TCPFormer: Learning Temporal Correlation with Implicit Pose Proxy for 3D Human Pose Estimation

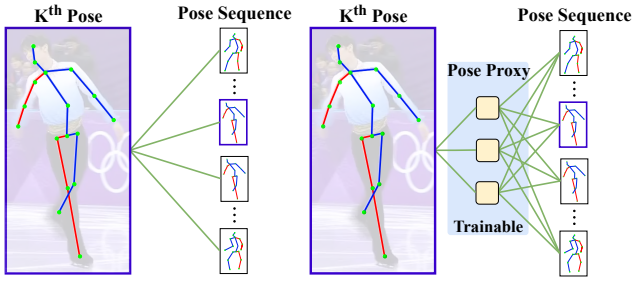
Jiajie Liu[1], **Mengyuan Liu**[1*], **Hong Liu**[1], **Wenhao Li**[2]

[1]State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School
[2]Nanyang Technological University

Abstract

Recent multi-frame lifting methods have dominated the 3D human pose estimation. However, previous methods ignore the intricate dependence within the 2D pose sequence and learn single temporal correlation. To alleviate this limitation, we propose TCPFormer, which leverages an implicit pose proxy as an intermediate representation. Each proxy within the implicit pose proxy can build one temporal correlation therefore helping us learn more comprehensive temporal correlation of human motion. Specifically, our method consists of three key components: Proxy Update Module (PUM), Proxy Invocation Module (PIM), and Proxy Attention Module (PAM). PUM first uses pose features to update the implicit pose proxy, enabling it to store representative information from the pose sequence. PIM then invocates and integrates the pose proxy with the pose sequence to enhance the motion semantics of each pose. Finally, PAM leverages the above mapping between the pose sequence and pose proxy to enhance the temporal correlation of the whole pose sequence. Experiments on the Human3.6M and MPI-INF-3DHP datasets demonstrate that our proposed TCPFormer outperforms the previous state-of-the-art methods.

Code — https://github.com/AsukaCamellia/TCPFormer



(a) Previous Methods (b) Our TCPFormer

Figure 1: **An illustration of our motivation.** Given a pose sequence of length T, we take the individual pose within the pose sequence as an example. (a) In previous methods, one pose establishes the temporal correlation with the pose sequence only in one *1-to-T* mapping. (b) We introduce an implicit pose proxy to act as an intermediate representation. Each proxy within the implicit pose proxy of length L can establish one *1-to-T* mapping, which facilitates learning more comprehensive temporal correlation.

Introduction

3D human pose estimation has always been a crucial problem in computer vision, which aims to locate the 3D joint positions of a human body (Moon and Lee 2020; Pavlakos, Zhou, and Daniilidis 2018; Chen et al. 2021). Nowadays, 3D human pose estimation finds widespread applications in various scenarios, including motion prediction (Wang et al. 2023), action recognition (Zhang et al. 2022a), and human-robot interaction (Gong et al. 2022; Ye et al. 2021). Given the widespread usage of 2D human pose detectors (Chen et al. 2018; He et al. 2017; Newell, Yang, and Deng 2016; Sun et al. 2019) and the task-relatedness between 2D pose and 3D pose, most research follows a 2D-to-3D lifting pipeline (Zheng et al. 2021; Li et al. 2022a,b; Zhang et al. 2022b; Wang et al. 2024), where 2D keypoints are first detected and then lifted to the 3D space. Despite the considerable success achieved, this task remains an ill-posed problem and inherently suffers from depth ambiguity. An extensive body of literature focuses on exploiting temporal information between adjacent frames to mitigate this issue, ranging from earlier methods (Pavllo et al. 2019; Liu et al. 2020b; Chen et al. 2021) use temporal convolution and subsequent attempts (Cai et al. 2019; Hu et al. 2021; Wang et al. 2020) use graph convolution. Recently, transformer (Vaswani et al. 2017) have achieved significant success in both natural language preprocess (Brown et al. 2020; Devlin et al. 2018) and computer vision (Dosovitskiy et al. 2021; Carion et al. 2020). For the 3D human pose estimation task, many works (Li et al. 2022b; Zhang et al. 2022b; Shan et al. 2022; Tang et al. 2023) leverage the powerful sequence modeling capability of transformer to extend their input from the limited neighboring frames to long-term sequences for advanced accuracy, e.g., 243 video frames for MixSTE (Zhang et al. 2022b) and STCFormer (Tang et al. 2023); large as 351 frames for MHFormer (Li et al. 2022b).

Despite their achievements, a potential concern has gradually emerged: with the massive increase in the number of input frames, the performance improvement becomes slow. For instance, PoseFormerV2 (Zhao et al. 2023b) achieved an error reduction of 0.8mm when expanding the input from 81

*Corresponding author: liumengyuan@pku.edu.cn

frames to 243 frames. StridedTrans (Li et al. 2022a) achieved a marginal 0.3mm error reduction when expanding the input from 243 frames to 351 frames, while MHFormer (Li et al. 2022b) achieved an even smaller error reduction of 0.2mm with the same input expansion. These key observations point towards a problem that restricts most methods from effectively modeling the temporal correlation within the 2D pose sequence. In this work, we are trying to solve this problem.

As illustrated in Figure 1a, we discover that most of the aforementioned multi-frame methods only establish one *1-to-T* mapping for each pose within the pose sequence, where T denotes the length of pose sequence. However, due to the extensive number of frames, only establishing one *1-to-T* mapping can not comprehensively reflect the complex temporal correspondence within the pose sequence.

To address this limitation, we propose a novel method to learn **T**emporal **C**orrelation with Implicit Pose **P**roxy, dubbed **TCPFormer**. As illustrated in Figure 1b, we introduce an implicit pose proxy to act as the intermediate representation. We first establish a *1-to-L* mapping to build the relationship between the individual pose and the implicit pose proxy, where L denotes the length of the implicit pose proxy. Then, each proxy within the implicit pose proxy will interact with the pose sequence and build multiple *1-to-T* mapping to help model learning more comprehensive temporal correlation. Moreover, our implicit pose proxy is trainable and will be continuously optimized during the training process.

