A Structure-aware and Motion-adaptive Framework for 3D Human Pose Estimation with Mamba

Ye Lu[1*] Jie Wang [2*] Jianjun Gao[1] Rui Gong[1] Chen Cai [1] Kim-Hui Yap[1]

[1] Nanyang Technological University [2] Beijing Institute of Technology

{lu0001ye@e.,gaoj0018@e.,gong0084@e.,e190210@e.,ekhyap@}ntu.edu.sg {jwang991020}@gmail.com

Abstract

Recent Mamba-based methods for the pose-lifting task tend to model joint dependencies by 2D-to-1D mapping with diverse scanning strategies. Though effective, they struggle to model intricate joint connections and uniformly process all joint motion trajectories while neglecting the intrinsic differences across motion characteristics. In this work, we propose a structure-aware and motion-adaptive framework to capture spatial joint topology along with diverse motion dynamics independently, named as SAMA. Specifically, SAMA consists of a Structure-aware State Integrator (SSI) and a Motion-adaptive State Modulator (MSM). The Structure-aware State Integrator is tasked with leveraging dynamic joint relationships to fuse information at both the joint feature and state levels in the state space, based on pose topology rather than sequential state transitions. The Motion-adaptive State Modulator is responsible for joint-specific motion characteristics recognition, thus applying tailored adjustments to diverse motion patterns across different joints. Through the above key modules, our algorithm enables structure-aware and motion-adaptive pose lifting. Extensive experiments across multiple benchmarks demonstrate that our algorithm achieves advanced results with fewer computational costs.

1. Introduction

Monocular 3D Human Pose estimation is a fundamental computer vision task, aiming to estimate 3D human poses in 3D space from single-view 2D images or videos. This technique serves as the foundation for a diverse range of applications, including action recognition [33, 35] and human-computer interaction [4, 5, 31]. Approaches to this task generally fall into two categories: directly estimating 3D poses from images or videos [3, 15, 22, 29], detecting 2D poses with off-the-shelf detectors and lifting them into 3D. Due to its more dependable performance, the 2D-to-3D pose lifting has become the mainstream based on robust 2D pose estimators. However, monocular 2D pose often suffers from depth ambiguity, where one single 2D pose can correspond to multiple 3D poses, making it difficult to accurately recover 3D poses from a single frame of 2D keypoints. Current methods address this issue by leveraging temporal information from videos to capture joint dependencies across space and time, achieving significant progress.

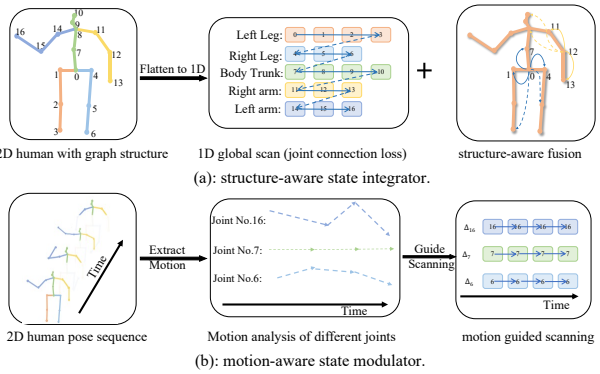

Figure 1. (a) Illustration of structure-aware state integrator. On top of the linear scanning, we aggregate joints based on their connections, supplementing the necessary learnable topology information. (b) Representation of motion-aware modulator. We identify the distinct motion characteristics of different joints and adaptively generate timescales Δ to guide the model in capturing the unique motion features of these joints.

Recently, Mamba-based methods [12, 36] have been introduced to the pose-lifting task using state space models [6, 9, 10], leveraging their linear complexity and effectively capturing detailed spatio-temporal joint dependencies. Despite employing different scanning methods [12, 36], these approaches have limitations in effectively capturing complex joint interactions. Their uniform treatment of joint trajectories tends to overlook the inherent variations in motion patterns across different joints, as shown in Fig. 1. In the spatial domain, human joints are naturally connected by

*Equal contribution

a specific graph structure, where each joint maintains connections with a varying number of neighbor joints. Simply flattening this graph-structured pose into 1D data disrupts its inherent topology, resulting in the loss of crucial structural information and ultimately degrading pose estimation performance. In the temporal domain, joint motions vary significantly, with arms and legs exhibiting high flexibility and large ranges, while the trunk remains more constrained. Previous methods process all joint motion trajectories uniformly, ignoring their intrinsic motion differences, resulting in insufficient learning and suboptimal motion representation. Thus, preserving pose topology and adaptively capturing joint-specific motion dynamics remains a challenge in these Mamba-based methods.

To address these limitations, we propose a structure-aware and motion-adaptive framework named as SAMA, as shown in Fig. 1. It contains a structure-aware state integrator that efficiently fuses dynamic joint relations into the state space. Additionally, it includes a motion-adaptive state modulator to model joint-specific motion dynamics. To incorporate structure-aware joint relationships, the proposed SSI fuses dynamic pose topology within both joint features and states in the state space. Specifically, we introduce a learnable adjacency matrix that encodes both the inherent joint connectivity and the learned global dependencies. This matrix guides the construction of a structure-aware embedding to enhance pose representation and facilitates state fusion in the state space. By integrating structural features, SSI mitigates the limitation of conventional state-space models that rely solely on sequential reasoning. To capture joint-specific motion dynamics, our MSM adaptively regulates the timescale in the SSM, enabling the model to effectively adjust to varying motion patterns across joints. Specifically, it aggregates neighboring frame joint features to learn a joint-specific timescale, which adapts the model's reliance on the previous joint state and current joint input based on the unique motion characteristics of each joint. This adaptive dependency allows MSM to dynamically model diverse joint motion patterns. By integrating SSI and MSM, our model captures the intrinsic connectivity between joints and adaptively learns the motion trajectory characteristics of different joints, achieving significant performance gains with minimal computational costs.

