gmail.com chen.chen@crcv.ucf.edu

A. Overview

The supplementary material includes sections as follows:

- Section B: A formal introduction to Discrete Cosine Transform.
- Section C: Datasets and evaluation metrics.
- Section D: More implementation details.
- Section E: Comparisons of PoseFormerV2 and a simple baseline model purely in the frequency domain.
- Section F: Generalization of our approach to more models.
- Section G: Visualizations and analysis.
- Section H: Broader impacts and limitations.

B. Discrete Cosine Transform

We now give a formal introduction to DCT. Given a 2D joint sequence denoted by $\mathbf{x} \in \mathbb{R}^{F \times J \times 2}$, where F is the sequence length and J is the joint number in each frame, the trajectory of the x (or y) coordinate of the j-th joint denoted as $\mathbf{x}_{j,0} \in \mathbb{R}^F$ (or $\mathbf{x}_{j,1} \in \mathbb{R}^F$, both denoted by $\hat{\mathbf{x}}_j$ for convenience) is a 1D time series and we apply DCT to each trajectory ($J * 2$ trajectories in total) individually.

For trajectory $\hat{\mathbf{x}}_j$, the i-th DCT coefficient is calculated as

$$C_{j,i} = \sqrt{\frac{2}{F}} \sum_{f=1}^{F} x_{j,f} \frac{1}{\sqrt{1+\delta_{i1}}} \cos\left(\frac{\pi}{2F}(2f-1)(i-1)\right) , \quad (1)$$

where $\delta_{i1} = 1 \; when \; i = 1, \; otherwise \; \delta_{i1} = 0$. Each time step in trajectory yields one DCT coefficient, *i.e.*, $i \in \{1, 2, \cdots, F\}$. DCT coefficients encode multiple levels of temporal information in the input time series. Specifically, low-frequency coefficients (*i.e.*, when i is small) encode the rough contour of the input sequence while high-frequency coefficients (*i.e.*, for the large i) encode details, *e.g.*, jitters or sharp changes in the input sequence. The original input sequence in the time domain can be restored using Inverse Discrete Cosine Transform (IDCT), which is given by

$$x_{j,f} = \sqrt{\frac{2}{F}} \sum_{i=1}^{F} C_{j,i} \frac{1}{\sqrt{1+\delta_{i1}}} \cos\left(\frac{\pi}{2F}(2f-1)(i-1)\right) , \quad (2)$$

and $f \in \{1, 2, \cdots, F\}$. DCT is lossless if we keep all its coefficients intact. In practice, we can slightly lossily recover the input sequence using only a few low-frequency coefficients and set other coefficients to zero. It is worth noting that the recovered curve would be smoother compared to the original one since we discard some of the high-frequency coefficients. This property of DCT is desirable – only a small proportion of DCT coefficients are enough to represent the whole input sequence, even in a cleaner manner. This motivates us to use such representation to efficiently operate long sequences while improving the robustness of the model to low-quality 2D detection where high-frequency noise often occurs.

C. Datasets and Evaluation Metrics

Human3.6M is the most widely used benchmark for 3D human pose estimation. Over 3.6 million video frames are captured indoors from 4 cameras at different places. This dataset contains 11 subjects performing 15 different actions, *e.g.*, "Walking" and "Phoning". We train our model on 5 subjects (S1, S5, S6, S7, S8) and use other 2 subjects (S9, S11) for testing, following [1, 6, 10, 15].

MPI-INF-3DHP is collected in both controlled indoor environments and challenging outdoor environments. It also provides different subjects and actions from multiple camera views similar to Human3.6M.

Evaluation Metrics. We report two common metrics, MPJPE and P-MPJPE [14] on Human3.6M. MPJPE (Mean Per Joint Position Error, referred to as Protocol 1) measures the mean Euclidean distance between the estimated 3D pose and the ground truth 3D pose. P-MPJPE (Protocol 2) applies a rigid transformation to the estimated 3D pose and

*Work was done while Qitao was an intern mentored by Chen Chen.

the distance is computed between the aligned estimated 3D pose and the ground truth 3D pose.

For the MPI-INF-3DHP dataset, we report MPJPE, Percentage of Correct Keypoint (PCK) within the 150mm range, and Area Under Curve (AUC) as in [1, 5, 12].