Specifically, we first propose a Proxy Update Module (PUM). PUM adaptively encodes useful and representative information from the pose sequence to update the pose proxy through the cross-attention mechanism (Vaswani et al. 2017). Although the information in the pose proxy has been updated, we have not yet transmitted it to each pose within the pose sequence. Therefore, we propose a Proxy Invocation Module (PIM) that uses the pose proxy as the key and value to enhance the feature representation ability of the pose sequence. In addition, we propose a Proxy Attention Module (PAM). PAM skillfully leverages the two cross-attention matrices of PUM and PIM to get an aggregation matrix and flexibly fuses it with the original self-attention matrix to obtain a more effective and comprehensive temporal correlation.

We extensively evaluate our TCPFormer on two widely used benchmark datasets, Human3.6M (Ionescu et al. 2013) and MPI-INF-3DHP (Mehta et al. 2017). Empirical evaluations show that our approach outperforms the previous state-of-the-art methods. Comprehensive ablation studies are also presented to evaluate the contribution of each component. Our **contributions** can be summarized as follows:

- To the best of our knowledge, we are the first to introduce the implicit pose proxy to 3D human pose estimation. Our method leverages the implicit pose proxy as an intermediate representation to effectively model the complex temporal correlation within the pose sequence.
- We design three novel modules: Proxy Update Module, Proxy Invocation Module, and Proxy Attention Module. These three modules present a unique way to effectively enhance the feature of pose sequence and learn a more comprehensive temporal correlation.
- Extensive experiments conducted on Human3.6M and MPI-INF-3DHP two challenging datasets for 3D human pose estimation demonstrate that our method achieves superior performances than the previous methods.

Related Work

3D Human Pose Estimation

Early works (Ionescu, Carreira, and Sminchisescu 2014; Ionescu et al. 2013; Ramakrishna, Kanade, and Sheikh 2012; Agarwal and Triggs 2005; Andriluka, Roth, and Schiele 2009; Ionescu, Li, and Sminchisescu 2011) of monocular 3D human pose estimation primarily focus on exploiting spatial prior information in the form of human skeletal structure and motion features. With the development of deep learning, more deep neural network-based methods have been introduced and can be divided into two mainstream types: one-stage manner and two-stage manner. One-stages approaches (Kanazawa et al. 2018; Pavlakos et al. 2017; Sun et al. 2018) directly estimate the 3D pose from the input image without the intermediate 2D pose representation. Different from the one-stage manner, two-stage methods (Fang et al. 2018; Martinez et al. 2017; Zhao et al. 2019; Liu et al. 2020a; Xu and Takano 2021) first obtain 2D joint coordinates in the image and then leverage the task-relevant positional information to lift the 2D joint coordinates to 3D poses. With the reliable achievement of 2D human pose detectors (Chen et al. 2018; He et al. 2017; Newell, Yang, and Deng 2016; Sun et al. 2019), these 2D-to-3D lifting methods outperform one-stage approaches. However, they still inherently suffer from the problem of depth ambiguities. To address this problem, some studies (Liu et al. 2020b; Pavllo et al. 2019) have made preliminary explorations in utilizing temporal information. Liu *et al.* (Liu et al. 2020b) extend the temporal convolutional network by introducing the attention mechanism. The aforementioned methods utilize limited temporal information, which is unable to effectively facilitate 3D human pose estimation.

Transformer-based Methods

For the 3D human pose estimation task, PoseFormer (Zheng et al. 2021) firstly introduces transformer architecture to leverage spatial and temporal dependency. MHFormer (Li et al. 2022b) addresses the depth ambiguity by learning spatio-temporal representations of multiple pose hypotheses. MixSTE (Zhang et al. 2022b) constructs a mixed spatio-temporal transformer to capture the temporal motion of different body joints. P-STMO (Shan et al. 2022) is the first approach that introduces the pre-training technique to 3D human pose estimation. PoseFormerV2 (Zhao et al. 2023b) improves PoseFormer by utilizing a frequency-domain representation of input joint sequences. STCFormer (Tang et al. 2023) decomposes spatio-temporal attention and integrates the structure-enhanced positional embedding. MotionBERT (Zhu et al. 2023) and UPS (Foo et al. 2023) both train a unified model for multi-task. However, these methods still have limitation in directly modeling the complex temporal correlation of the pose sequence due to sequence length. We conduct further exploration to address this limitation in this paper.