We have extensively validated the effectiveness of our proposed method on multiple datasets, including Human3.6M and more challenging in-the-wild MPI-INF-3DHP. Our method surpasses the previous state-of-the-art (SOTA) methods with fewer parameters and MACs, as shown in Fig. 2. Our experiment results also demonstrate that the proposed modules, SSI and MSM, improve the performance of diverse models, showing their generalization. Our contributions can be summarized as follows:

- We present a new framework, SAMA, which incorporates dynamic joint relations into the state space and captures joint-specific motion dynamics.
- We propose a method that adaptively captures spatiotemporal dependencies and dynamically adjusts the timescale for modeling joint-specific motion dynamics, based on local motion patterns through SSI and MSM.
- We demonstrate the effectiveness of SAMA through extensive experiments across diverse datasets.

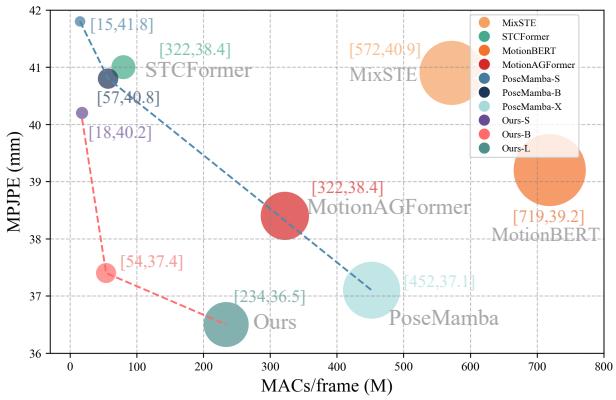

Figure 2. Comparisons of various 3D Human Pose Estimation methods on Human3.6M (↓). MACs/frame represents multiply-accumulate operations per output frame. Radius denotes the parameters. Our method achieves superior results with fewer parameters and computation costs.

2. Related Work

2.1. 2D-to-3D Pose Lifting

Monocular 3D human pose estimation can be divided into two categories: direct 3D human pose estimation and 2D-to-3D pose lifting. Direct regression methods predict 3D human poses from 2D images or videos. End-to-end approaches [23, 26, 28] directly regress 3D poses from images or other raw data but require high computational costs and yield suboptimal results due to operating directly in the image space. In contrast, 2D-to-3D pose lifting methods, which first detect 2D poses and then reconstruct 3D poses from these estimations, have demonstrated superior performance over direct regression approaches. The existing pose lifting methods are classified into two types: Transformer-based methods and GCN-based methods. Transformers [14, 19, 43] are extensively used in pose-lifting tasks for capturing spatial and temporal joint correlations, leveraging their strong global modeling ability. PoseFormer [41] is the first to employ spatial and temporal Transformers separately to capture intra-frame joint dependencies and pose correlations across different frames. MixSTE [34] is a sequence-to-sequence model that alternates between spatial and temporal blocks to capture joint dependencies, and it proposes separately modeling the temporal correlations of different joints. GCN-based methods leverage the connec-

tion of human joints through bones, establishing essential spatial constraints and temporal coherence. SemGCN [38] proposes learning the relationships between directly connected joints and joints that are not physically connected, taking into account dynamic poses across various datasets and real-world applications. In GraFormer [40], the ChebGConv block was introduced to enable information exchange among nodes that lack direct connections, thereby capturing subtle relationships that may not be readily apparent. Overall, Transformer-based methods face challenges to model pose structure and suffer from quadratic complexity, while GCN-based methods lack global modeling capability. In this manuscript, we introduce a novel Mamba-based approach that not only captures the dynamic structure of poses but also incorporates global modeling capabilities.

2.2. Mamba-based Models in Human-Centric Tasks

Mamba [9] achieves Transformer-like capabilities with linear complexity by incorporating a data-dependent selective mechanism and a hardware-aware algorithm to facilitate highly efficient training and inference processes. Based on that, Mamba2 [6] reveals the connections between SSMs and attention with specific structured matrix and explore larger and more expressive state spaces through introducing State Space Duality. In human centric tasks, SSMs have been widely utilized with their strong global modeling ability and linear complexity. Motion Mamba [37] enhances temporal and spatial modeling, while Hamba [7] integrates graph learning with SSMs for structured joint relations. For 2D-to-3D pose lifting task, previous works have leveraged state-space models to model spatiotemporal joint dependencies. PoseMamba [11] proposes a global-local spatiotemporal modeling approach within Mamba framework to address the 2D-to-3D pose lifting task. Posemagic [36] propose a attention-free hybrid spatiotemporal architecture adaptively combining Mamba with GCN. However, these methods merely apply Mamba to the 2D-to-3D pose lifting task without accounting for the unique motion characteristics of human pose sequences and the inherent connections between joints in state space. In this manuscript, we introduce the structure-aware state integrator and the motion-adaptive state modulator to enhance Mamba's ability to capture the unique motion patterns of human pose sequences and the intrinsic connections between joints in state space.

3. Method

3.1. Preliminaries

Mamba in pose lifting. SSMs are widely applied in sequential data analysis and the modeling of continuous linear time-invariant (LTI) systems. This dynamic system can be described by the linear state transition and observation equations: $h'(t) = Ah(t) + Bx(t), y(t) = Ch(t) + Dx(t)$, where $A \in \mathbb{C}^{N\times N}, B, C \in \mathbb{C}^{N}, D \in \mathbb{C}^{1}$ are trainable parameters, $x(t)$ denotes the input sequence, $y(t)$ means the output sequence, and $h(t)$ represents state variable.