D. More Implementation Details

Our method is built upon PoseFormerV1 [15]. Aiming at better demonstrating the effectiveness of our DCT coefficient representation of input sequences and providing fair comparisons to PoseFormerV1, we **directly** adopt optimal hyper-parameters for model architecture from PoseFormerV1, although further investigation may bring additional improvements.

Model architecture hyper-parameters. The embedded feature dimension c in the spatial transformer is 32 and the layer number of the spatial transformer and feature-fusion transformer is 4, following [15]. Plus, the design of Spatial-Temporal Positional Embedding is also adopted from [15].

Experimental settings. Our experiments are conducted with Pytorch [9] on a single NVIDIA RTX 3090. For both training and testing, we apply horizontal flipping augmentation following [1, 6, 10, 15]. We train our model using the AdamW [7] optimizer for 80 epochs with a weight decay of 0.1. The initial learning rate is set to 8e-4 with an exponential learning rate decay schedule and the decay factor is 0.99. We adopt the CPN [2] 2D pose detection on Human3.6M, following [1, 6, 10]. As for the MPI-INF-3DHP dataset, we use ground truth 2D detection, following [5, 8].

E. Simple Baseline

In our approach, the temporal encoder of PoseFormerV1 [15] is reformulated as a Time-Frequency Feature Fusion module and we show that the low-frequency coefficients of the input sequence help improve the efficiency of the model to process long sequences and its robustness against noisy joint detection. Given the effectiveness of this representation, readers may raise a question: Why not entirely extract features from DCT coefficients of the input sequence but additionally combine them with features in the time domain? Here we design a baseline model where we simply replace the input to PoseFormerV1 [15] (joint coordinates in the time domain) with low-frequency DCT coefficients of the input sequence. The full sequence length and the number of the retained DCT coefficients (denoted as n) are kept the same for the baseline model and our approach. For convenience, the number of frames (f) as input into the spatial encoder of PoseFormerV2 is set to n. We provide quantitative results to demonstrate that this straightforward approach does not work well, especially when the ratio between the full sequence length and n is increased (see Table 1). The features of only a few central frames in the sequence significantly boost accuracy, *e.g.*, with 3 central frames of the full input sequence of length 81, the MPJPE is reduced from 49.7mm to 47.1mm (5.2%↓, the 3rd row in Table 1).

Table 1. Comparisons of PoseFormerV2 and a simple baseline. The evaluation is performed on Human3.6M (Protocol 1, MPJPE) [3] and the Frame Number (f) is only applicable to PoseFormerV2.

Frame Number (f)	Coefficient Number (n)	Full Length	Baseline	PoseFormerV2
3	3	9	50.2	49.5 (0.7↓)
3	3	27	48.7	47.9 (0.8↓)
3	3	81	49.7	47.1 (2.6↓)
9	9	27	48.8	47.6 (1.2↓)
9	9	81	47.8	46.0 (1.8↓)

Intuitively, the spatial encoder of PoseFormerV2 that encodes joint coordinates of a few central frames in the time domain helps capture the fine-grained human motions, benefiting 3D pose estimation for the frame at the sequence center. In contrast, low-frequency coefficients of the input sequence filter out high-frequency noise and human motion details (*e.g.*, fast motions) that may be informative to human pose estimation (*i.e.*, the over-smoothing problem). Therefore, features from the time domain and frequency domain, *i.e.*, the joint coordinate of central frames and low-frequency coefficients of the sequence, carry complementary semantics. These considerations necessitate our proposed Time-Frequency Feature Fusion design.

F. Generalization to More Models

In the main text, we focus on improving PoseFormerV1 [15] from a barely explored frequency-domain perspective. In this part, we show that the proposed frequency-domain approach also generalizes well to other existing state-of-the-art methods, *e.g.*, MixSTE [13] and MHFormer [4]. Since these approaches [4, 13] also apply self-attention along the time dimension to all frames as PoseFormerV1 [15], the proposed method can be easily incorporated into their model without complex redesigns for model architecture. For fair comparisons, we **directly** adopt optimal hyper-parameters (*e.g.*, the layer number, channel dimension) for these original methods. Further tuning of hyper-parameters may bring additional improvements.