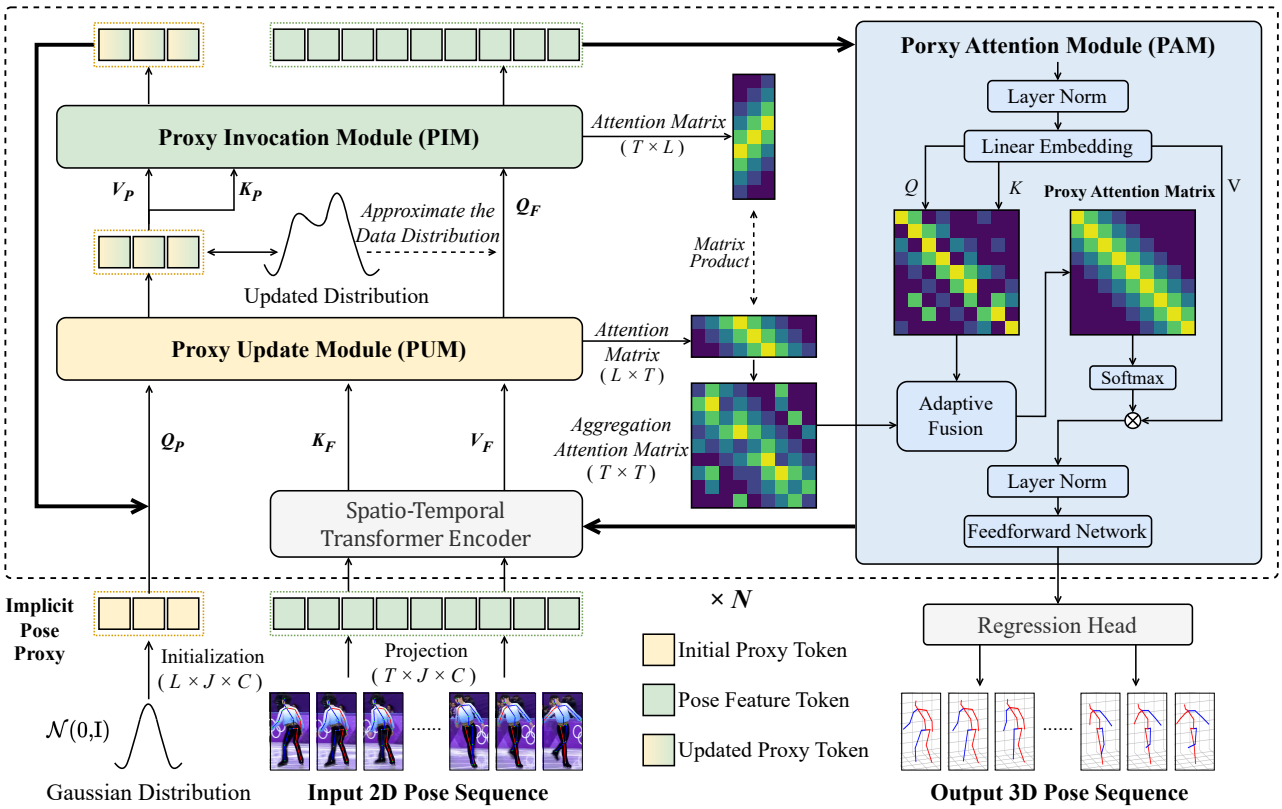


Figure 2: **Overview of our method.** We first extract the spatio-temporal information through a spatio-temporal encoder. Then, we introduce an implicit pose proxy which is initialized by Gaussian distribution. These features and proxy are then handed to the proxy update module to update the implicit pose proxy. Next, the proxy invocation module uses the updated pose proxy to enhance the feature of the pose sequence. We obtain an aggregation attention matrix through two cross attention matrices and send it with the pose sequence feature to the proxy attention module to learn comprehensive temporal correlation. After repeating the above processes N times, we use a regression head to obtain the 3D pose sequence.

Method

Given a 2D pose sequence $X \in \mathbb{R}^{T\times J\times C_{\text{in}}}$, our goal is to estimate the 3D pose sequence $Y \in \mathbb{R}^{T\times J\times C_{\text{out}}}$. Here, T refers to the number of input frames, and J refers to the number of joints. C_{in} and C_{out} denote the dimension of input 2D pose and output 3D pose.

Overview

An overview of our pipeline is illustrated in Fig. 2. Firstly, we use a base spatio-temporal encoder to extract information from X and obtain a basic feature $B \in \mathbb{R}^{T\times J\times C_f}$, where C_f denotes the dimension of the hidden feature. We reshape B into $F \in \mathbb{R}^{J\times T\times C_f}$ as the per-joint temporal features. F is then fed into our Proxy Update Module with the implicit pose proxy $P \in \mathbb{R}^{J\times L\times C_f}$ which is initialized by Gaussian distribution. Here, L indicates the time dimension of implicit pose proxy (further discussion in the ablation study). This step will update implicit pose proxy P. Then, the updated implicit pose proxy P and feature F will be sent to our Proxy Invocation Module to generate the enhanced features $\widetilde{F} \in \mathbb{R}^{J\times T\times C_f}$. During above process, we can obtain the cross-attention matrix $M_{P\to F} \in \mathbb{R}^{L\times T}$ and $M_{F\to P} \in \mathbb{R}^{T\times L}$. We aggregate them to generate a aggregation attention matrix $M_{F\to F} \in \mathbb{R}^{T\times T}$. The enhanced feature $\widetilde{F}$ and the aggregation attention matrix $M_{F\to F} \in \mathbb{R}^{T\times T}$ are sent to Proxy Attention Module to produce the final feature $\overline{F} \in \mathbb{R}^{J\times T\times C_f}$. In the end, we use a regression head to generate the final 3D pose sequence Y.

Proxy Update Module

For the proxy update module, we first use three linear matrices W_Q^U,W_K^U and W_V^U to map proxy P to queries Q_P, the input feature F to keys K_F and values V_F as follows:

$$Q_P = P \cdot W_Q^U \quad K_F = F \cdot W_K^U \quad V_F = F \cdot W_V^U \tag{1}$$

Then we use the Softmax attenion (Vaswani et al. 2017), which has been widely used in modern transformer designs to get the cross attention matrix $M_{P\to F}$ as follows:

$$M_{P\to F} = Softmax(Q_P \cdot K_F^T / \sqrt{C_f}) \tag{2}$$

Finally, we fuse the $M_{P\to F}$ and V_F to get the updated implicit pose proxy P as follows:

$$P = P + M_{P\to F} \cdot V_F \tag{3}$$

Proxy Invocation Module

After updating the implicit pose proxy, We use inverse processes to integrate the updated implicit pose proxy P into the feature F. Similarly, we use three linear matrices W_Q^I,W_K^I and W_V^I to project the feature F to queries Q_F, the updated implicit pose proxy P to keys K_P, and values V_P as:

$$Q_F = F \cdot W_Q^I \quad K_P = P \cdot W_K^I \quad V_P = P \cdot W_V^I \tag{4}$$

Then, we use the Softmax function to get the cross attention matrix $M_{F\to P}$ as follows:

$$M_{F\to P} = Softmax(Q_F \cdot K_P^T / \sqrt{C_f}) \tag{5}$$

Finally, we utlize the $M_{F\to P}$ and V_P to get the enhanced pose sequence feature $\widetilde{F}$ as follows:

$$\widetilde{F} = F + M_{F\to P} \cdot V_P \tag{6}$$

Proxy Attention Module

After the above process, we can obtain two cross-attention matrices $M_{F\to P}$ and $M_{P\to F}$. We skillfully leverage them to further learn temporal correlation. Concretely, we use matrix multiplication to obtain an aggregation attention matrix as:

$$M = M_{F\to P} \cdot M_{P\to F} \tag{7}$$

Then, we map the enhanced feature $\widetilde{F}$ to queries $Q_{\widetilde{F}}$, keys $K_{\widetilde{F}}$, and values $V_{\widetilde{F}}$ as follows:

$$Q_{\widetilde{F}} = \widetilde{F} \cdot W_Q^A \quad K_{\widetilde{F}} = \widetilde{F} \cdot W_K^A \quad V_{\widetilde{F}} = \widetilde{F} \cdot W_V^A \tag{8}$$