In the pose lifting task, the input is a sequence of 2D discrete poses $C_{n,t} \in \mathbb{R}^{N\times T\times 2}$ and the output is a sequence of 3D discrete poses $O_{n,t} \in \mathbb{R}^{N\times T\times 3}$, where N denotes the number of joints in a single frame, and T signifies the total number of frames. To adapt SSMs to this discrete sequence input in the deep learning framework, PoseMamba utilized the Zero-Order Hold (ZOH) discretization, following the setting of Mamba. It discretizes the continuous-time system by assuming the input remains constant within each time interval and introducing a timescale Δ which represents the interval between adjacent timesteps. The ZOH method is applied to compute the discrete system parameters as follows: $\overline{\mathbf{A}} = e^{\Delta\mathbf{A}}, \overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(e^{\Delta\mathbf{A}} - I)\Delta\mathbf{B}$.

In addition, PoseMamba follows context-aware and adaptive SSMs in Mamba through modifying the parameter $\Delta, \overline{\mathbf{B}}, \overline{\mathbf{C}}$ as functions of the input sequence x_t, resulting a data-independent parameters $\mathbf{\Delta}_t = s_\Delta(x_t)$, $\overline{\mathbf{B}}_t = s_B(x_t)$, and $\mathbf{C}_t = s_C(x_t)$. Following previous methods [16, 34, 44], PoseMamba models spatial and temporal joint dependencies separately. In the spatial modeling, PoseMamba processes joints feature in one frame $X_n \in \mathbb{R}^{N\times d}$, where d denotes the dimension of features. The discrete spatial state transition equation and observation equation are formulated as:

$$h_n = \overline{\mathbf{A}}_n h_{n-1} + \overline{\mathbf{B}}_n x_n, \quad y_n = \mathbf{C}_n h_n. \tag{1}$$

In the temporal modeling, PoseMamba processes joints feature in a joint motion trajectory $X_t \in \mathbb{R}^{T\times d}$. The discrete version of the temporal state transition equation and observation equation is similar to Eq. (1).

Mamba2 and State Space Duality. Based on Mamba, Mamba2 draws connection between SSMs and Transformers by introducing Structured State Space Duality (SSD). Different from Mamba1, Mamba2 restrict $\overline{A} = \alpha_t * I$, where I denotes Identity Matrix, leading to the formulation of causal linear attention. Due to the aforementioned connection between SSMs and Transformers, the SSD mixer family of Mamba-2 has been shown to be equivalent to sequentially-semi-separable matrices. The SSD can be expressed as:

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t, \quad y_t = Ch_t, \tag{2}$$

The quadratic form of Eq. (2) can be reformulated as:

$$y_t = P \cdot (C^T B)x, \tag{3}$$

where P_{ij} is defined as follows: $P_{ij} = \overline{A}_{i+1} \times \cdots \times \overline{A}_j$ if $i > j$, $P_{ij} = 1$ if $i = j$, and $P_{ij} = 0$ if $i < j$. Hence, Mamba2 network is regarded as a causal linear attention with a learnable causal mask. In this work, we employ the SSD in Mamba2 as the baseline to construct SAMA due to its training stability and ease of implementation.

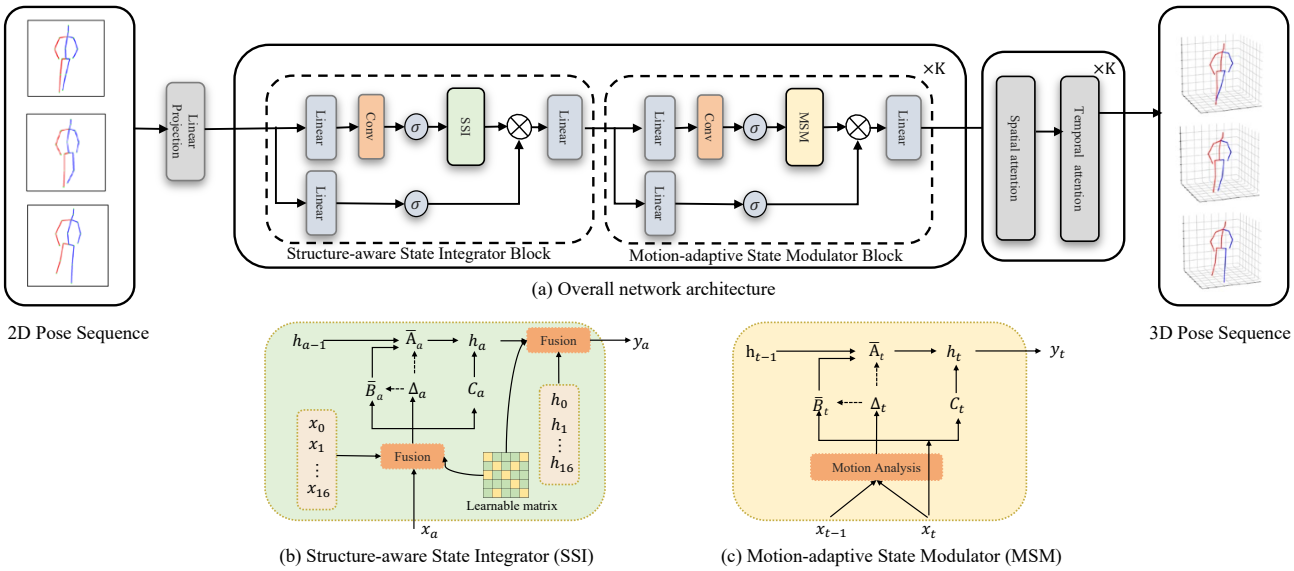

Figure 3. The overview of our proposed SAMA. (a): Our Network Structure. The core part is the alternative stack of structure-aware state integrator block and motion-adaptive state modulator block. (b): Our structure-aware state integrator with structure-aware fusion in state space. (c): Our motion-adaptive state modulator with adaptively joint motion modeling.

