PoseMamba: Monocular 3D Human Pose Estimation with Bidirectional Global-Local Spatio-Temporal State Space Model

Yunlong Huang, Junshuo Liu, Ke Xian*, Robert Caiming Qiu

Huazhong University of Science and Technology, Wuhan, China
{huangyunlong, junshuo_liu, kxian, caiming}@hust.edu.cn

Abstract

Transformers have significantly advanced the field of 3D human pose estimation (HPE). However, existing transformer-based methods primarily use self-attention mechanisms for spatio-temporal modeling, leading to a quadratic complexity, unidirectional modeling of spatio-temporal relationships, and insufficient learning of spatial-temporal correlations. Recently, the Mamba architecture, utilizing the state space model (SSM), has exhibited superior long-range modeling capabilities in a variety of vision tasks with linear complexity. In this paper, we propose PoseMamba, a novel purely SSM-based approach with linear complexity for 3D human pose estimation in monocular video. Specifically, we propose a bidirectional global-local spatio-temporal SSM block that comprehensively models human joint relations within individual frames as well as temporal correlations across frames. Within this bidirectional global-local spatio-temporal SSM block, we introduce a reordering strategy to enhance the local modeling capability of the SSM. This strategy provides a more logical geometric scanning order and integrates it with the global SSM, resulting in a combined global-local spatial scan. We have quantitatively and qualitatively evaluated our approach using two benchmark datasets: Human3.6M and MPI-INF-3DHP. Extensive experiments demonstrate that PoseMamba achieves state-of-the-art performance on both datasets while maintaining a smaller model size and reducing computational costs.

Introduction

3D human pose estimation from monocular observations is a fundamental task in computer vision with various real-world applications (Mehta et al. 2017b; Wiederer et al. 2020; Czech et al. 2022; Bauer et al. 2023; Munea et al. 2020). Typically, this involves two separate steps: 2D pose detection to locate keypoints on the image plane, followed by 2D-to-3D lifting to determine joint positions in 3D space from 2D keypoints. Recovering accurate 3D pose from 2D keypoints is challenging due to depth ambiguity and self-occlusion in monocular data. To address these challenges, significant advancements in deep learning approaches have been made, consistently improving performance (Liu et al. 2020; Chen et al. 2020; Zeng et al. 2020; Wang et al. 2020).

*Corresponding author

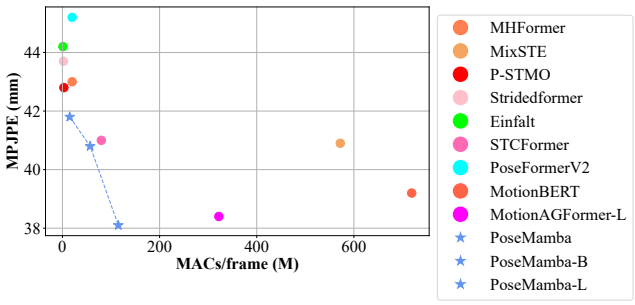


Figure 1: Comparisons of recent 3D human pose estimation techniques on Human3.6M (Ionescu et al. 2013) (lower is better). MACs/frame represents multiply-accumulate operations for each output frame. Our PoseMamba method presents various versions and achieves superior results, while maintaining computational efficiency.

Recently, transformers (Vaswani et al. 2017) have demonstrated significant potential in 3D human pose estimation. Its self-attention mechanism enables it to efficiently capture spatio-temporal relationships for this domain. For example, PoseFormer (Zheng et al. 2021) leverages spatio-temporal information to estimate more accurate central-frame pose in video sequence. MHFormer (Li et al. 2022b) learns spatio-temporal representations of multiple pose hypotheses in an end-to-end manner. MixSTE (Zhang et al. 2022) proposes an alternating design using a transformer-based seq2seq model to capture the coherence between sequences. However, applying full attention mechanisms to long 2D keypoints sequence results in a notable rise in computational requirements, due to the quadratic complexity of attention calculations in both computation and memory. This naturally raises the question: *how can a method be designed to function with linear complexity while still preserving the advantages of capturing spatio-temporal information?*

We observe recent progress in state space models (Gu and Dao 2023; Wang et al. 2023; Islam and Bertasius 2022), particularly with the emergence of the structured state space sequence model (S4) (Gu, Goel, and Ré 2021) as a promising architecture for sequence modeling. Building upon S4, Mamba (Gu and Dao 2023) incorporates time-varying parameters into the SSM, introducing an efficient hardware-

aware algorithm with global receptive fields and linear complexity. Recently, a few concurrent approaches (Zhu et al. 2024; Liu et al. 2024) have focused on 2D vision tasks, such as classification and segmentation.

