Pose Magic: Efficient and Temporally Consistent Human Pose Estimation with a Hybrid Mamba-GCN Network

Xinyi Zhang[1], **Qiqi Bao**[2], **Qinpeng Cui**[1], **Wenming Yang**[1], **Qingmin Liao**[1]*

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Zhejiang University of Science & Technology, Hangzhou, China
xinyi-zh22@mails.tsinghua.edu.cn, nora919530829@163.com, cqp22@mails.tsinghua.edu.cn, yang.wenming@sz.tsinghua.edu.cn, liaoqm@tsinghua.edu.cn

Abstract

Current state-of-the-art (SOTA) methods in 3D Human Pose Estimation (HPE) are primarily based on Transformers. However, existing Transformer-based 3D HPE backbones often encounter a trade-off between accuracy and computational efficiency. To resolve the above dilemma, in this work, we leverage recent advances in state space models and utilize Mamba for high-quality and efficient long-range modeling. Nonetheless, Mamba still faces challenges in precisely exploiting local dependencies between joints. To address these issues, we propose a new attention-free hybrid spatiotemporal architecture named Hybrid **Ma**mba-**GC**N (Pose Magic). This architecture introduces local enhancement with GCN by capturing relationships between neighboring joints, thus producing new representations to complement Mamba's outputs. By adaptively fusing representations from Mamba and GCN, Pose Magic demonstrates superior capability in learning the underlying 3D structure. To meet the requirements of real-time inference, we also provide a fully causal version. Extensive experiments show that Pose Magic achieves new SOTA results ($\downarrow 0.9mm$) while saving 74.1% FLOPs. In addition, Pose Magic exhibits optimal motion consistency and the ability to generalize to unseen sequence lengths.

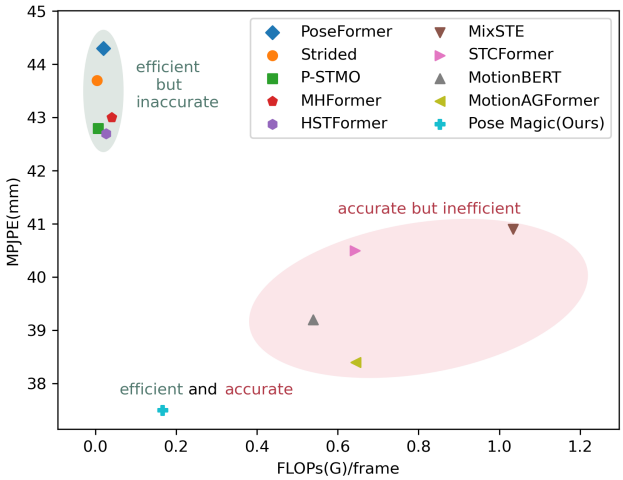


Figure 1: Comparisons of transformer-based methods on Human3.6M (↓). FLOPs/frame denotes floating point operations per output frame. The proposed Pose Magic attains superior results, while maintaining computational efficiency.

Introduction

Monocular 3D Human Pose Estimation (HPE) aims to capture positions of joints on the human skeleton in 3D space from images or videos. It is widely applied in action recognition (Peng et al. 2024), action correction and online coaching (Dittakavi et al. 2022), and augmented/virtual reality (Yuan et al. 2023). With these extensive applications, the demand for more accurate, computationally efficient and temporally consistent models continues to grow.

Impressive progress has been made in monocular 3D HPE (Zhu et al. 2023; Mehraban, Adeli, and Taati 2024). But it inherently remains an ill-posed problem due to depth ambiguity. To mitigate this issue in a single frame, some efforts have been made to comprehensively exploit the spatial and temporal relationships between joints contained in the input video. Current state-of-the-art (SOTA) methods (Li et al. 2022a,b; Zhang et al. 2022; Zhu et al. 2023) use Transformers (Vaswani et al. 2017) to capture spatiotemporal information from 2D pose sequences. Their outstanding performance is largely attributed to the powerful self-attention mechanism and the ability to capture long-range dependencies. However, despite the long-range receptive field, Transformer-based methods have high computational complexity (quadratic growth with the number of frames), making it challenging to deploy on devices with limited computational resources. Moreover, employing efficient techniques such as token pruning (Ma et al. 2022; Li et al. 2023b) for 3D HPE often sacrifices a globally effective receptive field. These methods do not inherently resolve the trade-off between accuracy and computational efficiency.

Recently, achievements in State Space Models (SSMs), particularly Mamba (Gu and Dao 2023; Liu et al. 2024; Zhu et al. 2024), have established them as an efficient and effective backbone for building deep networks. To address the above issue, **we propose a novel attention-free spatiotemporal architecture.** Specifically, both spatial and temporal Mamba blocks are included, which effectively learn rich

*Corresponding author

spatiotemporal features. This new attention-free structure inherits the computational efficiency advantages of Mamba. Additionally, due to Mamba's inherent continuous-time formulation, the consistency and smoothness in pose estimation over time are improved.

However, standard Mamba still primarily accommodates long-range dependencies and pays less attention to local dependencies, which is detrimental to capturing human motion. Human movement inherently comprises local spatial and temporal dependencies (Mehraban, Adeli, and Taati 2024). To address the above challenge, **we propose a hybrid spatiotemporal architecture** combining Mamba and Graph Convolutional Networks (GCNs) (Kipf and Welling 2016), named Hybrid **Ma**mba-**GC**N (Pose Magic), to learn 3D pose representations from global to local. Specifically, Pose Magic consists of two streams: the Mamba stream and the GCN stream, which learn global and local information, respectively. In the GCN stream, GCN sequentially learns spatial and temporal relationships similar to the Mamba stream. We then employ adaptive fusion to aggregate features from both Mamba and GCN streams. By doing so, we ensure a balanced and comprehensive representation of human motion, thereby improving the accuracy of 3D HPE. Additionally, GCN has lower computational complexity, making our model still computationally efficient, as shown in Fig.1. We also introduce a fully causal version of Pose Magic, designed to predict each current timestep using only past and present frames, without forecasting future frames. This is crucial for real-time applications.