3.2. Overall Architecture

As illustrated in Fig. 3 (a), our network processes a 2D pose sequence $C_{n,t} \in \mathbb{R}^{N\times T\times 2}$ and outputs a 3D pose sequence $O_{n,t} \in \mathbb{R}^{N\times T\times 3}$. Firstly, a linear projection layer is used to project the input into high dimension feature $X \in \mathbb{R}^{N\times T\times d}$. In contrast to previous methods, the spatial and temporal position embeddings are not added to the high-dimensional features, because models such as SSM are already capable of capturing token positional order, making the additional positional information redundant. Next, several layers of structure-aware state integrator and motion-adaptive state modulator capture dynamic spatial and temporal joint correlations in an alternating manner. SSI is designed to enable the fusion of joint features and hidden states among joints. Meanwhile, MSM considers the differences in motion characteristics among joints by learning the timescale from the joint motion information to dynamically learn each joint's unique motion properties.

3.3. Structure-aware State Integrator

Structure-aware state integrator is designed to effectively capture the spatial dependencies between adjacent joints within the latent state space, as shown in Fig. 3 (b). To achieve this goal, unlike previous methods that repeatedly scan using different approaches, we introduce a structure-aware state transition into original Mamba formulas. We first construct a learnable matrix to dynamically model the relationships between joints. Then, we use the designed matrix to aggregate joint features and state information.

Construction of the learnable adjacency matrix. To efficiently model joint connections in the state space, a learnable adjacency matrix M is defined as follows:

$$M = \text{softmax}(D^{-\frac{1}{2}}(M_o + I)D^{-\frac{1}{2}}), \tag{4}$$

where D denotes the degree of each joint and I represents the identity matrix. $M_o \in \mathbb{R}^{N\times N}$ means the adjacency matrix and $M \in \mathbb{R}^{N\times N}$ represent a learnable adjacency matrix with global perception and enhanced attention to the connected joint. In Eq. (4), we normalize the adjacency matrix based on joint degrees, as different joints have varying connections. Given the diversity of human actions, we set the normalized adjacency matrix as a learnable parameter to adapt to this variability. Additionally, Eq. (4) provides an initialization for M.

Structure-aware joint feature and state fusion. By using the learnable adjacency matrix, we can achieve the fusion of joint features and states. Since the aggregation is implemented using an $M \in \mathbb{R}^{N\times N}$ matrix, we can save more computational cost compared to the previous method of repeated scanning. The process of the structure-aware joint feature and state fusion can be described by four equations: the joint feature fusion equation, the state transition equation, the structure-aware state fusion equation and the observation equation. In the joint feature fusion equation, we first add structure-aware information to the input through the learnable matrix in Eq. (4):

$$x'_a = x_a + \sum_{k=0}^{N-1} M_{ak}x_k \tag{5}$$

where x_a is the feature of a-th joint, x'_a is the feature of a-th joint after structure-aware joint fusion. Then, we compute the state h_a based on the state transition equation: $h_a = \bar{A}_a h_{a-1} + \bar{B}_a x'_a$. In addition, we also update the

hidden state of joints by incorporating other joint hidden states through the adjacent matrix with the structure-aware state fusion equation :

$$H_a = h_a + \sum_{k=0}^{N-1} M_{ak} h_k \tag{6}$$

where, h_a is the original hidden state, H_a is the structure-aware hidden state. Finally, we employ the observation equation: $y_a = C_a H_a$, where y_a is the output feature of (a)-th joint. Compared with Eq. (1), we can observe that the joint feature and hidden state are directly influenced by other joints, especially the connected joints. However, in previous methods [12, 36], the current joint could only be influenced by joints with a smaller index in the scan.

3.4. Motion-adaptive State Modulator

The previous Mamba-based method, when modeling the temporal motion of joints, ignored the differences in motion characteristics among different joints and simply fed the raw joint trajectories into the SSM. MSM is designed to adaptively learn the motion characteristics of different joints, capturing their unique dynamics and improving motion representation, as shown in Fig. 3 (c). We first propose capturing the motion characteristics of different joints and using these characteristics to dynamically learn the timescale, which controls the model's reliance on the current input and previous state. Then, we introduce two simple methods to model the timescale based on motions.

Motion-aware timescale. The timescale Δ, which controls the balance between how much to focus or ignore the current input, is an important parameter in Mamba and Mamba2. Typically, the timescale is designed as a learnable parameter determined by each token in other tasks. However, the joint motion trajectories exhibit different characteristics across different joints. Specifically, joints in the legs and arms exhibit high motion intensity, so a larger timescale should be used at certain moments to focus on the current input. On the other hand, joints in the body trunk have lower motion intensity, so a smaller timescale should be used to maintain continuity and preserve the state. Different from the previous method, which ignores the motion characteristics of different joints, we use the features of adjacent joints as input to learn the timescale:

$$\Delta_t = S_\Delta(x_t, x_{t-1}) \tag{7}$$

where S_Δ denotes a learnable function, with x_t and x_{t-1} representing the features of the same joint at adjacent time steps. This design enables the timescale to adapt dynamically to varying joint motion characteristics, ensuring a more flexible and responsive modeling of joint dynamics.

Practical implementation. We employ two different functions to model the timescale Δ: point-wise convolution, and linear transformation. For the point-wise convolutions, we use a kernel size of 2 in the temporal dimension and apply zero padding at the start to capture local motion patterns. For the linear transformation, we concatenate adjacent joint features along the feature dimension, with zero padding applied at the start to preserve all the features.

3.5. Network Architecture.

The overall architecture is illustrated in Fig. 3 (a). We alternately stack structure-aware state integrator and motion-adaptive state modulator for K layers. Following Jamba [18], we integrate K layers of spatial and temporal attention to further enhance joint correlation modeling.

