Spatio-Temporal Transformer with Rotary Position Embedding and Bone Priors for 3D Human Pose Estimation

Cheng Chen[1,3], Jiang Liu[2], LiaoYuan Zeng[1,3], Fang Duan[4], Sean McGrath[5], Tian Dan[*1]

[1]Yibin Institute of UESTC , [2]Southwest Jiaotong University , [3]University of Electronic Science and Technology of China , [4]University of Bath , [5]University of Limerick

[1,3]202122011822@std.uestc.edu.cn, [2]liujiang@swjtu.edu.cn, [1,3]lyzeng@uestc.edu.cn, [4]fd506@bath.ac.uk, [5]sean.mcgrath@ul.ie, [1]tiandan@uestc.edu.cn

1 Introduction

- In 3D human pose estimation, the effective use of temporal and spatial information is key. Transformers have shown considerable potential in this field. However, existing models often utilize basic temporal position embedding, which restricts their ability to fully leverage temporal information.
- Additionally, while human body information like bone lengths are known in some cases, current networks do not incorporate this prior information, leading to limitations in estimation accuracy.
- To address these issues, we propose a transformer-based network for 3D human pose estimation that uses cross-attention with Rotary Position Embedding (RoPE). This network integrates RoPE with windows mechanism, allowing for flexible inference across varying sequence lengths while maintaining strong relative position awareness.
- Furthermore, we introduce bone length prior input to the network, and a cross-attention to integrate bone constraints into 3D pose estimation. Experimentally, our approach demonstrates that the inclusion of bone length information and longer sequences significantly reduces estimation errors, while improving the continuity of pose sequences.
- Notably, the performance surpasses state-of-the-art methods, showcasing the benefits of incorporating bone priors and advanced position embedding into 3D human pose estimation.

2 Approach- Framework Part: S-T Transformer With RoPE and Bone Priors

The Input of the Network

A continuous 2D pose sequence $C \in \mathbb{R}^{T\times N\times 7}$ with bone length priors.

N is the number of joints.
T represents the number of frames.
"**5**" means the dim of the features is five.

Bone Length Prior

The input is composed of two parts:
(1) Original 2D pose sequence $U \in \mathbb{R}^{T\times N\times 2}$.
(2) Second part is constructed from a hypergraph bone $B \in \mathbb{R}^{T\times N\times 5}$, with the human bone directed graph G is shown in Figure 1.

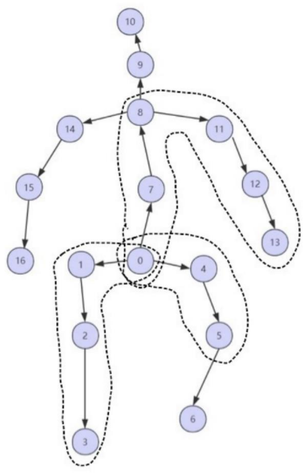

Figure 1: Bone Length Prior

Spatial, Temporal and Bone Transformer Block

The architecture of the network adopts a Spatio-Temporal-Bone(STB) Transformer structure, utilized the bone length information from the bone hypergraph to constrain the bone lengths between different joints.

- **Spatial Transformer Block(ST Block)**
  Purpose: To capture the information necessary to reconstruct the 3D human pose from the features $F_S \in \mathbb{R}^{N\times d_m}$ between different joints within the same frame.
  d_m is the dim of the features.
  N is the number of joints.
- **Temporal Transformer Block(TT Block)**
  Purpose: To capture the information necessary to reconstruct the 3D human pose from the features $F_T \in \mathbb{R}^{T\times d_m}$ between different frames within the same joints.
  where d_m is the dim of the features, T is the number of frames.
- **Bone Transformer Block(BT Block)**
  Purpose: To utilize cross-attention to perceive the prior information of bone length and guide the bone vectors composed of different joints to conform to the prior bone length.
- **Position Embedding**
  We employ the Rotational Position Encoding (RoPE) with windows mechanism, which facilitates the input of variable-length sequences and distinguishes the positional semantics between adjacent and distant frames.