In summary, our main contributions are as follows:

- We propose a novel attention-free hybrid network, Pose Magic. It leverages Mamba for high-quality and efficient long-range modeling. Additionally, GCN is utilized to enhance Mamba's performance by effectively capture local dependencies between joints. To the best of our knowledge, this is the first attempt to apply Mamba to 3D HPE tasks.
- We propose two versions of Pose Magic, achieved by designing the Bidirectional Magic Block and Unidirectional (causal) Magic Block, corresponding to offline and real-time inference, respectively.
- Extensive experiments on two popular benchmarks demonstrate that Pose Magic achieves state-of-the-art results while maintaining efficiency with fewer parameters and lower computational complexity.
- Pose Magic can model the natural smoothness of human motion, achieving optimal motion consistency and generalizing well to unseen sequence lengths.

Related Work

3D Human Pose Estimation

Based on the number of camera views used, 3D HPE can be categorized into monocular (Li et al. 2022b; Zhang et al. 2022; Zhu et al. 2023) and multi-view methods (Qiu et al. 2019; Zhang et al. 2024b,a). Due to the accessibility of a single camera in real-world scenarios, more attention has been paid to monocular 3D HPE, i.e., estimating 3D human pose from monocular images or videos. Mainstream monocular 3D HPE methods follow a two-stage paradigm. First, a plug-and-play 2D pose detector (Newell, Yang, and Deng 2016; Chen et al. 2018) is used to locate the 2D positions of joints. Then, the 2D pose is lifted to the 3D pose. In this work, we focus on the latter challenging step, known as the 2D-to-3D lifting process, following Zhu et al. (2023); Mondal, Alletto, and Tome (2024).

Transformer-based Methods for 3D HPE

Numerous studies have demonstrated Transformer (Vaswani et al. 2017) exhibits superior performance in 3D HPE. PoseFormer (Zheng et al. 2021) is the first purely Transformer-based model, utilizing a spatial-temporal transformer to model joint relationships within each frame as well as temporal associations across frames. Subsequently, many studies have focused on enhancing spatiotemporal feature encoding. MixSTE (Zhang et al. 2022) proposed learning distinct motion trajectories for each joint. HSTFormer (Qian et al. 2023) introduced a hierarchical Transformer encoder to capture multi-level spatiotemporal correlations from local to global. MotionBERT (Zhu et al. 2023) presented a dual-stream spatiotemporal Transformer that separately attends to long-range spatiotemporal relationships between stable and dynamic joints. To overcome self-occlusion and depth ambiguity, MHFormer (Li et al. 2022b) learned spatiotemporal multi-hypothesis representations through Transformers.

However, performance gains come with significant computational overhead. Consequently, researchers have begun exploring efficient methods. Strided (Li et al. 2022a) designed a cross-row Transformer encoder to aggregate redundant sequences, while Shan et al. (2022); Zeng et al. (2022a); Einfalt, Ludwig, and Lienhart (2023) improved efficiency by uniform sampling of video sequences. In terms of token pruning, PPT (Ma et al. 2022) selected important tokens based on attention scores. TCFormer (Zeng et al. 2022b) proposed a token clustering Transformer to merge tokens. In contrast to spatial reduction methods, Hourglass Tokenizer (Li et al. 2023b) selected representative frame tokens in the temporal domain. However, these methods may sacrifice accuracy as they lose contextual cues.

State Space Models

Recent studies (Gu, Goel, and Ré 2021; Gu et al. 2022; Gupta, Gu, and Berant 2022; Smith, Warrington, and Linderman 2022) have explored SSMs as an effective alternative to Transformers for efficient modeling. For instance, Gu, Goel, and Ré (2021); Gu et al. (2022) proposed the S4 model and its diagonal version S4D, addressing issues of computational efficiency and long sequence dependencies. However, these models process all inputs in the same manner, which limits their data modeling capabilities. To enhance content-awareness, Mamba (Gu and Dao 2023) integrated time-varying parameters into the SSM framework and proposed a hardware-aware algorithm for efficient training and inference. Some researchers have applied SSMs to computer vision tasks. Vim (Zhu et al. 2024) introduced a bidirectional SSM block (Wang et al. 2022) for efficient and versatile visual representation, achieving performance comparable to

ViT (Dosovitskiy et al. 2020). Mondal, Alletto, and Tome (2024) applied S4D to accelerate training and real-time inference in 3D HPE, albeit at the cost of sacrificing accuracy. In this work, we explore adapting Mamba to 3D HPE, effectively achieving a balance between accuracy and efficiency.

Graph Convolutional Networks

Human motion inherently involves local spatial and temporal dependencies (Mehraban, Adeli, and Taati 2024). Therefore, it is crucial to model local dependencies. GCNs are computational efficient networks that specialize in local dependencies, achieving notable success in 3D HPE (Ci et al. 2019; Zhao et al. 2019; Yu et al. 2023). GLA-GCN (Yu et al. 2023) used global spatiotemporal representation and local joint representation to achieve a lower memory load while maintaining accuracy. MotionAGFormer (Mehraban, Adeli, and Taati 2024) integrated local spatial and temporal relationships using GCNs to complement the global information extracted by Transformers. However, existing methods do not consider temporal causality, which is impractical for real-time applications. In this work, we introduce local enhancement with GCN to complement the global outputs of Mamba. On this basis, we explore causal GCN methods by designing causal adjacency matrices.

Method

Preliminaries

Selective Structured State Space Models. SSMs (Gu, Goel, and Ré 2021; Gu and Dao 2023) define a continuous system that maps a 1D sequence $x(t) \in \mathbb{R}$ to $y(t) \in \mathbb{R}$ through implicit latent states $h(t) \in \mathbb{R}^N$. This process is described by ordinary differential equations as follows:

$$\begin{aligned} h'(t) &= \boldsymbol{A}h(t) + \boldsymbol{B}x(t), \\ y(t) &= \boldsymbol{C}h(t), \end{aligned} \tag{1}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are learned matrices. Instead of directly initializing $\boldsymbol{A}$ randomly, a popular strategy is to impose a diagonal structure on $\boldsymbol{A}$ (Gu et al. 2022; Gupta, Gu, and Berant 2022; Smith, Warrington, and Linderman 2022).