MixSTE [13] adopts the spatial-temporal architecture as PoseFormerV1 [15]. Compared to the spatial-then-temporal paradigm of PoseFormerV1, MixSTE alternately uses spatial and temporal transformer encoders. Similarly, we centrally sample a few video frames from a longer sequence as input into the spatial encoders of MixSTE. For temporal encoders, we append the time-domain features (the output of the spatial encoders) with the embedding of low-frequency coefficients of the complete input sequence. The comparisons of MixSTE and its improved version are presented in Fig. 1, 2. Original MixSTE is highly com-

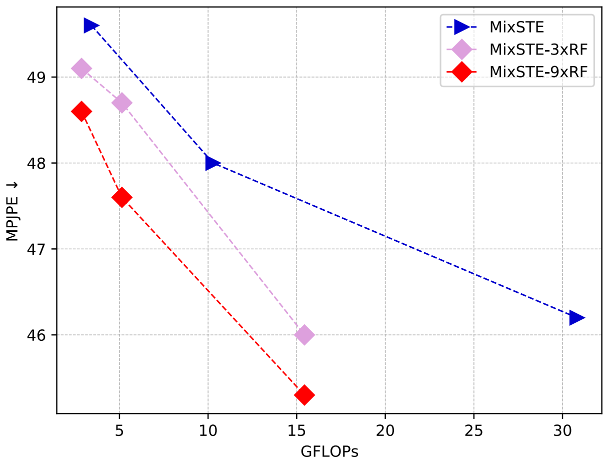

Figure 1. Comparisons of MixSTE [13] and its improved version with frequency representations of the sequence on Human3.6M [3]. RF: Receptive Field and $k\times$RF indicate that the RF of MixSTE is expanded by $k\times$ with a few low-frequency DCT coefficients of the full sequence. The proposed approach helps MixSTE gain a better speed-accuracy trade-off. (Best viewed in color)

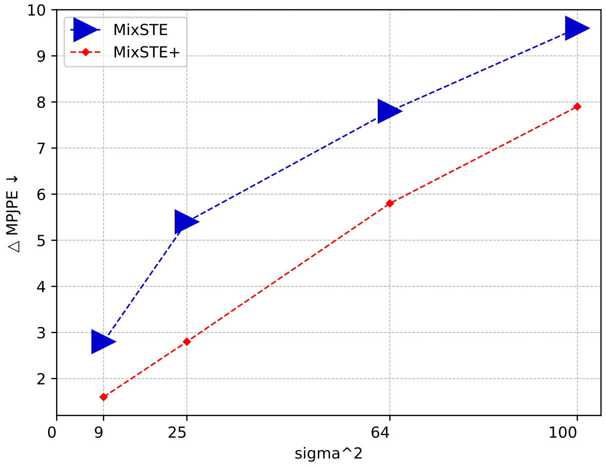

Figure 2. Comparisons of MixSTE [13] and its improved version using low-frequency DCT coefficients of the sequence in terms of robustness to noise on Human3.6M [3]. Zero-mean Gaussian noise of standard deviation $sigma$ is added to ground truth 2D detection, and we show their performance drop (△MPJPE) as $sigma$ increases. The size of markers indicates the computational cost of models.

putationally expensive and our approach improves its efficiency and accuracy simultaneously, *e.g.*, MixSTE achieves 46.2mm MPJPE taking 30.8 GFLOPs, while its improved version achieves 45.3mm MPJPE with 15.4 GFLOPs (2× faster and 1.9%↑ error reduction, see the bright red curve in Fig. 1). We also show that our method improves the robustness of MixSTE against noisy 2D joint detection (Fig. 2). Specifically, we add zero-mean Gaussian noise to the ground-truth 2D joint sequence of 27 frames on Human3.6M [3]. The improved MixSTE (denoted as MixSTE+) uses 3 central frames as input to its spatial encoders and the first 3 DCT coefficients as a cleaner global representation of the full sequence. MixSTE+ suffers from less performance drop while being 6× more efficient (30.8 GFLOPs *vs*. 5.1 GFLOPs, indicated by the marker size).