3.6. Overall Learning Objectives

Following the previous method [44], we train the model with a end-to-end manner. The final loss is defined as:

$$\mathcal{L} = \mathcal{L}_w + \lambda_m \mathcal{L}_m + \lambda_n \mathcal{L}_n, \tag{8}$$

where L_w is weighted MPJPE, L_m denotes MPJVE, and L_n represents Normalized MPJPE. We set λ_m to 20 and λ_n to the default value of 0.5, respectively.

4. Experiments

We first introduce the experimental setup in §4.1. Then we assess the performance of our method across various datasets, including indoor Human3.6M in §4.2, and more challenging in-the-wild datatset MPI-INF-3DHP in §4.3. Lastly, we provide ablative analyses in §4.4.

4.1. Experimental Setup

Datasets. We conduct experiments on two widely used datasets, Human3.6M [13] and MPI-INF-3DHP [21].

- **Human3.6M** is the most commonly used indoor dataset for monocular 3D human pose estimation task, containing 3.6 million human poses and corresponding images. It includes 11 subjects performing 15 daily activities. Following established protocols in recent studies [12, 44, 44], we take data from subjects 1, 5, 6, 7, 8 for training, and subjects 9, 11 for testing. We take Mean Per-Joint Position Error (MPJPE, $mm, \downarrow$) and Pose-aligned MPJPE (P-MPJPE, $\%, \downarrow$) as the main evaluation matrices. More details are in the supplementary materials.
- **MPI-INF-3DHP** is another challenging large-scale dataset captured in both indoor and outdoor environments, comprising over 1.3 million frames from 8 subjects performing 8 activities. We take Mean Per-Joint Position Error (MPJPE, $mm, \downarrow$), Percentage of Correct Keypoints (PCK, $\%, \uparrow$) and Area Under Curve (AUC, $\%, \uparrow$) as the main evaluation matrices.

Implementation details. Our model, is trained end-to-end, following distinct protocols for dataset as detailed below:

Table 1. Quantitative comparisons on Human3.6M. T: Number of input frames. CE: Estimating center frame only. MACs/frame: multiply-accumulate operations per output frame. P1: MPJPE (mm). P2: P-MPJPE (mm). $P1^{\dagger}$: P1 on 2D ground truth. (*) denotes using HRNet for 2D pose estimation. The best and second-best scores are in bold and underlined, respectively.

Method	T	CE	Param(M)	MACs(G)	MACs/frame(M)	P1↓ /P2↓	$P1^{\dagger}$↓
*MHFormer [CVPR2022] [16]	351	✓	30.9	7.0	20	43.0/34.4	30.5
Stridedformer [TMM2022] [17]	351	✓	4.0	0.8	2	43.7/35.2	28.5
Einfalt *et al.* [WACV2023] [8]	351	✓	10.4	0.5	1	44.2/35.7	-
STCFormer [CVPR2023] [27]	243	×	4.7	19.6	80	41.0/32.0	21.3
STCFormer-L [CVPR2023] [27]	243	×	18.9	78.2	321	40.5/31.8	-
PoseFormerV2 [CVPR23] [39]	243	✓	14.4	4.8	20	45.2/35.6	-
GLA-GCN [ICCV2023] [32]	243	✓	1.3	1.5	6	44.4/34.8	21.0
MotionBERT [ICCV2023] [44]	243	×	42.3	174.8	719	39.2/32.9	17.8
HDFormer [IJCAI2023] [1]	96	×	3.7	0.6	6	42.6/33.1	21.6
MotionAGFormer-L [WACV2024] [20]	243	×	19.0	78.3	322	38.4/32.5	17.3
KTPFormer [CVPR2024] [24]	243	×	35.2	76.1	313	40.1/31.9	19.0
PoseMagic [AAAI2025] [36]	243	×	14.4	20.29	84	37.5/-	-
PoseMamba-S [AAAI2025] [12]	243	×	0.9	3.6	15	41.8/35.0	20.0
PoseMamba-B [AAAI2025] [12]	243	×	3.4	13.9	57	40.8/34.3	16.8
PoseMamba-X [AAAI2025] [12]	243	×	26.5	109.9	452	37.1/31.5	14.8
SAMA-S (Ours)	243	×	1.1	3.9	16	40.6/34.0	20.2
SAMA-B (Ours)	243	×	3.3	11.7	48	37.7/32.0	13.6
SAMA-L (Ours)	243	×	17.3	53.2	219	<u>36.9</u>/<u>31.3</u>	<u>11.9</u>
SAMA-S (Ours)	351	×	1.1	6.3	18	40.2/33.8	19.5
SAMA-B (Ours)	351	×	3.3	18.9	54	37.4/31.7	12.4
SAMA-L (Ours)	351	×	17.3	82.1	234	**36.5/31.0**	**11.4**
vs. prev. SoTA	-	-	***↓11.2***	***↓27.8***	***↓218***	***↓0.6/↓0.5***	***↓3.4***

- **Human3.6M**: We train the model for 80 epochs using the AdamW optimizer with a batch size of 8. We set the sequence length to 351 and 243. The initial learning rate is established at 5e-5 with an exponential learning rate decay schedule, utilizing a decay factor of 0.99. Following previous method [12, 36, 44], we utilize SHNet [30] to extra 2D human poses and ground true input from Human3.6M for fair comparison.
- **MPI-INF-3DHP**: Our model is trained for 90 epochs using the AdamW optimizer and the batch size is set as 16. Following the previous work [12, 36], the sequence length is set as 81. The initial learning rate is established at 5e-4 with an exponential learning rate decay schedule, utilizing a decay factor of 0.99. We employ the 2D ground true pose from MPI-INF-3DHP as input.

Baselines. We compare our method with the state-of-the-art PoseMamba and PoseMagic.