Driven by the successes of SSM in 2D image processing, we propose Pose State Space Model, **PoseMamba**, featuring bidirectional global-local spatial-temporal modeling with linear complexity for 3D human pose estimation. Our pilot tests indicate that relying solely on Mamba (Gu and Dao 2023) may not yield optimal performance, likely due to its unidirectional modeling approach. To address this, we introduce a bidirectional global-local spatial-temporal modeling framework, where global modeling captures the full-body pose, and local modeling focuses on limbs and their movements. We enhance local modeling with a reordering strategy that integrates with global SSM, resulting in a combined spatial scan. Experimental results on Human3.6M and MPI-INF-3DHP demonstrate that PoseMamba outperforms previous state-of-the-art (SOTA) methods while using fewer parameters and MACs, highlighting the potential of SSM in 3D human pose estimation, as shown in Figure 1.

Our main contributions are:

- We introduce a novel bidirectional global-local spatio-temporal modeling approach and logical geometric scanning strategy within the Mamba framework for 3D HPE.
- PoseMamba enables effective learning of global-local spatial-temporal information with linear complexity, leveraging human skeleton geometry.
- *Efficiency and Flexibility:* **i)** PoseMamba is lightweight and faster, being 2.8× faster than MotionAGFormer and reducing GPU memory usage by 64.7% during batch inference for 3D pose estimation. **ii)** Various versions of PoseMamba are available to balance accuracy and speed based on user needs.
- Our PoseMamba model achieves state-of-the-art results on both Human3.6M and MPI-INF-3DHP datasets without unnecessary complexity.

Code — https://github.com/nankingjing/PoseMamba

Related Work

3D Human Pose Estimation

Existing methods can be categorized based on input type: multi-view and monocular approaches. Multi-view methods (Zhang et al. 2021; Reddy et al. 2021; Chun, Park, and Chang 2023) require multiple cameras, posing practical challenges. They can also be classified into direct 3D HPE methods (Pavlakos, Zhou, and Daniilidis 2018; Sun et al. 2018; Zhou et al. 2019; Huang et al. 2023) and 2D-3D lifting methods, which use 2D pose detectors (Chen et al. 2018; Sun et al. 2019; Newell, Yang, and Deng 2016) to elevate 2D coordinates to 3D (Zhao et al. 2023; Zhu et al. 2023; Zhang et al. 2022). Existing works (Holmquist and Wandt 2023; Shan et al. 2023) use multi-hypothesis approach to improve depth ambiguity in 3D HPE. DSED (Liu et al. 2022) addresses the self-occlusion problem in 3D HPE by explicitly reasoning about occlusion relationships in multi-person scenarios. While HumMUSS (Mondal, Alletto, and Tome 2024) first explores bidirectional SSM modeling for human motion understanding, our work introduces a novel bidirectional global-local spatio-temporal approach and logical geometric scanning strategy tailored for 3D HPE. While PoseMagic (Zhang et al. 2024) introduces a hybrid Mamba-GCN architecture, its heavy reliance on GCN for capturing local details can be limiting. In contrast, our PoseMamba captures local movement details more effectively through its bidirectional global-local spatio-temporal modeling method and logical geometric scanning strategy.

State Space Models

Mamba (Gu and Dao 2023) achieved breakthroughs with linear-time inference and efficient training. MoE-Mamba (Pióro et al. 2024) combined Mixture of Experts with Mamba, enhancing scalability. Vision Mamba (Zhu et al. 2024) and VMamba (Liu et al. 2024) employed bidirectional SSM blocks and cross-scan modules, respectively, but the potential of Mamba in 3D human pose estimation remains unexplored. Our approach compares unidirectional with bidirectional scanning, addressing inaccuracies in limb recognition and enhancing spatial scanning for 3D human pose estimation.

Preliminaries

State Space Model We can think of SSM as linear time-invariant system that maps input $x(t) \in \mathbb{R}^L$ to output $y(t) \in \mathbb{R}^L$ via hidden state $h(t) \in \mathbb{C}^N$. It can be described as linear ordinary differential equations:

$$\begin{aligned} \dot{h}(t) &= \boldsymbol{A}h(t) + \boldsymbol{B}x(t) \\ y(t) &= \boldsymbol{C}h(t) + Dx(t) \end{aligned} \tag{1}$$

Here, $\dot{h}(t)$ is the time derivative of $h(t)$, and $\boldsymbol{A} \in \mathbb{C}^{N\times N}$, $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{C}^N$, and $D \in \mathbb{C}^1$ are the weighting parameters.