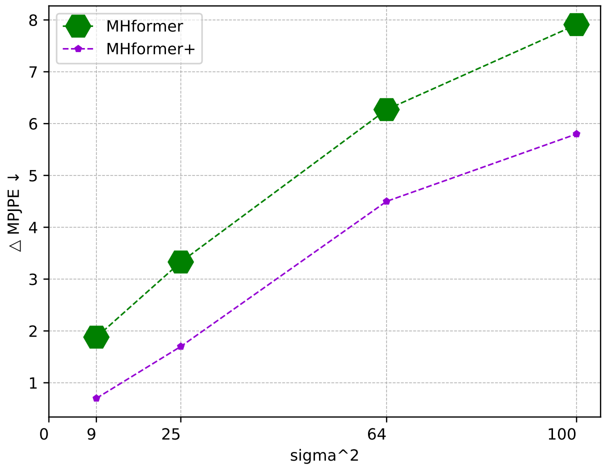

Figure 3. Comparisons of MHFormer [4] and its improved version using low-frequency DCT coefficients of the sequence in terms of robustness to noise on Human3.6M [3]. Experimental settings follow Fig. 2. The size of markers indicates the computational cost of models.

MHFormer [4] introduces multiple hypotheses into its architecture to model depth ambiguity of body parts and uncertainty of joint detectors and is thus relatively robust (experimental results are available in the main paper). Besides, MHFormer also includes spatial-temporal transformer modules as in PoseFormerV1 [15]. To further verify the universality of our approach, we similarly improve MHFormer following MixSTE+. Experimental evidence shows that the proposed method promotes the robustness of MHFormer while reducing its computational cost (see Fig. 3), even though it already equips itself with prior knowledge of noisy joint detection. Therefore, this result demonstrates that, in terms of improvements in the robustness of models, our method is compatible with other approaches.

We have so far generalized our approach to other two transformer-based methods, *i.e.*, MixSTE [13] and MHFormer [4]. We may explore the generalization of the proposed method to a wider range of model architectures in the future, such as CNN-based and GNN-based methods in 3D human pose estimation. Moreover, we believe our method can also be utilized in other tasks, especially skeleton-based ones where the computational cost of long-sequence processing and the quality of human skeleton representations can become problems.