To enhance computational efficiency, it is necessary to discretize continuous variables $\boldsymbol{A}$ and $\boldsymbol{B}$. The choice of discretization criteria is varied, with Zero-Order Hold (Iserles 2009) being a common approach.

$$\overline{\boldsymbol{A}} = \exp(\Delta\boldsymbol{A}) \quad \overline{\boldsymbol{B}} = (\Delta\boldsymbol{A})^{-1}(\exp(\Delta\boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta\boldsymbol{B}, \tag{2}$$

where Δ represents the step size, and $\boldsymbol{I}$ is the identity matrix.

Thus, the discrete version of Eq.(1) can be written as:

$$\begin{aligned} h_t &= \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t \\ y_t &= \boldsymbol{C}h_t. \end{aligned} \tag{3}$$

This Linear Time-Invariant (LTI) discrete system allows efficient computation in recursive or convolution forms, scaling linearly or near-linearly with sequence lengths.

However, the selective state space model Mamba (Gu and Dao 2023) highlights that LTI limits the model's ability in data modeling. This is because matrices in SSMs remain unchanged regardless of input, thereby lacking content-based inference capabilities. Therefore, Mamba removes the constraint of LTI and introduces time-varying parameters, i.e.,

$$\begin{aligned} \boldsymbol{B} = \mathrm{Linear}_N(x), \boldsymbol{C} = \mathrm{Linear}_N(x), \\ \Delta = \mathrm{softplus}(\mathrm{Parameter} + \mathrm{Broadcast}_D(\mathrm{Linear}_1(x))), \end{aligned} \tag{4}$$

where Linear_d is a parameterized projection layer projecting to dimension d, and $\mathrm{softplus}$ is an activation function.

However, the absence of LTI leads to a loss of equivalence with convolution, affecting training efficiency. To address this issue, Mamba introduces a hardware-aware algorithm, enabling parallelization. As a result, Mamba enables high-quality and efficient long-sequence dynamic modeling.

Graph Convolutional Networks. Unlike Mamba capturing global information, GCNs (Kipf and Welling 2016) excel at aggregating local dependencies. A widely used GCN module (Luo et al. 2022) is defined as:

$$\mathrm{GCN}(x) = \sigma\Big(x + \mathrm{Norm}\Big(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}xW_1 + xW_2\Big)\Big), \tag{5}$$

where $\tilde{A} = A + \boldsymbol{I}$ represents the adjacency matrix with self-connections added. $\tilde{D}ii = \Sigma_j \tilde{A}_{ij}$ represents the summation of $\tilde{A}$ along its diagonal. W_1 and W_2 are trainable weight matrices. $\mathrm{Norm}(\cdot)$ and $\sigma(\cdot)$ denote Batch Normalization (Ioffe and Szegedy 2015) and ReLU activation function (Glorot, Bordes, and Bengio 2011), respectively.

The key to GCNs lies in constructing the adjacency matrix $\tilde{A}$, which should be customized based on specific tasks.

Overall Architecture

To accurately lift 2D skeleton sequences to 3D pose sequences in a monocular setup, we propose the attention-free architecture Pose Magic. It employs Mamba and GCN to lift motion sequences, as shown in Fig.2.

Formally, given a monocular video of 2D joints with confidence scores $X^{2D} \in \mathbb{R}^{T\times J\times 3}$, our goal is to predict 3D positions $\hat{X}^{3D} \in \mathbb{R}^{T\times J\times 3}$. Here, T and J are the number of frames and joints, respectively. First, a linear projection layer is used to map each joint in every frame to a d-dimensional feature $X(0) \in \mathbb{R}^{T\times J\times d}$. Then, positional embedding $P_{pos} \in \mathbb{R}^{T\times J\times d}$ is added. Next, N Magic Blocks effectively capture 3D structure of the skeleton sequence by computing $X(i) \in \mathbb{R}^{T\times J\times d}(i = 1, \ldots, N)$. Subsequently, $X(N)$ is mapped to a higher dimension $M \in \mathbb{R}^{T\times J\times d'}$ using a linear layer and a $\tanh$ activation. Finally, a regression head is employed to output the 3D pose $\hat{X}^{3D} \in \mathbb{R}^{T\times J\times 3}$.

To ensure the temporal smoothness of the predicted 3D poses, we use a positional loss $\mathcal{L}_{3D}$ and a velocity loss $\mathcal{L}_v$ for supervision, i.e.,

$$\mathcal{L} = \mathcal{L}_{3D} + \lambda\mathcal{L}_v, \tag{6}$$

where $\mathcal{L}_{3D} = \sum_{t=1}^{T}\sum_{j=1}^{J} \|\hat{X}^{3D}_{t,j} - X^{3D}_{t,j}\|$, $\mathcal{L}_v = \sum_{t=2}^{T}\sum_{j=1}^{J} \|\Delta\hat{X}^{3D}_{t,j} - \Delta X^{3D}_{t,j}\|$. $\Delta\hat{X}^{3D}_t = \hat{X}^{3D}_t - \hat{X}^{3D}_{t-1}$ and $\Delta X^{3D}_t = X^{3D}_t - X^{3D}_{t-1}$. λ is a balancing constant.

In the subsequent sections, we present the general architectures of bidirectional and unidirectional (causal) Magic Blocks, showing how to capture spatiotemporal information.