Discretization of SSM To handle discrete sequence inputs, continuous-time SSMs need to be discretized. The analytical solution to Equation (1) is given by:

$$h(t_b) = e^{\boldsymbol{A}(t_b - t_a)}(h(t_a) + \int_{t_a}^{t_b} \boldsymbol{B}(\tau)x(\tau)e^{-\boldsymbol{A}(\tau - t_a)}\, d\tau) \tag{2}$$

Subsequently, through sampling with step size $\boldsymbol{\Delta}$ (i.e., $d\tau|_{t_i}^{t_{i+1}} = \Delta_i$), $h(t_b)$ can be discretized as:

$$h_b = e^{\boldsymbol{A}(\sum_{i=a}^{b-1}\Delta i)}\left(h_a + \sum_{i=a}^{b-1} \boldsymbol{B}_i x_i e^{-\boldsymbol{A}(\sum_{j=a}^{i}\Delta j)}\Delta_i\right) \tag{3}$$

This approach utilizes the zero-order hold technique (Gu and Dao 2023). For $b = a + 1$, Equation (3) simplifies to:

$$h_{a+1} = \overline{\boldsymbol{A_a}}h_a + \overline{\boldsymbol{B_a}}x_a \tag{4}$$

Here, $\overline{\boldsymbol{A_a}} = e^{\boldsymbol{A}\Delta_a}$ corresponds to the ZOH discretization result (Gu and Dao 2023), while $\overline{\boldsymbol{B_a}} = \boldsymbol{B}_a\Delta_a$ essentially represents the first-order Taylor expansion of the ZOH-derived equivalent.

Selective Scan The weight matrix $\boldsymbol{B}$ in Equation (2) and Equation (3), along with $\boldsymbol{C}$, $\boldsymbol{D}$, and $\boldsymbol{\Delta}$, is tailored to be input-dependent to overcome the limitations of SSMs (Equation (1)) in capturing contextual details (Gu and Dao 2023). However, the introduction of time-varying SSMs presents a computational challenge because convolutions with dynamic weights are not supported, making them unsuitable for this purpose. Nonetheless, deriving the recurrence relation of h_b in Equation (3) enables efficient computation. Specifically, if we define $e^{\boldsymbol{A}(\Delta_a+\ldots+\Delta_{i-1})}$ as $\boldsymbol{p^i_{A,a}}$, its recurrence relation can be expressed as

$$\boldsymbol{p^i_{A,a}} = e^{\boldsymbol{A}\Delta_{i-1}}\boldsymbol{p^{i-1}_{A,a}} \tag{5}$$

Regarding the second term of Equation (3), we obtain

$$\boldsymbol{p^b_{B,a}} = e^{\boldsymbol{A}(\Delta_a+\ldots+\Delta_{b-1})}\sum_{i=a}^{b-1}\boldsymbol{B_i}x_i e^{-\boldsymbol{A}(\Delta_a+\ldots+\Delta_i)}\Delta_i \tag{6}$$

Therefore, utilizing the relationships derived in Equation (5) and Equation (6), the computation of $h_b = \boldsymbol{p^b_{A,a}}h_a + \boldsymbol{p^b_{B,a}}$ can be efficiently parallelized using associative scan algorithms (Martin and Cundy 2017; Smith, Warrington, and Linderman 2022), which are facilitated by various contemporary programming libraries.

PoseMamba

As illustrated in Figure 2, our network processes a concatenated 2D coordinate array $C_{T,J} \in \mathbb{R}^{T\times J\times 2}$ representing J joints across T frames. The input has a channel size of 2.

Initially, we project the input keypoint sequence $C_{T,J}$ into a high-dimensional feature $P_{T,J} \in \mathbb{R}^{T\times J\times d_m}$ with each joint represented by a feature dimension of d_m. Subsequently, we incorporate a spatial and a temporal position embedding matrix to preserve positional details across spatial and temporal domains. The proposed PoseMamba takes $P_{T,J}$ as input and focuses on capturing global bidirectional spatial-temporal information efficiently through Mamba blocks with linear complexity. Lastly, we employ a regression head to combine the encoder's outputs $Z \in \mathbb{R}^{T\times J\times d_m}$, adjusting the dimension from d_m to 3 to derive the 3D human pose sequence $Out \in \mathbb{R}^{T\times J\times 3}$.