- **PoseMamba**. Utilizing a global-local spatial-temporal SSM block, PoseMamba effectively models human joint correlations, while neglecting the inherent topology and ignores motion differences among joints.
- **PoseMagic**. Leveraging a hybrid Mamba-GCN architecture that explicitly captures the relationships between neighboring joints, PoseMagic incorporates a local enhancement module for structure modeling. Although effective at learning the underlying 3D structure, the approach uniformly treats all joints, thereby overlooking the distinct modeling requirements of joint motion.

4.2. Indoor Monocular 3D Human Pose Estimation

Quantitative comparison. The comparative performance of various methodologies in terms of indoor monocular 3D human pose estimation is systematically listed in Tab. 1. The results unequivocally demonstrate that our proposed method exhibits superior performance, registering an exemplary state-of-the-art MPJPE score of 36.5. In direct comparison with the sota method PoseMamba-X [12] with SAMA-L , our method exhibits a marked enhancement of $0.6mm$ MPJPE reduction. Moreover, our method consistently attains high accuracy results across settings of different sizes: $40.2mm$, $37.4mm$, for different variants SAMA-S / SAMA-B, respectively. Specifically, these variants surpass PoseMamba among models with comparable parameter scales. Furthermore, when aligning the estimated poses, our SAMA-L achieves a P-MPJPE of 31.0, reaching the advanced level. Across different model scales, our approach consistently outperforms PoseMamba. Lastly, with ground truth 2D poses as input, our SAMA-L achieves an MPJPE of $11.4mm$, marking a significant improvement over PoseMamba (11.4 *v.s.* 14.8). We attribute this to the core module of our algorithm, structure-aware state integra-

tor and motion-adaptive state modulator. They aggregate pose topology information and adaptively model the varying motion characteristics of different joints in state space.

Efficiency comparison. To showcase the efficiency of our method, we compare it with others in terms of parameter count and MACs per frame. Especially, our SAMA-B uses only 3.3M parameters (1/2 of PoseMamba-L) and 54M MACs per frame (less than half of PoseMamba-L). On the dataset with SHNet-detected 2D poses, it achieves $0.7mm$ lower prediction error than PoseMamba-L. When using 2D ground truth as input, it surpasses all previous models. Additionally, our SAMA-L achieves significantly lower parameter count and MACs per frame compared to the previous SOTA PoseMamba-X while maintaining superior accuracy, reducing the prediction error by 0.6mm with SHNet-detected 2D poses and 3.4mm with ground truth input. We attribute this to our module's structure-aware joint feature fusion and state fusion, which are based on a lightweight learnable adjacency matrix. Additionally, in MSM, we leverage basic functions to identify joint motion characteristics without introducing excessive computation.

4.3. In-the-wild 3D Human Pose Estimation

To evaluate robustness, we compare our SAMA's performance with other methods in Tab. 2 on MPI-INF-3DHP, which contains in-the-wild scenario. For a fair comparison, we follow the previous works [11, 20, 36] to take the ground true 2D keypoints as input and the sequence length is set as 81. Our SAMA achieves state-of-the-art performance with an MPJPE of 14.4 mm, compared to the previous best method, PoseMamba. Additionally, our method surpasses PoseMagic in terms of AUC and PCK by 0.2% and 0.7%, respectively. These results demonstrate the robustness of our method on the outdoor dataset MPI-INF-3DHP, while maintaining strong performance even with short sequences.

4.4. Ablation Study

We conduct a series of ablation study on Human3.6M [13] to validate the efficacy of our core algorithm designs with our SAMA-B as the base model.

Effect of our main components. To evaluate the impact of our core algorithm, we conducted an analysis by removing the structure-aware state integrator and motion-adaptive state modulator. As presented in Tab. 3, the baseline model, composed of stacked blocks without our proposed components, achieves an MPJPE of 39.9 mm. Incorporating the basic SSD module leads to a 0.6 mm reduction in MPJPE. Built on this setting, SSI yields a performance improvement to 38.4 mm, attributed to its ability to enhance joint correlation modeling via a learnable adjacency matrix in the state space. Besides, MSM further improves performance to 37.4 mm, owing to its capability to adaptively capture motion patterns by controlling the timescale. The results also verify that combining SSI and MSM yields the best results, indicating the effectiveness of considering topology information aggregation in space and different joint motion characteristics in time.

Table 2. Quantitative comparisons on MPI-INF-3DHP dataset. The best performances are **bold**. MPJPE($mm, \downarrow$), PCK($\%, \uparrow$) and AUC($\%, \uparrow$) are reported. T denotes the number of input frames.

Method	T	PCK↑	AUC ↑	MPJPE ↓
Anatomy3D [TCSVT2021] [2]	81	87.8	53.8	79.1
PoseFormer [ICCV2021] [42]	9	88.6	56.4	77.1
MixSTE [CVPR2022] [34]	27	94.4	66.5	54.9
MHFormer [CVPR2022] [16]	9	93.8	63.3	58.0
P-STMO [ECCV2022] [25]	81	97.9	75.8	32.2
GLA-GCN [ICCV2023] [32]	81	98.5	79.1	27.8
STCFormer [CVPR2023] [27]	81	98.7	83.9	23.1
PoseFormerV2 [CVPR2023] [39]	81	97.9	78.8	27.8
MotionAGFormer [WACV2024] [20]	81	98.2	85.3	16.2
KTPFormer [CVPR2024] [24]	81	98.9	85.9	16.7
PoseMagic [AAAI2025] [36]	81	98.8	87.6	14.7
PoseMamba [AAAI2025] [11]	81	-	-	14.5
SAMA (Ours)	81	**99.0**	**88.3**	**14.4**