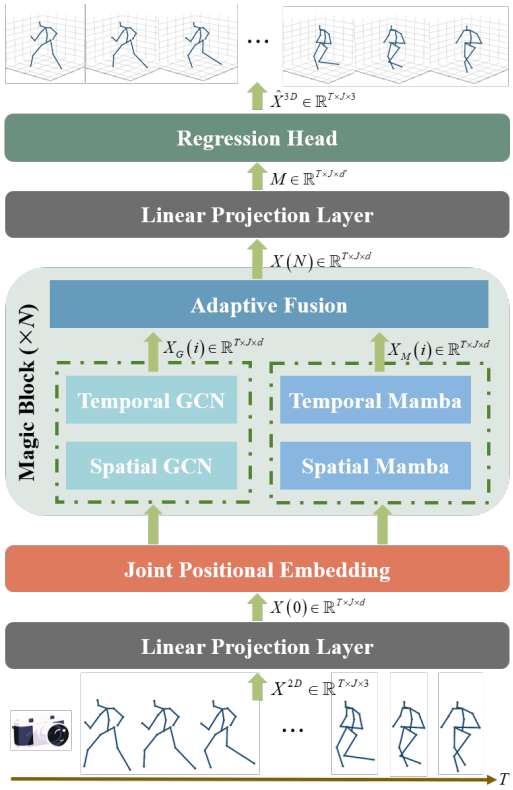
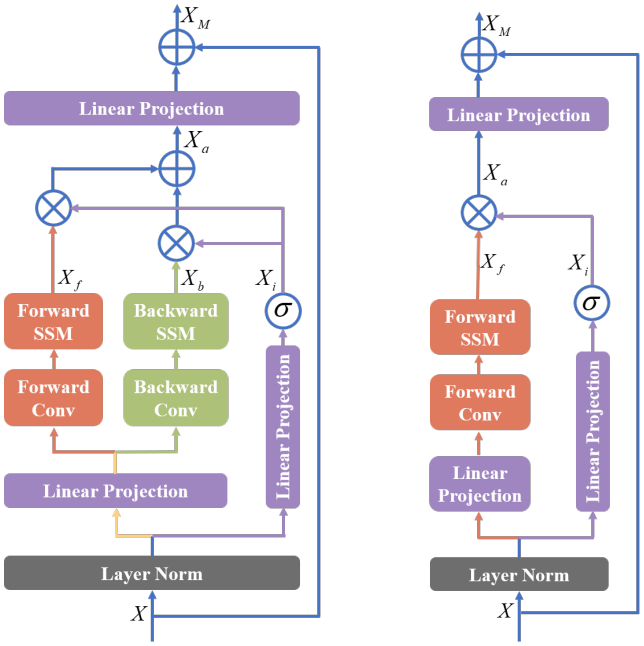


Figure 2: Overview of Pose Magic. It consists of N dual-stream Magic Blocks, with GCN capturing local information and Mamba capturing global information. Spatial GCN/Mamba models connections among joints within a frame, while the Temporal one tracks each joint's motion over time.



(a) Bidirectional Mamba **(b) Unidirectional Mamba**

Figure 3: Different Mamba structures. (a) Bidirectional: process information forward, backward and independently. (b) Unidirectional: process information forward and independently. Here, current information only relates to present and past data, making it suitable for real-time applications.

Bidirectional Magic Block

As shown in Fig.2, Magic Block comprises two streams: Mamba stream and GCN stream. Mamba stream leverages its high-quality and efficient long-range modeling capabilities to capture global dependencies. Meanwhile, GCN stream enhances the power of Mamba by effectively capturing local dependencies between joints. This dual-stream architecture ensures comprehensive and accurate 3D pose estimation. Specifically, Spatial Mamba/GCN treats different joints as individual tokens, effectively capturing the structural relationships of joints within a frame. Temporal Mamba/GCN considers each frame as a single token, thereby capturing the motion trajectory of joints over time. Finally, features captured by both streams are adaptively fused.

Mamba Stream. To bidirectionally capture spatiotemporal information, we adopt the structure of Zhu et al. (2024), as shown in Fig.3(a). This stream has three parallel processing paths: forward, backward, and independent.

Formally, given an input $X \in \mathbb{R}^{B\times L\times d}$, where B is the batch size, L is the length of the sequence and d is the dimension. First, information is processed forward and backward along the sequence dimension, respectively:

$$X_f = \mathrm{SSM}_f(\sigma(\mathrm{Norm}_l(X)W_{p1}W_f)), \tag{7}$$

$$X_b = \mathrm{flip}(\mathrm{SSM}_b(\sigma(\mathrm{flip}(\mathrm{Norm}_l(X)W_{p1})W_b))). \tag{8}$$

The final path processes information independently:

$$X_i = \sigma(\mathrm{Norm}_l(X)W_{p2}), \tag{9}$$

where $\mathrm{Norm}_l(\cdot)$ represents Layer Normalization, $\sigma(\cdot)$ is the GELU activation function (Hendrycks and Gimpel 2016), and $\mathrm{flip}(\cdot)$ indicates flipping along the sequence dimension. $W_{p1}, W_{p2}, W_f, W_b \in \mathbb{R}^{d\times d}$ are learnable matrices.

Next, information from the forward, backward and independent paths is aggregated via multiplicative gating:

$$X_a = X_f \odot X_i + X_b \odot X_i, \tag{10}$$

where $\odot$ denotes a Hadamard Product.

Finally, a skip connection is added to compute the output:

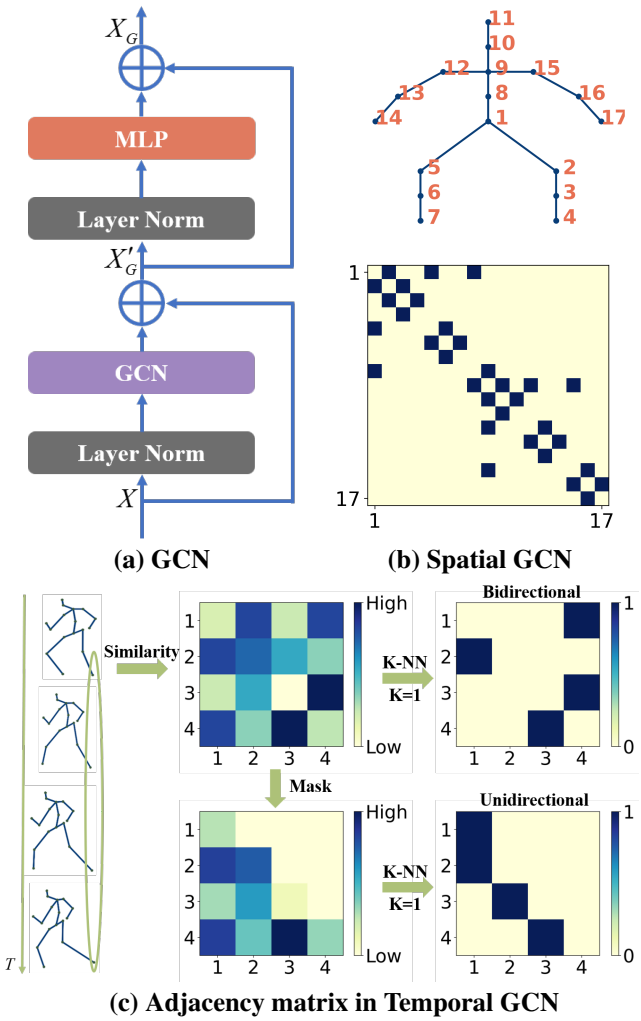
$$X_M = X + X_a W_{p3}, \tag{11}$$

where $W_{p3} \in \mathbb{R}^{d\times d}$ is a learnable matrix.