The aggregation attention matrix is then adaptively fused with the original self attention matrix as follows:

$$\overline{M} = Sigmoid(\mu) \cdot M + (1 - Sigmoid(\mu)) \cdot Q_{\widetilde{F}} \cdot K_{\widetilde{F}}^T \tag{9}$$

where μ is a learnable parameter for each layer during the training process to adaptively learn suitable fusion ratios. Finally, we get the final pose sequence feature $\overline{F}$ as follows:

$$\overline{F} = \widetilde{F} + Softmax(\overline{M} / \sqrt{C_f}) \cdot V_{\widetilde{F}} \tag{10}$$

Regression Head and Loss Function

After the above process is repeated several times, we project feature $\overline{F}$ to a higher dimension by applying a linear layer and tanh activation to compute the motion semantics and use a linear transformation layer as regression head to estimate the final 3D pose sequence Y.

Our network has two optimization objectives and is trained in an end-to-end manner. We use L2 loss to minimize the errors between predictions and ground truths:

$$\mathcal{L}_{3D} = \frac{1}{JT} \sum_{j=1}^{J} \sum_{t=1}^{T} \left\| \widehat{Y}_{j,t} - Y_{j,t} \right\|_2 \tag{11}$$

where $\widehat{Y}_{j,t}$ and $Y_{j,t}$ are the ground truth and estimated 3D pose of the j-th joint in t-th frame. In addition, the temporal consistency loss (TCLoss) in (Hossain and Little 2018) is introduced to produce smooth poses. Specifically, the TCLoss can be formulated as follows:

$$\mathcal{L}_T = \frac{1}{J(T-1)} \sum_{j=1}^{J} \sum_{t=2}^{T} \left\| \Delta\widehat{Y}_{j,t} - \Delta Y_{j,t} \right\|_2 \tag{12}$$

where $\Delta\widehat{Y}_t = \widehat{Y}_t - \widehat{Y}_{t-1}$, $\Delta Y_t = Y_t - Y_{t-1}$. The final loss function $\mathcal{L}$ is then defined as :

$$\mathcal{L} = \mathcal{L}_{3D} + \lambda \mathcal{L}_T \tag{13}$$

where λ is used to balance position accuracy and motion smoothness.

Experiments

Datasets and Evaluation Metrics

We comprehensively evaluate our model on two large-scale 3D human pose estimation datasets: Human3.6M (Ionescu et al. 2013) and MPI-INF-3DHP (Mehta et al. 2017).

Human3.6M is the most popular benchmark for indoor 3D human pose estimation, which contains approximately 3.6 million frames captured by 4 cameras at different views. This dataset contains 11 subjects performing 15 typical actions.

MPI-INF-3DHP is a recently proposed large-scale challenging dataset with both indoor and outdoor scenes. The training set comprises 8 subjects, covering 8 activities, ranging from walking and sitting to complex exercise poses and dynamic actions. The test set covers 7 activities, containing three scenes: green screen, non-green screen, and outdoor environments.

Evaluation Metrics. For the Human3.6M dataset, we use two evaluation metrics: MPJPE and P-MPJPE. MPJPE (Mean Per Joint Position Error) is computed as the mean Euclidean distance between the estimated joints and the ground truth in millimeters after aligning their root joints. P-MPJPE (Procrustes-MPJPE) is the MPJPE after the estimated joints align to the ground truth via a rigid transformation. For the MPI-INF-3DHP dataset, following previous works (Shan et al. 2022; Zhou, Yin, and Li 2024; Zhu et al. 2023), we use ground truth 2D pose as input and report MPJPE, Percentage of Correct Keypoint (PCK) with the threshold of 150mm, and Area Under Curve (AUC) as the evaluation metrics.

Implementation Details

We consider the layers N of modules, the number H of heads in attention block, the size C of hidden feature, the temporal dimension L of implicit pose proxy, and the initialization distribution D of proxy as free parameters. The performances of the versions with ($N = 16, H = 8, C = 128, L = T/3, D = Gaussian$) are reported. Our model is implemented using PyTorch and executed on a server equipped with 2 NVIDIA 4090 GPUs. We apply horizontal flipping augmentation for both training and testing following (Tang et al. 2023; Zhu et al. 2023; Foo et al. 2023; Zhao et al. 2023a). For model training, we set each mini-batch as 16 sequences. The network parameters are optimized using AdamW (Loshchilov and Hutter 2017) optimizer over 90 epochs with a weight decay of 0.01. The initial learning rate is set to 5e-4 with an exponential learning rate decay schedule and the decay factor is 0.99. In the experiments on Human3.6M, two kinds of input are utilized including the 2D ground truth and the Stacked Hourglass (Newell, Yang, and Deng 2016) 2D pose detection, following (Zhu et al. 2023; Ci et al. 2019). For MPI-INF-3DHP, 2D ground truth is used following previous works (Cai et al. 2024; Zhang et al. 2022b; Li et al. 2023; Zhu et al. 2023; Li et al. 2024).