Generalization evaluation. To evaluate the generalization capability of our approach, we integrate our core discrete joint modeling component into other methods. Specifically, we prepend our SSI and MSM to the networks without modifying the remaining architecture. For a fair comparison, we adopt their default implementation settings, including hyperparameters and augmentation strategies. Tab. 6 presents the comparative results on Human3.6M. As observed, our approach significantly enhances the performance of the baseline estimation networks, achieving reductions of 0.6, 1.2, and 0.9 mm in MPJPE for MixSTE [34], MotionBERT [44], and MotionAGFormer [20], respectively. These consistent performance improvements illustrate the wide potential benefit of our algorithm. Notably, 'MotionAGFormer + `Ours`' achieves an MPJPE of 37.5 mm, which is on par with the advanced methods. This result is particularly impressive, considering the fact that the improvement is solely achieved by integrating our module, without any additional modifications. The success of our approach can be attributed to the fact that our algorithm not only effectively complements the topological connections between joints but also takes into account the distinct motion characteristics of different joints, further enhancing overall performance.

Comparison with various spatial learning methods. To demonstrate the effectiveness of our SSI, we replaced the spatial dependency learning part of our model with the previous methods, bi-directional scanning method in PoseMagic and global-local scanning method in PoseMamba. The bi-directional scanning method sequentially processes joint indices in both descending and ascending

Table 3. Ablation of the main components in our method.

Vanilla SSD	SSI	MSM	MPJPE
-	-	-	39.9
✓	-	-	39.3
✓	✓	-	38.4
✓	-	✓	38.5
✓	✓	✓	**37.4**

Table 4. Comparison with other spatial scanning methods.

Spatial Learning	MPJPE	MACs
bi-direction [36]	38.2	58.12
global-local [11]	37.9	58.12
vanilla + SSI (Ours)	**37.4**	53.95

Table 5. Effect of different motion detection function.

Motion Learning	MPJPE
Baseline	38.4
Linear	38.0
Point-wise Conv	**37.4**

Table 6. Generalization of our algorithm.

Method	MPJPE
MixSTE[CVPR2022] [34]	40.9
MixSTE + **Ours**	40.3 ↓*0.6*
MotionBERT[ICCV2023] [44]	39.2
MotionBERT + **Ours**	38.0 ↓*1.2*
MotionAGFormer[WACV2024] [20]	38.4
MotionAGFormer + **Ours**	37.5 ↓*0.9*

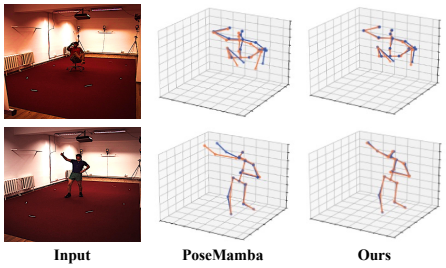

Figure 4. Visual comparable results of estimated 3D poses between PoseMamba and ours.

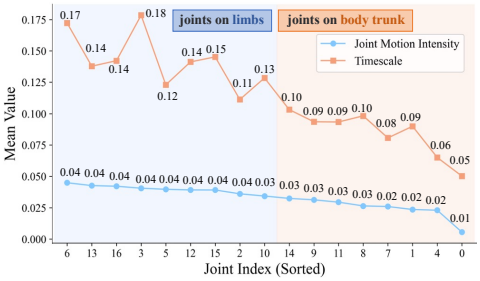

Figure 5. Statistical motion intensity and timescale Δ results across different joints.

orders, thereby neglecting the intrinsic connectivity among joints. Besides, the global-local strategy employs a predefined local motion-specific scanning pattern, which yields only marginal performance gains at the expense of considerable computational cost. As shown in Tab. 4, our approach, which integrates a simple vanilla scanning method with the SSI, achieves the best MPJPE result of 37.4 mm with lower computational cost, demonstrating greater efficiency compared to the more complex scanning strategies. This result underscores the capability of our SSI in effectively capturing dynamic spatial joint dependencies.

Effect of motion-adaptive state modulator. We visualize the effect of motion-adaptive state modulator in Fig. 5. MSM leverages the motion characteristics between adjacent frames to learn a timescale that dynamically balances the influence of the previous state and the current input for the current frame's output, thereby capturing richer joint motion features. As shown in the figure, joints on limbs (*e.g.*, joint 3, 6, 13, 16, 5 and 12), which exhibit greater average motion intensity, correspond to larger timescales, while joints on the body trunk (*e.g.*, joint 0, 1, 4, 7 and 8), which move less, correspond to smaller timescales. This correlation between motion intensity and timescale confirms the fundamental rationale behind our design of motion-adaptive state modulator. Specifically, our model leverages motion information so that larger motion amplitudes correspond to larger timescales. This allows the model to reduce reliance on the previous state when encountering intense motion, preventing it from erroneously smoothing the motion trajectory in such cases.

Effect of motion capture method. We explore two simple functions to capture motion cues between adjacent joints to regulate the timescale, using SAMA-B without motion capturing as the baseline, as shown in Tab. 5. Point-wise convolution (1D conv, kernel size 2) captures local motion patterns, enabling dynamic timescale adjustments. A simple linear layer preserves complete adjacent joint features, enhancing joint dependency modeling. Both methods use zero padding on the left and improve performance, demonstrating the effectiveness of joint-specific motion information in regulating timescales. In practical applications, we adopt point-wise convolution for implementation.

Visualization of estimated poses. Fig. 4 illustrates the 3D pose predictions of PoseMamba and our method, where blue / orange denotes the ground truth / estimated poses, respectively. It reveals that the estimated poses generated by our approach demonstrate superior accuracy compared to those of PoseMamba, particularly in the highly dynamic limb regions. This highlights the effectiveness of our joint-specific modeling strategy, enabling more precise motion capture and consequently enhancing overall performance.