GCN Stream. Unlike Mamba stream aggregating global information, GCN stream focuses on local spatial and temporal relationships to incorporate pose-specific priors. Our GCN structure is shown in Fig.4(a). Given $X \in \mathbb{R}^{B\times L\times d}$, the output X_G is obtained by

$$\begin{aligned} X'_G &= X + \mathrm{GCN}(\mathrm{Norm}(X)), \\ X_G &= X'_G + \mathrm{MLP}(\mathrm{Norm}(X'_G)), \end{aligned} \tag{12}$$

where $\mathrm{GCN}(\cdot)$ is defined in Eq.(5).



(a) GCN (b) Spatial GCN

(c) Adjacency matrix in Temporal GCN

Figure 4: (a) GCN structure. (b) Spatial GCN uses the Human3.6M skeleton as the adjacency matrix. (c) Temporal GCN uses K-NN for connection edges based on joint similarity across frames. After K-NN, each row connects to K columns. Top: bidirectional adjacency matrix. Bottom: unidirectional adjacency matrix: K-NN after a causal mask.

As mentioned in *Preliminaries*, adjacency matrices in GCNs should be customized based on specific requirements. In Spatial GCN, joint topology is chosen as the adjacency matrix (Fig.4(b)). For Temporal GCN, we first calculate the similarity of individual joints across different frames:

$$\mathrm{Sim}(X_{t_i}, X_{t_j}) = (X_{t_i})^T X_{t_j}. \tag{13}$$

Then, k nearest neighbors are selected as connected nodes in the temporal graph (Fig.4(c)). In a bidirectional setup, the model can access information from the entire sequence, thus a node at a certain frame may relate to past frames as well as future frames.

Adaptive Fusion. Following Mehraban, Adeli, and Taati (2024), we employ adaptive fusion to aggregate features extracted from Mamba and GCN streams:

$$\begin{aligned} X(i) &= \alpha_M(i) \odot X_M(i-1) + \alpha_G(i) \odot X_G(i-1), \\ \alpha_M(i), \alpha_G(i) &= \mathrm{softmax}(W[X_M(i–1), X_G(i–1)]), \end{aligned} \tag{14}$$

where $X(i)$ denotes feature embeddings extracted at depth i, $X_M(i-1)$ and $X_G(i-1)$ represent features extracted from Mamba and GCN streams respectively at depth $i-1$. W is a learnable matrix.

Unidirectional Magic Block

To ensure the temporal causality, we introduce the Unidirectional Magic Block, which also includes the branching and fusion of Mamba and GCN.

Mamba Stream. As shown in Fig.3(b), for an input $X \in \mathbb{R}^{B\times L\times d}$, Unidirectional Mamba combines information along the sequence dimension L but only learns in the forward and independent directions. In this case, Eq.(10) is rewritten to aggregate the forward and independent paths:

$$X_a = X_f \odot X_i. \tag{15}$$

GCN Stream. The structure of Unidirectional GCN remains as shown in Fig.4(a), but its temporal adjacency matrix undergoes some changes. Due to the causal nature of the model, which can only observe present and past data, the Unidirectional adjacency matrix is obtained by applying a causal mask to the similarity matrix, followed by filtering via K-nearest neighbors (K-NN). Specifically, we zero out similarities corresponding to frames occurring after the current time. This prevents K-NN from selecting temporal connections from future frames.

Adaptive Fusion. Features extracted from Mamba and GCN streams are also aggregated according to Eq.(14).

Experiments

Datasets and Evaluation Metrics

Experiments are conducted on widely used 3D HPE benchmark datasets, including Human 3.6M (Ionescu et al. 2013) and MPI-INF-3DHP (Mehta et al. 2017).

Human3.6M (H3.6M) is one of the largest indoor motion capture datasets, consisting of 3.6 million frames from 11 actors across 15 scenarios. Following Cai et al. (2024), our model is trained on subjects S1, S5, S6, S7 and S8, and tested on subjects S9 and S11. We evaluate our single-view temporal 3D HPE model using three metrics: Mean Per Joint Position Error (MPJPE, mm) (Pavllo et al. 2019) for joint accuracy, Mean Per Joint Velocity Error (MPJVE, mm/s) and Acceleration Error (ACC-ERR, mm/s^2) (Mehraban, Adeli, and Taati 2024) for temporal consistency and smoothness, which are essential for video applications. MPJVE and ACC-ERR correspond to the MPJPE of the first and second derivatives of 3D pose sequences.

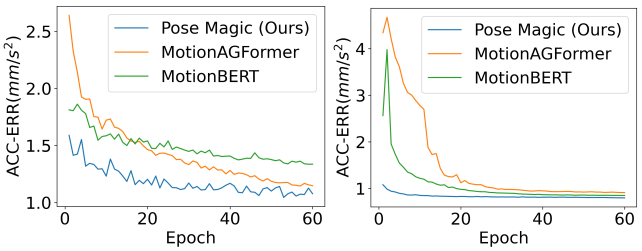
MPI-INF-3DHP (3DHP) is a large dataset collected in various indoor and outdoor environments. It includes over 1.3 million frames, capturing 8 actors performing 8 different activities. Following Shan et al. (2023), MPJPE, Percentage of Correct Keypoint (PCK) within 150 mm, and the Area Under the Curve (AUC) are reported as evaluation metrics.