Method	Venue	Seq2Seq	T	Parameter	MACs	MACs/frames	MPJPE ↓	P-MPJPE ↓
MHFormer (Li et al. 2022b)	CVPR'22	✗	351	30.9M	7.1G	7096M	43.0	34.4
MixSTE (Zhang et al. 2022b)	CVPR'22	✓	243	33.6M	139.0G	572M	40.9	32.6
P-STMO (Shan et al. 2022)	ECCV'22	✗	243	6.2M	0.7G	740M	42.8	34.4
STCFormer (Tang et al. 2023)	CVPR'23	✓	243	4.7M	19.6G	80M	41.0	32.0
PoseFormerV2 (Zhao et al. 2023b)	CVPR'23	✗	243	14.3M	0.5G	528M	45.2	35.6
GLA-GCN (Yu et al. 2023)	ICCV'23	✗	243	1.3M	1.5G	1556M	44.4	34.8
MotionBERT (Zhu et al. 2023)	ICCV'23	✓	243	42.3M	174.8G	719M	<u>39.2</u>	32.9
KTPFormer (Peng, Zhou, and Mok 2024)	CVPR'24	✓	243	33.7M	69.5G	286M	40.1	<u>31.9</u>
TCPFormer (Ours)	-	✓	81	35.0M	36.4G	150M	40.5	33.7
TCPFormer (Ours)	-	✓	243	35.1M	109.2G	449M	**37.9**	**31.7**

Table 1: Quantitative comparisons on Human3.6M dataset. T is the number of input frames. Seq2seq refers to estimating 3D pose sequences rather than only the center frame. MACs/frames represents multiply-accumulate operations for each output frame. The best result is shown in bold, and the second-best result is underlined. Our TCPFormer achieves the best performance with smaller parameters and computational cost compared with MotionBERT (Zhu et al. 2023).

MPJPE (GT)	T	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
MHFormer (Li et al. 2022b)	351	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2	37.0	39.3	30.0	31.0	29.4	22.2	23.0	30.5
MixSTE (Zhang et al. 2022b)	243	21.6	22.0	20.4	21.0	20.8	24.3	24.7	21.9	26.9	24.9	21.2	21.5	20.8	14.7	15.7	21.6
P-STMO (Shan et al. 2022)	243	28.5	30.1	28.6	27.9	29.8	33.2	31.3	27.8	36.0	37.4	29.7	29.5	28.1	21.0	21.0	29.3
STCFormer (Tang et al. 2023)	243	20.8	21.8	20.0	20.6	23.4	25.0	23.6	19.3	27.8	26.1	21.6	20.6	19.5	14.3	15.1	21.3
PoseFormerV2 (Tang et al. 2023)	243	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLA-GCN (Yu et al. 2023)	243	20.1	21.2	20.0	19.6	21.5	26.7	23.3	19.8	27.0	29.4	20.8	20.1	19.2	12.8	13.8	21.0
MotionBERT (Zhu et al. 2023)	243	<u>16.7</u>	19.9	<u>17.1</u>	<u>16.5</u>	<u>17.4</u>	<u>18.8</u>	<u>19.3</u>	20.5	24.0	<u>22.1</u>	<u>18.6</u>	<u>16.8</u>	<u>16.7</u>	<u>10.8</u>	<u>11.5</u>	<u>17.8</u>
(Peng, Zhou, and Mok 2024)	243	19.6	<u>18.6</u>	18.5	18.1	18.7	22.1	20.8	<u>18.3</u>	<u>22.8</u>	22.4	18.8	18.1	18.4	13.9	15.2	19.0
TCPFormer (Ours)	81	19.2	19.6	20.2	18.3	20.7	24.0	20.9	20.3	26.9	26.9	21.6	18.5	19.6	15.7	15.4	20.5
TCPFormer (Ours)	243	**15.0**	**15.9**	**16.1**	**14.2**	**15.7**	**16.4**	**16.2**	**16.9**	**22.1**	**20.6**	**16.7**	**13.3**	**13.8**	**9.2**	**9.9**	**15.5**

Table 2: Results on Human3.6M dataset in millimeters under MPJPE using ground truth 2D pose as input. T is the number of input frames. The best result is shown in bold, and the second-best result is underlined. Our TCPFormer reduces 2.3mm (12.9%) MPJPE compared with MotionBERT (Zhu et al. 2023) and achieves the best performance in all actions.

Comparison with State-of-the-art Methods

Due to page limitations, we only present the best results of other methods along with their corresponding number of input frames (T). Our TCPFormer achieves the best performance across different numbers of input frames.

Results on Human3.6M. We compare our TCPFormer with previous methods on the Human3.6M dataset. Table 1 summarizes the performance comparisons in terms of MPJPE and P-MPJPE errors of all 15 actions and the computational cost of each method. Our method achieves state-of-the-art performance with an MPJPE of 37.9mm and P-MPJPE of 31.7mm with T = 243. It is worth noting that our method with T = 81 input frames still achieves competitive performance with an MPJPE error of 40.5mm and surpasses the performance of most methods with a higher number of input frames. For example, this result outperforms PoseformerV2 (Zhao et al. 2023b) (40.5mm v.s. 45.2mm) with 243 frames, and MHFormer (Li et al. 2022b) even with 351 frames (40.5mm v.s. 43.0mm). Our model also achieves a good balance between performance and computational cost. TCPFormer achieves the best performance with lower parameters (35.1M v.s. 42.5M) and MACs per frame (449M v.s. 719M) compared with MotionBERT (Zhu et al. 2023).

To explore the lower bound of our method, we also directly use the 2D ground truth as input. As shown in the Table 2, our method with T = 243 achieves the best performance with an MPJPE of 15.5mm in all actions. Similarly, our method with T = 81 input frames surpasses the performance of most methods with a higher number of input frames. For example, this result outperforms STCFormer (Tang et al. 2023) (20.5mm v.s. 21.3mm) with 243 frames, and MHFormer (Li et al. 2022b) with 351 frames (20.5mm v.s. 30.5mm).

Results on MPI-INF-3DHP. To demonstrate the generalization capability of our model, we evaluate our model on the challenging MPI-INF-3DHP dataset, which includes more complex scenes and motions. Following previous works (Zheng et al. 2021; Zhang et al. 2022b; Shan et al. 2022; Li et al. 2024), we use ground truth 2D as input and set the number of input frames as 9, 27, or 81 due to the shorter video sequences. As observed in Table 3, our method with T = 81 achieves the best performance with the PCK of 99.0%, AUC of 87.7%, and MPJPE of 15.0mm. More remarkably, our method with T = 9 input frames still outperforms the previous state-of-the-art model STCFormer (Tang et al. 2023) with T = 81 input frames, despite having only one-ninth of the input frames (9 frames v.s. 81 frames).