Loss Function

Defined of Loss Function $\mathcal{L}$: $\mathcal{L} = \mathcal{L}_{MPJPE} + \lambda_t \mathcal{L}_{MPJVE} + \lambda_m \mathcal{L}_{MPBLE}$

$\mathcal{L}_{MPJPE}$ represents the Mean Per Joint Position Error (MPJPE) [4] loss.
$\mathcal{L}_{MPJVE}$ denotes the Mean Per Joint Velocity Error (MPJVE) [5] loss.
$\mathcal{L}_{MPBLE}$ indicates the Mean Per Bone Length Error (MPBLE).
λ_t and λ_m are the Lagrange Multiplier, used to adjusted loss ratio.

3 Experiment and Result

Datasets

Human3.6M [6] is the most widely utilized indoor dataset for 3D human pose estimation tasks. Our approach is the same as before [1, 2, 4, 5].

Implementation Details

- **To analyze the performance of our model**: Using the input 2D pose from a 2D pose detector or 2D ground truth.
- **The bone length input**: The bone length ground truth as input in both the 2D pose detector and 2D ground truth.

Comparison with State-of-the-art Methods

"Figure 3" shows the MPJVE result of our method.

- **Compared with other baselines:**
  (1) Our method achieves the best results under most of the actions, outperforming the MixSTE by 6.6% in terms of total error, and by 2.5% compared to MotionBERT.
  (2) Our method achieves results quite close to the best model at T = 243 and takes the best performance at T = 1000.
- **On most of the motions**, our model achieves the best performance, which demonstrates:
  (1) Our method is significantly effective in estimating the correct 3D human poses.
  (2) Our method has strong generalization capabilities under most of the motions.
- **Under longer sequence**, our model still achieves performance gains over longer sequences.
- **Compared with MixSTE:**
  (1) our method achieves results quite close to the best model at T = 243 and takes the best performance at T = 1000.
  (2) our results are 15.2% better, and longer sequences also give better results, with our model outperforming the best by 7.3% at T=1000.
- **Preliminary summary**
  RoPE can better provide temporal position information and help to maintain the temporal continuity of the joint motion.

Ablation Study

- As shown in **Figure 2,**
  Relative error(%) of bone lengths for estimated human pose with our model are mostly better than those of MixSTE.
- As shown in **Table 1**, we have tested different Component.
  (1) Bone Transformer Block plays an important role in reducing the estimation error of 3D human pose.
  (2) Compared to the baseline with RoPE, the error is reduced by 15%.
  (3) Our Loss also plays a crucial role and the model can't converge to the correct weights without it.
- As shown in **Table 2**, We conducted comparative experiments with W-RoPE on MixSTE. W-RoPE applied on temporal information can provide better positional information to the model and reduce MPJPE of the 3D HPE.
- As shown in **Table 3,**
  Different channels and input lengths lead to different performances(from 48.2mm to 36.9mm).

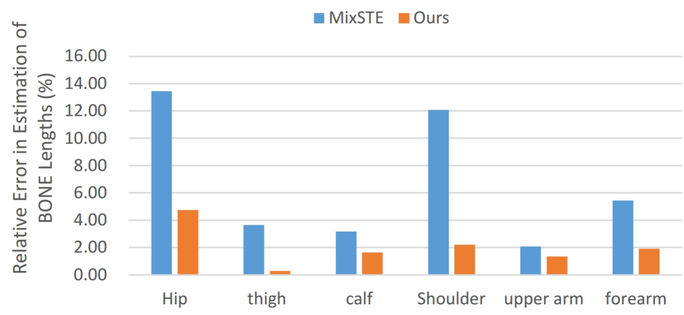

Figure 2: The relative error(%) of bone lengths for estimated human pose between our method and MixSTE on Directions S9 Human3.6M using CPN 2D pose as inputs.

Table 1: Ablation study for each component used in our method.