Method	T	Params	FLOPs	MPJPE ↓
Causal				
MotionBERT-scatch (Zhu et al. 2023)*	243	16.00	131.09	44.7
(Mehraban, Adeli, and Taati 2024)*	243	19.00	156.63	42.6
Ours	243	14.21	40.58	41.7
Bidirectional				
PoseFormer (Zheng et al. 2021)	81	9.60	1.63	44.3
(Einfalt, Ludwig, and Lienhart 2023)	351	10.36	1.00	44.2
Strided (Li et al. 2022a)	351	4.35	1.60	43.7
MHFormer (Li et al. 2022b)	351	31.52	14.15	43.0
TPC w. MHFormer (Li et al. 2023b)	351	31.52	8.22	43.0
P-STMO (Shan et al. 2022)	243	7.01	1.74	42.8
HSTFormer (Qian et al. 2023)	81	22.72	2.12	42.7
HDFormer (Chen et al. 2023)	96	3.70	-	42.6
HoT w. MixSTE (Li et al. 2023b)	243	35.00	167.52	41.0
MixSTE (Zhang et al. 2022)	243	33.78	277.25	40.9
STCFormer (Tang et al. 2023)	243	18.93	156.22	40.5
TPC w. MixSTE (Li et al. 2023b)	243	33.78	251.29	39.9
MotionBERT-scatch (Zhu et al. 2023)	243	16.00	131.09	39.2
HoT w. MotionBERT (Li et al. 2023b)	243	16.35	63.21	39.8
TPC w. MotionBERT (Li et al. 2023b)	243	16.00	91.38	39.0
(Mehraban, Adeli, and Taati 2024)	243	19.00	156.63	38.4
Ours	243	14.42	40.58	37.5

Table 1: Comparison of parameters (M), FLOPs (G) and MPJPE with Transformer-based methods on Human3.6M. $T/*$ denote the number of frames / our re-implementation. Red: Best. Blue: Second best.

Method	T	MPJPE↓	PCK↑	AUC↑
Causal				
MotionBERT-scatch (Zhu et al. 2023)	243	25.4	97.9	85.2
(Mondal, Alletto, and Tome 2024)(scatch)	243	24.6	98.2	85.6
(Mehraban, Adeli, and Taati 2024)*	81	18.7	98.4	85.0
Ours	81	16.1	99.1	86.7
Bidirectional				
PoseFormer (Zheng et al. 2021)	81	77.1	88.6	56.4
TPC w. MHFormer (Li et al. 2023b)	9	58.4	94.0	63.3
MHFormer (Li et al. 2022b)	9	58.0	93.8	63.3
MixSTE (Zhang et al. 2022)	27	54.9	94.4	66.5
HoT w. MixSTE (Li et al. 2023b)	27	53.2	94.8	66.5
(Einfalt, Ludwig, and Lienhart 2023)	81	46.9	95.4	67.6
HSTFormer (Qian et al. 2023)	81	41.4	97.3	71.5
HDFormer (Chen et al. 2023)	96	37.2	98.7	72.9
P-STMO (Shan et al. 2022)	81	32.2	97.9	75.8
PoseFormerV2 (Zhao et al. 2023)	81	27.8	97.9	78.8
GLA-GCN (Yu et al. 2023)	81	27.7	98.5	79.1
STCFormer (Tang et al. 2023)	81	23.1	98.7	83.9
(Mondal, Alletto, and Tome 2024)(scatch)	243	18.7	99.0	87.1
MotionBERT-scatch (Zhu et al. 2023)	243	18.2	99.1	88.0
(Mehraban, Adeli, and Taati 2024)	81	16.2	98.2	85.3
Ours	81	14.7	98.8	87.6

Table 2: Comparison on MPI-INF-3DHP. T denotes the number of input frames. * indicates our re-implementation. Red: Best. Blue: Second best.



(a) Causal models (b) Bidirectional models

Figure 5: Comparison results of ACC-ERR on Human3.6M.

Implementation Details

Model Variants. Based on whether causality is considered, we build two different models: a bidirectional model and a causal model. The difference lies in the Temporal Magic Block: bidirectional model uses a Bidirectional Magic Block, while causal model uses a Unidirectional Magic Block. Both models use a bidirectional Spatial Magic Block. For all experiments, the number of layers is $N = 26$, with a hidden dimension of $d = 128$, a motion semantic dimension of $d' = 512$, and the number of temporal neighbors in the GCN stream is $k = 2$.

Experimental Settings. Our model is implemented using PyTorch (Paszke et al. 2017) and trained on two GeForce RTX 3090 GPUs. Following Zhu et al. (2023), horizontal flip augmentation is applied during both training and testing. For training, each mini-batch consists of 8 sequences. The AdamW optimizer (Loshchilov and Hutter 2017) is utilized for 90 epochs with a weight decay of 0.01. The initial learning rate is $8e^{-4}$, with an exponential decay schedule and a decay factor of 0.99. To ensure a fair comparison with previous studies (Zhu et al. 2023; Mehraban, Adeli, and Taati 2024), detected 2D poses from an off-the-shelf 2D pose detector (Newell, Yang, and Deng 2016) are used for H3.6M, while ground truth 2D poses are used for 3DHP.

Comparison with the State-of-the-art Methods

Results on Human3.6M. Current SOTA performance on H3.6M is achieved by Transformer-based architectures, but these come with high computational complexity. We categorize methods based on causality and compare our approach with them in Table 1. In bidirectional settings, Pose Magic not only outperforms previous Transformer-based methods but also significantly reduces computational costs. Specifically, our Pose Magic achieves $37.5mm$ in MPJPE, improving accuracy by $0.9mm$ compared to the previous SOTA (Mehraban, Adeli, and Taati 2024), while saving 74.1% in FLOPs and improving Params by $4.58G$. Causal methods are more suitable for real-time scenarios where future frame information is unavailable. For strong baselines, we also train their non-causal variants. It can be observed that Pose Magic also outperforms existing methods in causal settings.

Results on MPI-INF-3DHP. We further evaluate our method on 3DHP, as shown in Table 2. Under both causal and non-causal settings, our method consistently outperforms other approaches across all metrics, demonstrating its effectiveness in both indoor and outdoor scenarios.

Temporal Consistency and Smoothness

We evaluate the ability to recover smooth 3D human motion from videos using MPJVE and ACC-ERR, as shown in Table 3 and Fig.5. Our method achieves lower MPJVE and ACC-ERR, and converges faster. These results indicate that our method effectively models the natural smoothness of human motion by learning kinematic characteristics of human movement and current frame features in long-term relationships. We attribute this temporal coherence advantage to the inherent continuous-time of Mamba.

Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Causal																
MotionBERT-scatch (Zhu et al. 2023)*	2.6	2.9	2.1	2.9	2.1	2.8	2.6	2.9	1.7	2.5	2.1	2.0	3.6	2.6	2.3	2.5
(Mehraban, Adeli, and Taati 2024)*	2.3	2.5	1.8	2.6	1.8	2.4	2.2	2.6	1.4	2.1	1.8	1.7	3.2	2.4	2.1	2.2
Ours	2.2	2.4	1.8	2.5	1.7	2.2	2.1	2.5	1.4	2.0	1.7	1.6	3.0	2.3	2.0	2.1
Bidirectional																
PoseFormer (Zheng et al. 2021)	3.2	3.4	2.6	3.6	2.6	3.0	2.9	3.2	2.6	3.3	2.7	2.7	3.8	3.2	2.9	3.1
VPose (Pavllo et al. 2019)	3.0	3.1	2.2	3.4	2.3	2.7	2.7	3.1	2.1	2.9	2.3	2.4	3.7	3.1	2.8	2.8
Trajectory Pose (Lin and Lee 2019)	2.7	2.8	2.1	3.1	2.0	2.5	2.5	2.9	1.8	2.6	2.1	2.3	3.7	2.7	3.1	2.7
Anatomy3D (Chen et al. 2021)	2.7	2.8	2.0	3.1	2.0	2.4	2.4	2.8	1.8	2.4	2.0	2.1	3.4	2.7	2.4	2.5
MHFormer (Li et al. 2022b)	2.6	2.7	1.9	2.8	1.9	2.3	2.3	2.6	1.7	2.4	2.0	2.1	3.2	2.7	2.3	2.4
MHFormer++ (Li et al. 2023a)	2.5	2.6	1.9	2.8	1.9	2.2	2.3	2.6	1.7	2.4	1.9	2.0	3.1	2.5	2.2	2.3
MixSTE (Zhang et al. 2022)	2.5	2.7	1.9	2.8	1.9	2.2	2.3	2.6	1.6	2.2	1.9	2.0	3.1	2.6	2.2	2.3
MotionBERT-scatch (Zhu et al. 2023)	1.8	2.1	1.5	2.0	1.5	1.9	1.8	2.1	1.2	1.8	1.5	1.4	2.6	2.0	1.7	1.8
(Mehraban, Adeli, and Taati 2024)	1.8	2.0	1.4	2.0	1.5	2.0	1.8	2.0	1.1	1.7	1.4	1.4	2.5	2.0	1.7	1.8
Ours	1.6	1.8	1.4	1.8	1.4	1.6	1.6	1.9	1.1	1.6	1.3	1.3	2.3	1.9	1.6	1.6

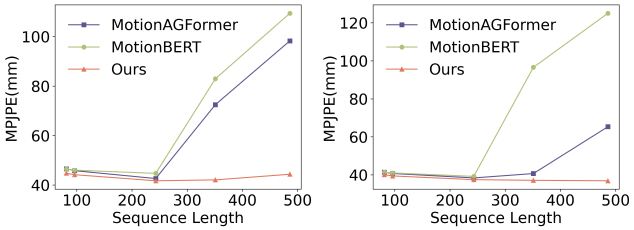
Table 3: Comparison results of MPJVE on Human3.6M. * indicates our re-implementation. Red: Best. Blue: Second best.

Method	MPJPE↓	MPJVE↓
GCN only	54.2	4.6
Mamba only	41.2	1.8
GCN → Mamba (Sequential)	44.2	1.8
Mamba → GCN (Sequential)	46.8	3.9
Mamba → GCN (Parallel)	37.5	1.6

Table 4: Comparison of different integration. All models are trained on Human3.6M with bidirectional settings.

Method	MPJPE↓	MPJVE↓
No Embedding	37.8	1.6
Spatial Embedding only	37.5	1.6
Temporal Embedding only	38.0	1.7
Both Embeddings	38.8	1.7

Table 5: Different types of positional embedding. All models are trained on Human3.6M with bidirectional settings.



(a) Causal models (b) Bidirectional models

Figure 6: Generalization performance on Human3.6M for different sequence lengths.

Ablation Study

Effectiveness of the Proposed Magic Block. To verify the effectiveness of the proposed Magic Block, Table 4 shows alternative blocks. When only using GCN, MPJPE and MPJVE are $54.2mm$ and $4.6mm/s$, indicating its limited ability to accurately and smoothly capture the underlying 3D sequence structure. This is because GCN can only capture local dependencies. The combination of GCN and Mamba results in significant improvements. While local information is also available to Mamba, the parallel stream including GCN allows to balance the integration of local and global information. Compared to using Mamba alone, MPJPE and MPJVE are reduced by $3.7mm$ and $0.2mm/s$.

Impact of Positional Embedding. Table 5 explores the impact of positional embedding on accuracy and smoothness. It can be seen that incorporating temporal positional embedding on top of spatial positional embedding increases MPJPE by $1.3mm$. This is due to the non-permutation equivariant nature of GCN and the temporal continuity of Mamba. Unlike Transformer-based methods, our method inherently maintains the temporal sequence, thus eliminating the need for temporal positional embedding.

Generalization to Unseen Sequence Length. We further investigate the ability of Pose Magic to generalize to unseen sequence lengths. Pose Magic is initially trained on $T = 243$ and then tested on $T = 81, 96, 351, 486$. As shown in Fig.6, Pose Magic demonstrates consistent performance when test on shorter and longer sequences. Compared to MotionBERT (Zhu et al. 2023) and MotionAGFormer (Mehraban, Adeli, and Taati 2024), Pose Magic can successfully generalize to any unseen sequence with minimal generalization loss ($<3mm$), especially in encoding longer contexts. This is particularly beneficial for deployment scenarios where the sequence lengths do not match those used during training.