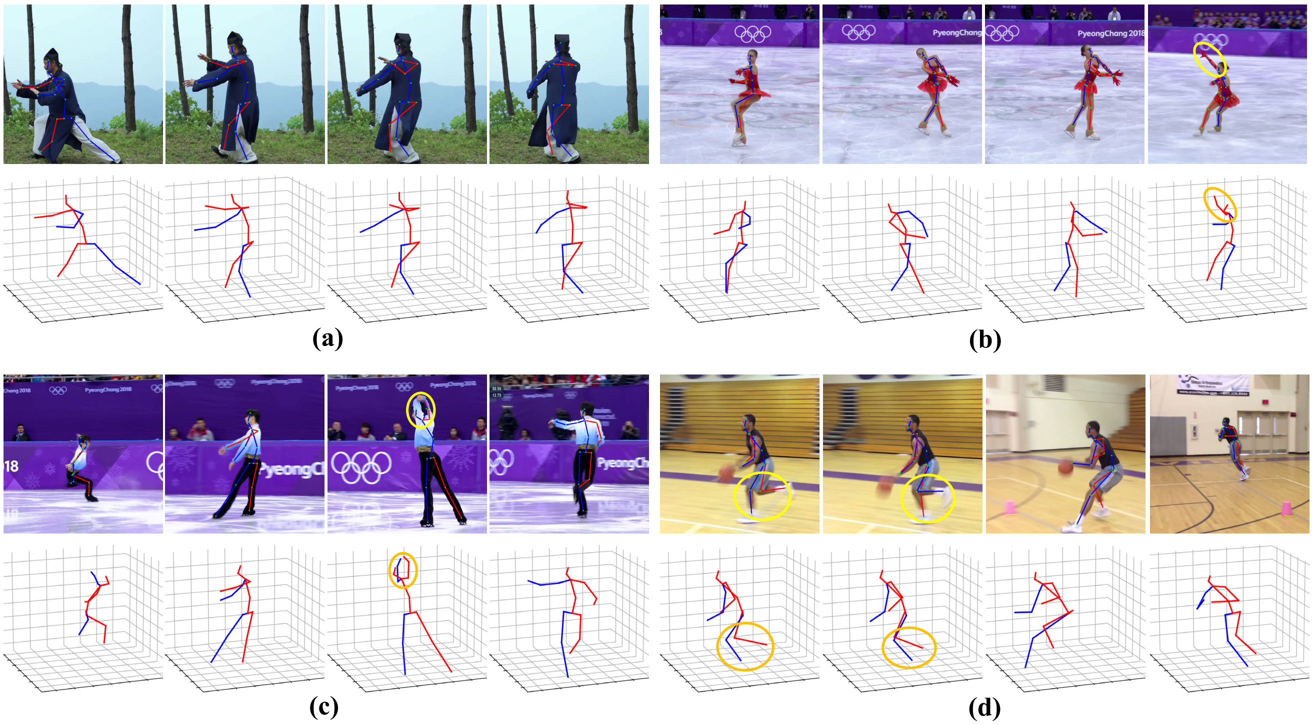

Figure 4. Qualitative results of PoseFormerV2 under challenging in-the-wild images: (a) Occlusions; (b)(c) Missed 2D joint detection; (d) Switched 2D joints. We highlight the unreliable 2D detection with light-yellow circles and corresponding 3D pose estimations with orange circles. PoseFormerV2 shows great robustness to imperfect 2D joint detection.

G. Visualizations and Analysis

In this section, we provide a series of qualitative results on challenging in-the-wild images to showcase the robustness of PoseFormerV2 in real scenarios.

Fig. 4 presents several representative hard cases with HRNet [11] 2D joint detection: (a) Occlusions where joints overlap with each other; (b)(c) Missed joints; (d) Switched joints. Specifically, the right arm of the person in the 4th image of (b) and the left arm of the person in the 3rd image of (c) are missed. Moreover, in the 2nd image of (d), two legs of the person are switched (highlighted with light-yellow circles). Despite the imperfect 2D joint input, PoseFormerV2 still infers correct positions for these joints in 3D space (marked with orange circles).

Analysis. The robustness of PoseFormerV2 is attributed to the usage of an appropriate representation – low-frequency DCT coefficients – of the input joint sequence, instead of hand-crafted modules that may bring additional computational cost such as the multi-hypothesis generation module in [4]. Low-frequency DCT coefficients provide a global vision of the input sequence and therefore the noise contained in individual video frames is dwarfed. This utilization of DCT coefficients also brings an extra advantage to PoseFormerV2, the temporal consistency of the estimated 3D pose between adjacent frames. We provide a video demo to illustrate that the proposed method keeps an excellent consistency (*i.e.*, temporal stability) under extremely corrupted 2D joint detection.

H. Broader Impacts and Limitations

Broader impacts. In this paper, we attempt to reconcile two critical issues in real-scenario applications of 3D HPE, *i.e.*, the efficiency of models to process long sequences for improved precision and their robustness against noisy 2D detection as high-quality joint sequences are hard to obtain. To encourage more real-world applications, we may shift our research focus from marginal improvements on carefully controlled datasets to overcoming the drawbacks of existing approaches in practical use. We expect more research to follow this line.

On the other hand, this work is done based on a scarcely investigated frequency method, *i.e.*, Discrete Cosine Transform (DCT) which plays an important role in conventional image compression algorithms. We hope this research will inspire more research to revisit traditional signal processing techniques as various data we treat in the deep learning era is actually signals of different forms. An appropriate combination of these choreographed techniques and recently developed deep learning approaches may bring surprising advantages.

Limitations. Our method includes two important hyperparameters – the number of sampled central frames and that of the kept DCT coefficients of the complete input sequence. Currently, they are chosen on the basis of experimental results or human experience for the trade-off between speed and accuracy. In the future, we may reshape them as learnable parameters that can be automatically learned from input data, or we may further theoretically formulate the optimal choices for them, thus removing the need for parameter-searching.

References

[1] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 1, 2

[2] 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. 2

[3] 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 TPAMI*, 2014. 2, 3

[4] 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 (CVPR)*, pages 13147–13156, June 2022. 2, 3, 4

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

[6] 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 *CVPR*, 2020. 1, 2

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[8] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 2

[9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 2

[10] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019. 1, 2

[11] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 4

[12] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*, pages 764–780. Springer, 2020. 2

[13] 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. 2, 3

[14] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey, 2020. 1

[15] 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 (ICCV)*, pages 11656–11665, October 2021. 1, 2, 3