Method	Venue	T	Seq2Seq	PCK ↑	AUC ↑	MPJPE ↓
MHFormer (Li et al. 2022b)	CVPR'22	9	✗	93.8	63.3	58.0
MixSTE (Zhang et al. 2022b)	CVPR'22	27	✓	94.4	66.5	54.9
P-STMO (Shan et al. 2022)	ECCV'22	81	✗	97.9	75.8	32.2
STCFormer (Tang et al. 2023)	CVPR'23	81	✓	98.7	83.9	23.1
PoseFormerV2 (Zhao et al. 2023b)	CVPR'23	81	✗	97.9	78.8	27.8
GLA-GCN (Yu et al. 2023)	ICCV'23	81	✗	98.5	79.1	27.8
MotionBERT (Zhu et al. 2023)	ICCV'23	-	✓	-	-	-
KTPFormer (Peng, Zhou, and Mok 2024)	CVPR'24	81	✓	<u>98.9</u>	85.9	<u>16.7</u>
TCPFormer (Ours)	-	9	✓	98.3	84.4	20.4
TCPFormer (Ours)	-	27	✓	98.7	<u>86.5</u>	17.8
TCPFormer (Ours)	-	81	✓	**99.0**	**87.7**	**15.0**

Table 3: Results on MPI-INF-3DHP under three evaluation metrics. T is the number of input frames. Seq2seq refers to estimating 3D pose sequences rather than only the center frame. The best result is shown in bold, and the second-best result is underlined.

Ablation Study

All experiments were conducted on the Human3.6M dataset with T = 243 as the number of input frames.

Step	Proxy	PUM	PIM	PAM	MPJPE ↓	P-MPJPE ↓
1	✓	-	-	-	42.2	34.6
2	✓	✓	-	-	39.5	32.6
3	✓	✓	✓	-	38.7	32.3
Ours	✓	✓	✓	✓	**37.9**	**31.7**

Table 4: The effectiveness of different components. All our proposed novel components exhibit improvements.

Impact of Each Component. As shown in Table 4, we validate the overall performance gain brought by the proposed implicit pose proxy (Proxy), proxy update module (PUM), proxy invocation module (PIM), and proxy attention module (PAM). Our baseline, which only introduces an implicit pose proxy without additional module design, achieves a result of 42.2mm MPJPE and 34.6mm P-MPJPE. By applying PUM, our method decreases 2.7mm MPJPE and 2.0mm P-MPJPE. Next, we integrate PIM into our method and achieve better results with 38.7mm MPJPE and 32.3mm P-MPJPE. Finally, we achieve the best performance with 37.9mm MPJPE and 31.7mm P-MPJPE by incorporating the PAM.

Length	Distribution	MPJPE ↓	P-MPJPE ↓
27	Gaussian	38.4	32.1
81	Random	39.1	32.5
81	Laplacian	38.2	32.0
81	**Gaussian**	**37.9**	**31.7**
243	Gaussian	38.6	32.7

Table 5: Analysis on implicit pose proxy. Distribution and Length denote the temporal dimension and initial distribution of our proposed implicit pose proxy.

Analysis on Implicit Pose Proxy. How to represent implicit pose proxy is crucial for our methods. We investigated the impact of the temporal dimension and initial distribution of our implicit pose proxy. For the temporal dimension, we set it to 27, 81, and 243 respectively. For the initial distribution, we provided gaussian distribution, laplace distribution, and random distribution. The results presented in Table 5 show that our method achieves the best performance when setting the temporal dimension of implicit pose proxy to 81 and using the gaussian distribution initialization.

PIM	PUM	MPJPE ↓	P-MPJPE ↓
MLP	MLP	40.8	33.8
CrossAttention	MLP	39.6	32.6
MLP	CrossAttention	39.5	32.8
CrossAttention	**CrossAttention**	**38.7**	**32.3**

Table 6: Analysis of the various micro designs within proxy update module and proxy invocation module.

Range	Strategy	MPJPE ↓	P-MPJPE ↓
0	Fixed	38.4	32.4
0	Trainable	38.2	32.0
(0, 1)	**Trainable**	**37.9**	**31.7**
(-1, 0)	Trainable	38.5	32.3
(-1, 1)	Trainable	37.9	32.0

Table 7: Analysis on the adaptive fusion. Range denotes the sampling range of μ. Strategy denotes whether μ is trainable.

Analysis on Micro Design. In this section, we further explore the effectiveness of various micro designs within the proxy update module (PUM) and proxy invocation module (PIM). As shown in Table 6, we achieve the best performance when both PUM and PIM use cross attention.

Analysis on Proxy Attention Module. We extensively investigated the fusion strategies of adaptive fusion within our proxy attention module. Specifically, we pay attention to the sampling range of μ and whether it is trainable. As shown in Table 7, we achieved the best performance when we allowed μ to be trainable and sampled it from (0, 1).