Conclusion

We propose a novel attention-free hybrid spatiotemporal architecture, Pose Magic, to address the trade-off between accuracy and efficient computation in 3D HPE. Specifically, we introduce the advanced state space model, Mamba, to effectively capture global dependencies. To complement this, we incorporate GCN to capture local joint relationships, enhancing neighborhood similarity and addressing local dependencies. This fusion improves the ability to understand the inherent 3D structure within input 2D sequences. Additionally, we provide a fully causal version of Pose Magic to perform real-time inference. Empirical evaluations demonstrate that Pose Magic achieves SOTA results while maintaining efficiency. Moreover, Pose Magic exhibits optimal motion consistency and smoothness, and can generalize to unseen sequence lengths.

Acknowledgments

This work was partly supported by the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen (No.KJZD20231023094700001) and Zhejiang Provincial Natural Science Foundation of China (No.LQN25F010018).

References

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.

Chen, H.; He, J.-Y.; Xiang, W.; Cheng, Z.-Q.; Liu, W.; Liu, H.; Luo, B.; Geng, Y.; and Xie, X. 2023. Hdformer: High-order directed transformer for 3d human pose estimation. *arXiv preprint arXiv:2302.01825*.

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 Transactions on Circuits and Systems for Video Technology*, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7103–7112.

Ci, H.; Wang, C.; Ma, X.; and Wang, Y. 2019. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2262–2271.

Dittakavi, B.; Bavikadi, D.; Desai, S. V.; Chakraborty, S.; Reddy, N.; Balasubramanian, V. N.; Callepalli, B.; and Sharma, A. 2022. Pose tutor: an explainable system for pose correction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3540–3549.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Einfalt, M.; Ludwig, K.; and Lienhart, R. 2023. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2903–2913.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; Gupta, A.; and Ré, C. 2022. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35: 35971–35983.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Gupta, A.; Gu, A.; and Berant, J. 2022. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35: 22982–22994.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456.

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 transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.

Iserles, A. 2009. *A first course in the numerical analysis of differential equations*. 44. Cambridge university press.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

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 Transactions on Multimedia*, 25: 1282–1293.

Li, W.; Liu, H.; Tang, H.; and Wang, P. 2023a. Multi-hypothesis representation learning for transformer-based 3D human pose estimation. *Pattern Recognition*, 141: 109631.

Li, W.; Liu, H.; Tang, H.; Wang, P.; and Van Gool, L. 2022b. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13147–13156.

Li, W.; Liu, M.; Liu, H.; Wang, P.; Cai, J.; and Sebe, N. 2023b. Hourglass Tokenizer for Efficient Transformer-Based 3D Human Pose Estimation. *arXiv preprint arXiv:2311.12028*.

Lin, J.; and Lee, G. H. 2019. Trajectory space factorization for deep video-based 3d human pose estimation. *arXiv preprint arXiv:1908.08289*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*.

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

Luo, C.; Song, S.; Xie, W.; Shen, L.; and Gunes, H. 2022. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*.

Ma, H.; Wang, Z.; Chen, Y.; Kong, D.; Chen, L.; Liu, X.; Yan, X.; Tang, H.; and Xie, X. 2022. Ppt: token-pruned pose transformer for monocular and multi-view human pose estimation. In *European Conference on Computer Vision*, 424–442.

Mehraban, S.; Adeli, V.; and Taati, B. 2024. MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6920–6930.

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 *2017 international conference on 3D vision*, 506–516.

Mondal, A. K.; Alletto, S.; and Tome, D. 2024. HumMUSS: Human Motion Understanding using State Space Models. *arXiv preprint arXiv:2404.10880*.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 483–499.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 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*, 7753–7762.

Peng, K.; Yin, C.; Zheng, J.; Liu, R.; Schneider, D.; Zhang, J.; Yang, K.; Sarfraz, M. S.; Stiefelhagen, R.; and Roitberg, A. 2024. Navigating open set scenarios for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4487–4496.

Qian, X.; Tang, Y.; Zhang, N.; Han, M.; Xiao, J.; Huang, M.-C.; and Lin, R.-S. 2023. Hstformer: Hierarchical spatial-temporal transformers for 3d human pose estimation. *arXiv preprint arXiv:2301.07322*.

Qiu, H.; Wang, C.; Wang, J.; Wang, N.; and Zeng, W. 2019. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4342–4351.

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 *European Conference on Computer Vision*, 461–478.

Shan, W.; Liu, Z.; Zhang, X.; Wang, Z.; Han, K.; Wang, S.; Ma, S.; and Gao, W. 2023. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14761–14771.

Smith, J. T.; Warrington, A.; and Linderman, S. W. 2022. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*.

Tang, Z.; Qiu, Z.; Hao, Y.; Hong, R.; and Yao, T. 2023. 3D human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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. *Advances in neural information processing systems*, 30.

Wang, J.; Yan, J. N.; Gu, A.; and Rush, A. M. 2022. Pretraining without attention. *arXiv preprint arXiv:2212.10544*.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8818–8829.

Yuan, Y.; Makoviychuk, V.; Guo, Y.; Fidler, S.; Peng, X.; and Fatahalian, K. 2023. Learning physically simulated tennis skills from broadcast videos. *ACM Trans. Graph*, 42(4).

Zeng, A.; Ju, X.; Yang, L.; Gao, R.; Zhu, X.; Dai, B.; and Xu, Q. 2022a. Deciwatch: A simple baseline for $10\times$ efficient 2d and 3d pose estimation. In *European Conference on Computer Vision*, 607–624.

Zeng, W.; Jin, S.; Liu, W.; Qian, C.; Luo, P.; Ouyang, W.; and Wang, X. 2022b. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11101–11111.

Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; and Yuan, J. 2022. 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*, 13232–13242.

Zhang, L.; Zhou, K.; Lu, F.; Zhou, X.-D.; and Shi, Y. 2024a. Deep Semantic Graph Transformer for Multi-View 3D Human Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7205–7214.

Zhang, X.; Cui, Q.; Bao, Q.; Yang, W.; and Liao, Q. 2024b. Geometry-Guided Diffusion Model with Masked Transformer for Robust Multi-View 3D Human Pose Estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 681–690.

Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; and Metaxas, D. N. 2019. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3425–3435.

Zhao, Q.; Zheng, C.; Liu, M.; Wang, P.; and Chen, C. 2023. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 11656–11665.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.

Zhu, W.; Ma, X.; Liu, Z.; Liu, L.; Wu, W.; and Wang, Y. 2023. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15085–15099.