S-T Transformer	RoPE	Bone Transformer	Our Loss	MPJPE (mm)
√				40.9
√	√			40.7
√	√	√		40.4
√	√		√	41.1
√	√	√	√	38.2

Table 2: Ablation study for RoPE in Spatio-Temporal Transformer(MixSTE) [1].

Input Length (T) , dimensionless	128	243	300
S-T Transformer, dimensionless	42.0	40.9	41.8
S-T Transformer with W-RoPE, dimensionless	41.4	10.7	40.6

Table 3: Ablation study for hyper-parameter.

Depth (d_t) (dimensionless)	Dimension (d_m) (dimensionless)	Input Length (T) (dimensionless)	MPJPE (mm)
8	128	27	48.2
8	256	27	46.8
8	512	27	44.5
8	512	243	38.2
8	512	700	37.3
8	512	1000	36.9
8	512	1500	36.9

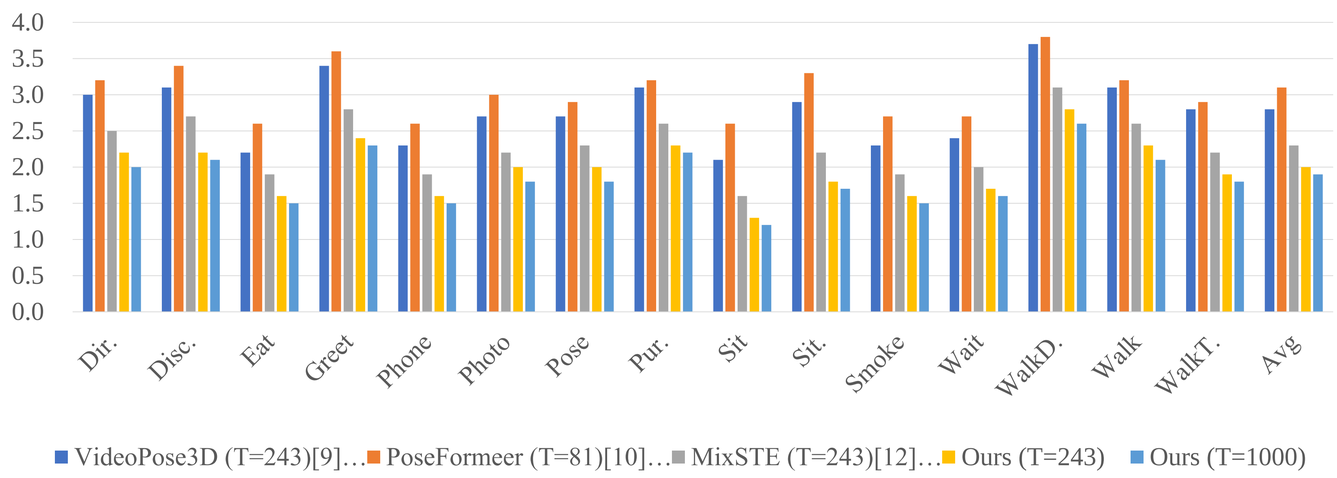

Figure 3: Detailed quantitative comparison results of MPJVE in millimeters (mm) on Human3.6M under Protocol #1 using CPN 2D keypoints as input.

4 Conclusion

- We propose a Spatio-Temporal Transformer with rotary position embedding and bone priors for 3D HPE.
- We have a comprehensive and systematic study of bone priors and RoPE in estimating reasonable 3D human pose.
- With bone priors and better motion continuity of joint, our method achieves accurate 2D-to-3D human pose estimation outperforming the start-of-the-art.

Reference List

[1] 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] 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*, 32(1):198–209, 2021.

[3] Hanyuan Chen, Jun-Yan He, Wangmeng Xiang, Zhi-Qi Cheng, Wei Liu, Hanbing Liu, Bin Luo, Yifeng Geng, and Xuansong Xie. Hdformer: High-order directed transformer for 3d human pose estimation. *arXiv preprint arXiv:2302.01825*, 2023.

[4] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017.

[5] 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.

[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.

BMVC 2024