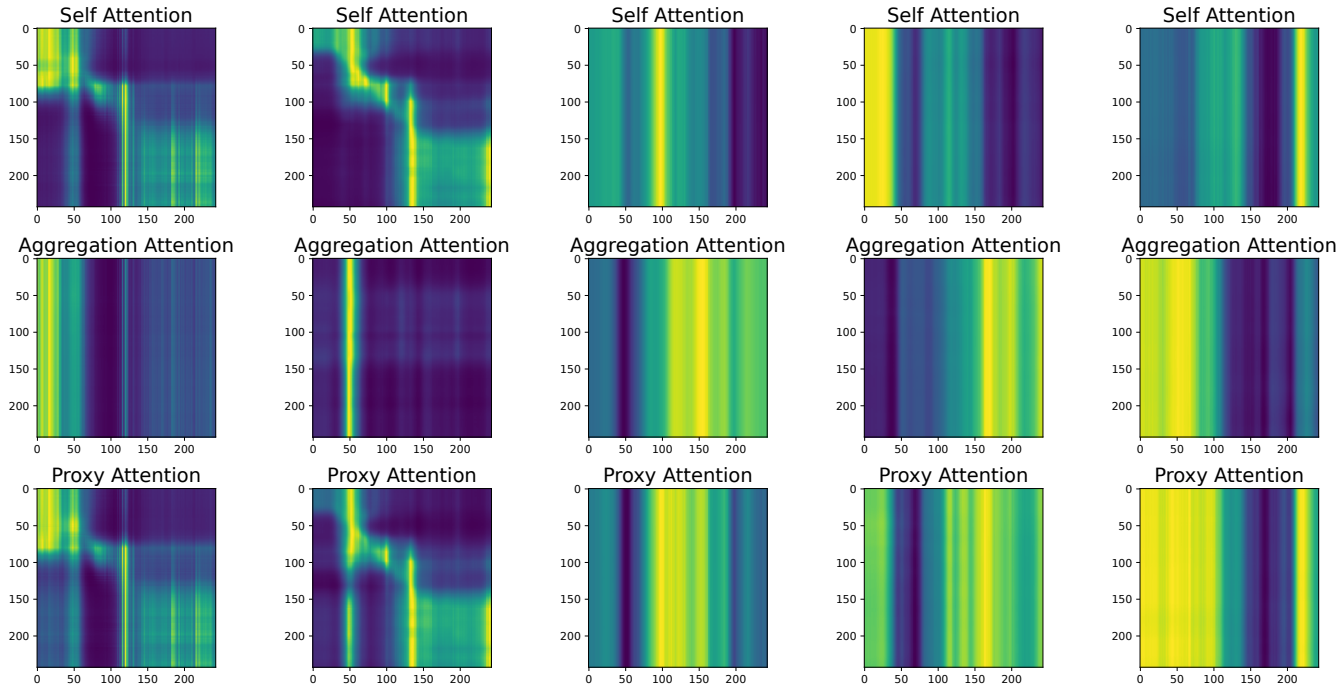


Figure 3: Visualizations of different attention matrices. The first row is the original self-attention matrix. The second row is the aggregation attention matrix. The third row is our proxy attention matrix. As expected, our proxy attention matrix effectively leverages the aggregation attention matrix to complement the missing parts of the original self attention matrix.

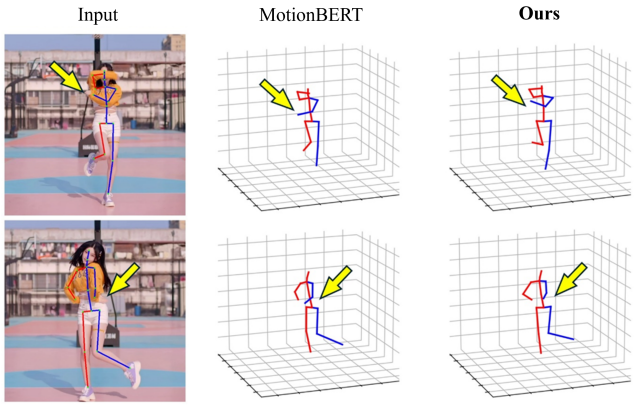


Figure 4: Qualitative comparisons of our TCPFormer with MotionBERT on in-the-wild videos. The yellow arrows indicate locations where our method achieves better results.

Qualitative Analysis. We visualized the original self attention matrix (first row), aggregation attention matrix (second row), and proxy attention matrix (third row) in Figure 3. All attention matrices are normalized to $[0, 1]$. As expected, our proxy attention matrix effectively leverages the aggregation attention matrix to complement the original self attention matrix. Furthermore, we also present 3D human pose estimation results by MotionBERT (Zhu et al. 2023) and our TCPFormer on the Human3.6M dataset and in-the-wild videos. As shown in Figure 4 and Figure 5, TCPFormer achieves better qualitative results compared with MotionBERT (Zhu et al. 2023).

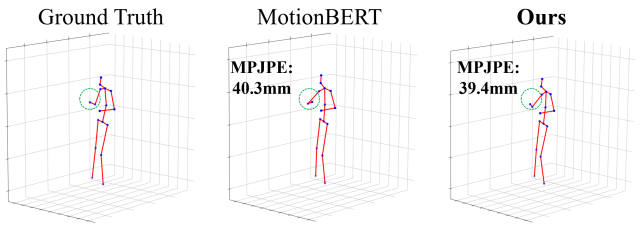


Figure 5: Qualitative comparisons of our TCPformer with MotionBERT on Human3.6M. The green circles indicate locations where our method achieves better results.

Conclusion

In this paper, we present TCPFormer, a novel method to learn temporal correlation with implicit pose proxy. Different from previous methods that learn complex temporal correlations only through single mapping, TCPFormer leverages the implicit pose proxy as an intermediate representation to skillfully model the complex temporal correlation within the pose sequence and effectively use the temporal information to facilitate 3D human pose estimation. The visualization results provide empirical evidence that our TCPFormer can build comprehensive temporal correlation within the 2D pose sequence. Extensive experimental results also show that our TCPFormer outperforms the previous state-of-the-art approaches on the Human3.6M and MPI-INF-3DHP datasets.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62203476), Natural Science Foundation of Guangdong Province (No. 2024A1515012089), Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (No. RCYX20231211090248064).

References

Agarwal, A.; and Triggs, B. 2005. Recovering 3D human pose from monocular images. *IEEE TPAMI*, 28(1): 44–58.

Andriluka, M.; Roth, S.; and Schiele, B. 2009. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 1014–1021.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.

Cai, Q.; Hu, X.; Hou, S.; Yao, L.; and Huang, Y. 2024. Disentangled Diffusion-Based 3D Human Pose Estimation with Hierarchical Spatial and Temporal Denoiser. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 882–890.

Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.-J.; Yuan, J.; and Thalmann, N. M. 2019. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In *ICCV*, 2272–2281.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229.

Chen, T.; Fang, C.; Shen, X.; Zhu, Y.; Chen, Z.; and Luo, J. 2021. Anatomy-aware 3D human pose estimation with bone-based pose decomposition. *IEEE TCSVT*, 32(1): 198–209.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 7103–7112.

Ci, H.; Wang, C.; Ma, X.; and Wang, Y. 2019. Optimizing network structure for 3D human pose estimation. In *ICCV*, 2262–2271.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Fang, H.-S.; Xu, Y.; Wang, W.; Liu, X.; and Zhu, S.-C. 2018. Learning pose grammar to encode human body configuration for 3D pose estimation. In *AAAI*, volume 32.

Foo, L. G.; Li, T.; Rahmani, H.; Ke, Q.; and Liu, J. 2023. Unified pose sequence modeling. In *CVPR*, 13019–13030.