5. Conclusion

In this work, we introduce a new algorithm tailored for lifting-based pose estimation. Our algorithm incorporates a structure-aware and motion-adaptive strategy, facilitating dynamic joint connection modeling and personalized motion adaptation, enabling more precise motion trajectory reconstruction while preserving intrinsic motion characteristics, thereby ensuring enhanced representation of joint dependencies. Experimental evaluations on comprehensive benchmarks manifest its superiority in accuracy and efficiency with reduced computational cost.

References

[1] 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.

[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] Xipeng Chen, Pengxu Wei, and Liang Lin. Deductive learning for weakly-supervised 3d human pose estimation via uncalibrated cameras. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 1089–1096. AAAI Press, 2021.

[4] Yujin Chen, Zhigang Tu, Liuhao Ge, Dejun Zhang, Ruizhi Chen, and Junsong Yuan. So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 6960–6969. IEEE, 2019.

[5] Yujin Chen, Zhigang Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and Junsong Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. IEEE Trans. Image Process., 30:4008–4021, 2021.

[6] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net, 2024.

[7] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. In Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024.

[8] Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023, pages 2902–2912. IEEE, 2023.

[9] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. CoRR, abs/2312.00752, 2023.

[10] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. Advances in neural information processing systems, 34:572–585, 2021.

[11] Yunlong Huang, Junshuo Liu, Ke Xian, and Robert Caiming Qiu. Posemamba: Monocular 3d human pose estimation with bidirectional global-local spatio-temporal state space model. CoRR, abs/2408.03540, 2024.

[12] Yunlong Huang, Junshuo Liu, Ke Xian, and Robert Caiming Qiu. Posemamba: Monocular 3d human pose estimation with bidirectional global-local spatio-temporal state space model. arXiv preprint arXiv:2408.03540, 2024.

[13] 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.

[14] Han Li, Bowen Shi, Wenrui Dai, Hongwei Zheng, Botao Wang, Yu Sun, Min Guo, Chenglin Li, Junni Zou, and Hongkai Xiong. Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In Proceedings of the AAAI conference on artificial intelligence, 2023.

[15] Sijin Li, Weichen Zhang, and Antoni B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pages 2848–2856. IEEE Computer Society, 2015.

[16] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 13137–13146. IEEE, 2022.

[17] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Trans. Multim., 25:1282–1293, 2023.

[18] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. arXiv preprint arXiv:2403.19887, 2024.

[19] Ye Lu, Jianjun Gao, Chen Cai, Ruoyu Wang, Duc Tri Phan, and Kim-Hui Yap. Hdplifter: Hierarchical dynamics perception for 2d-to-3d human pose lifting. In 2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024.

[20] Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024, pages 6905–6915. IEEE, 2024.

[21] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In 2017 international conference on 3D vision (3DV), pages 506–516. IEEE, 2017.

[22] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and

mesh estimation from a single RGB image. In Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII, pages 752–768. Springer, 2020.

[23] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 1263–1272. IEEE Computer Society, 2017.

[24] Jihua Peng, Yanghong Zhou, and P. Y. Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pages 1123–1132. IEEE, 2024.

[25] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-STMO: pre-trained spatial temporal many-to-one model for 3d human pose estimation. In Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V, pages 461–478. Springer, 2022.

[26] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI, pages 536–553. Springer, 2018.

[27] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pages 4790–4799. IEEE, 2023.

[28] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct prediction of 3d body poses from motion compensated sequences. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 991–1000. IEEE Computer Society, 2016.

[29] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 11179–11188. IEEE, 2021.

[30] Tianhan Xu and Wataru Takano. Graph stacked hourglass networks for 3d human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16105–16114, 2021.

[31] Mang Ye, He Li, Bo Du, Jianbing Shen, Ling Shao, and Steven C. H. Hoi. Collaborative refining for person re-identification with label noise. IEEE Trans. Image Process., 31:379–391, 2022.

[32] Bruce X. B. Yu, Zhi Zhang, Yongxu Liu, Sheng-Hua Zhong, Yan Liu, and Chang Wen Chen. GLA-GCN: global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pages 8784–8795. IEEE, 2023.

[33] Can Zhang, Tianyu Yang, Junwu Weng, Meng Cao, Jue Wang, and Yuexian Zou. Unsupervised pre-training for temporal action localization tasks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 14011–14021. IEEE, 2022.

[34] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 13222–13232. IEEE, 2022.

[35] Jiaxu Zhang, Gaoxiang Ye, Zhigang Tu, Yongtao Qin, Qianqing Qin, Jinlu Zhang, and Jun Liu. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. CAAI Trans. Intell. Technol., 7(1):46–55, 2022.

[36] Xinyi Zhang, Qiqi Bao, Qinpeng Cui, Wenming Yang, and Qingmin Liao. Pose magic: Efficient and temporally consistent human pose estimation with a hybrid mamba-gcn network. CoRR, abs/2408.02922, 2024.

[37] Zeyu Zhang, Akide Liu, Ian D. Reid, Richard I. Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part I, pages 265–282. Springer, 2024.

[38] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 3425–3435. Computer Vision Foundation / IEEE, 2019.

[39] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pages 8877–8886. IEEE, 2023.

[40] Weixi Zhao, Weiqiang Wang, and Yunjie Tian. Graformer: Graph-oriented transformer for 3d pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 20406–20415. IEEE, 2022.

[41] Ce Zheng, Sijie Zhu, Matías Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 11636–11645. IEEE, 2021.

[42] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 11656–11665, 2021.

[43] Hongwei Zheng, Han Li, Wenrui Dai, Ziyang Zheng, Chenglin Li, Junni Zou, and Hongkai Xiong. Hipart: Hier-

archical pose autoregressive transformer for occluded 3d human pose estimation. In Proceedings of the Computer Vision and Pattern Recognition Conference, 2025.

[44] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pages 15039–15053. IEEE, 2023.