Gong, J.; Fan, Z.; Ke, Q.; Rahmani, H.; and Liu, J. 2022. Meta agent teaming active learning for pose estimation. In *CVPR*, 11079–11089.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *ICCV*, 2961–2969.

Hossain, M. R. I.; and Little, J. J. 2018. Exploiting temporal information for 3D human pose estimation. In *ECCV*, 68–84.

Hu, W.; Zhang, C.; Zhan, F.; Zhang, L.; and Wong, T.-T. 2021. Conditional Directed Graph Convolution for 3D Human Pose Estimation. In *ACM MM*, 602–611.

Ionescu, C.; Carreira, J.; and Sminchisescu, C. 2014. Iterated second-order label sensitive pooling for 3D human pose estimation. In *CVPR*, 1661–1668.

Ionescu, C.; Li, F.; and Sminchisescu, C. 2011. Latent structured models for human pose estimation. In *ICCV*, 2220–2227.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 36(7): 1325–1339.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *CVPR*, 7122–7131.

Li, H.; Shi, B.; Dai, W.; Zheng, H.; Wang, B.; Sun, Y.; Guo, M.; Li, C.; Zou, J.; and Xiong, H. 2023. Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1296–1304.

Li, W.; Liu, H.; Ding, R.; Liu, M.; Wang, P.; and Yang, W. 2022a. Exploiting Temporal Contexts with Strided Transformer for 3D Human Pose Estimation. *IEEE TMM*, 25: 1282–1293.

Li, W.; Liu, H.; Tang, H.; Wang, P.; and Van Gool, L. 2022b. MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation. In *CVPR*, 13147–13156.

Li, W.; Liu, M.; Liu, H.; Wang, P.; Cai, J.; and Sebe, N. 2024. Hourglass Tokenizer for Efficient Transformer-Based 3D Human Pose Estimation. In *CVPR*, 604–613.

Liu, K.; Ding, R.; Zou, Z.; Wang, L.; and Tang, W. 2020a. A comprehensive study of weight sharing in graph networks for 3D human pose estimation. In *ECCV*, 318–334.

Liu, R.; Shen, J.; Wang, H.; Chen, C.; Cheung, S.-c.; and Asari, V. 2020b. Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction. In *CVPR*, 5064–5073.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2640–2649.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 506–516.

Moon, G.; and Lee, K. M. 2020. I2l-meshnet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single rgb image. In *ECCV*, 752–768.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *ECCV*, 483–499.

Pavlakos, G.; Zhou, X.; and Daniilidis, K. 2018. Ordinal depth supervision for 3D human pose estimation. In *CVPR*, 7307–7316.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 7025–7034.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 7753–7762.

Peng, J.; Zhou, Y.; and Mok, P. 2024. KTPFormer: Kinematics and Trajectory Prior Knowledge-Enhanced Transformer for 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1123–1132.

Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2012. Reconstructing 3D human pose from 2D image landmarks. In *ECCV*, 573–586.

Shan, W.; Liu, Z.; Zhang, X.; Wang, S.; Ma, S.; and Gao, W. 2022. P-STMO: Pre-Trained Spatial Temporal Many-to-One Model for 3D Human Pose Estimation. In *ECCV*.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 5693–5703.

Sun, X.; Xiao, B.; Wei, F.; Liang, S.; and Wei, Y. 2018. Integral human pose regression. In *ECCV*, 529–545.

Tang, Z.; Qiu, Z.; Hao, Y.; Hong, R.; and Yao, T. 2023. 3D Human Pose Estimation With Spatio-Temporal Criss-Cross Attention. In *CVPR*, 4790–4799.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, 5998–6008.

Wang, J.; Yan, S.; Xiong, Y.; and Lin, D. 2020. Motion guided 3D pose estimation from videos. In *ECCV*, 764–780.

Wang, X.; Fang, Z.; Li, X.; Li, X.; Chen, C.; and Liu, M. 2024. Skeleton-in-context: Unified skeleton sequence modeling with in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2436–2446.

Wang, X.; Zhang, W.; Wang, C.; Gao, Y.; and Liu, M. 2023. Dynamic Dense Graph Convolutional Network for Skeleton-based Human Motion Prediction. *IEEE Transactions on Image Processing (TIP)*.

Xu, T.; and Takano, W. 2021. Graph Stacked Hourglass Networks for 3D Human Pose Estimation. In *CVPR*, 16105–16114.

Ye, M.; Li, H.; Du, B.; Shen, J.; Shao, L.; and Hoi, S. C. 2021. Collaborative refining for person re-identification with label noise. *IEEE TIP*, 31: 379–391.

Yu, B. X.; Zhang, Z.; Liu, Y.; Zhong, S.-h.; Liu, Y.; and Chen, C. W. 2023. Gla-gcn: Global-local adaptive graph convolutional network for 3D human pose estimation from monocular video. In *ICCV*, 8818–8829.

Zhang, C.; Yang, T.; Weng, J.; Cao, M.; Wang, J.; and Zou, Y. 2022a. Unsupervised pre-training for temporal action localization tasks. In *CVPR*, 14031–14041.

Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; and Yuan, J. 2022b. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. In *CVPR*, 13232–13242.

Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; and Metaxas, D. N. 2019. Semantic graph convolutional networks for 3D human pose regression. In *CVPR*, 3425–3435.

Zhao, Q.; Zheng, C.; Liu, M.; and Chen, C. 2023a. A Single 2D Pose with Context is Worth Hundreds for 3D Human Pose Estimation. In *NeurIPS*.

Zhao, Q.; Zheng, C.; Liu, M.; Wang, P.; and Chen, C. 2023b. PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In *CVPR*, 8877–8886.

Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; and Ding, Z. 2021. 3D Human Pose Estimation with Spatial and Temporal Transformers. In *ICCV*, 11656–11665.

Zhou, F.; Yin, J.; and Li, P. 2024. Lifting by image–leveraging image cues for accurate 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7632–7640.

Zhu, W.; Ma, X.; Liu, Z.; Liu, L.; Wu, W.; and Wang, Y. 2023. MotionBERT: A Unified Perspective on Learning Human Motion Representations. In *ICCV*, 15085–15099.