Spatio-Temporal Encoder

Transformer-Based Spatio-Temporal Correlation Learning Prior transformer-based studies have primarily concentrated on utilizing multi-head self-attention mechanisms to understand spatio-temporal relationships, as illustrated in Fig. 3(a). The computation of attention for the query, key, and value matrices Q, K, V in each head is expressed as:

$$Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_m}})V, \tag{7}$$

where $\{Q,K,V\} \in \mathbb{R}^{O\times d_m}$, O indicates the number of tokens, and d_m is the dimension of each token.

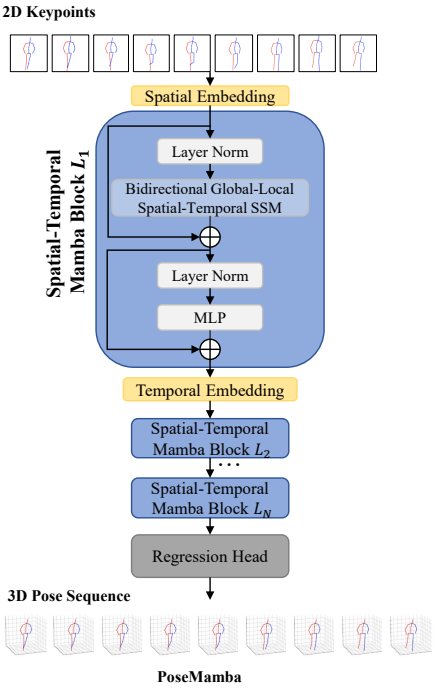
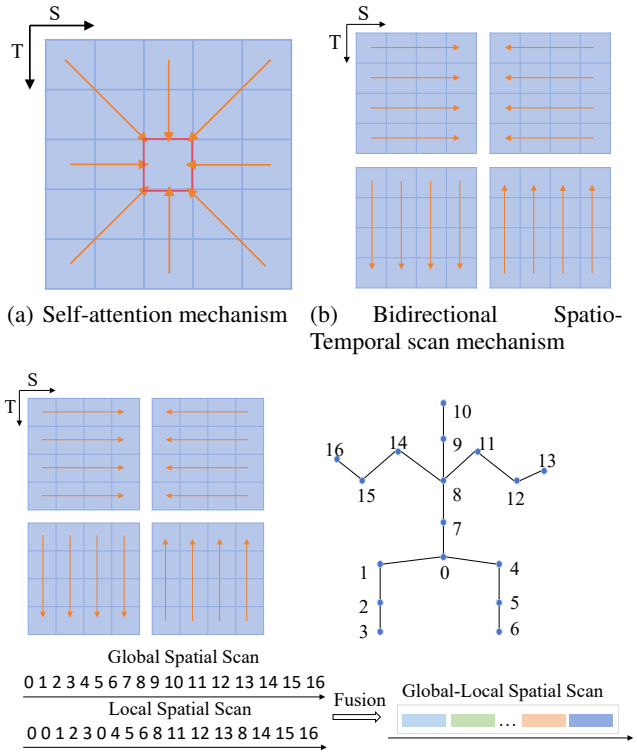


Figure 2: The pipeline of our PoseMamba. We start by using fully connected layer to project the input keypoint sequence, and then embed position and temporal embedding matrix into sequence. After that, we feed the sequence into the Mamba blocks.

Bidirectional Global-Local Spatio-Temporal Modeling In contrast to prior methods using attention mechanisms with quadratic computational complexity, we propose a state space model to encapsulate comprehensive spatio-temporal information at a linear complexity. Specifically, inspired by VMamba (Liu et al. 2024), before inputting the tokens into the S6 model, we reorganize the tokens in both spatial and temporal dimensions, specifically forward spatial scan, forward temporal scan, backward spatial scan, and backward temporal scan, as depicted in Figure 3(b). Subsequently, the resultant features are merged. This approach enables the model to obtain comprehensive bidirectional global spatio-temporal information from bidirectional spatial and temporal dimensions. Furthermore, the computational complexity remains at linear complexity in contrast to the self-attention operation with quadratic complexity in transformer Figure 3(a). To better demonstrate the benefits of bidirectional spatio-temporal modeling, we conduct experiments on four unidirectional spatio-temporal scan mechanisms, as depicted in Figure 4, which demonstrates that relying solely on Mamba can not achieve optimal performance.

Furthermore, to address the persistent challenge of inaccurate limb prediction, we introduce a novel reordering strategy designed to augment the local modeling capabilities of the state space model. This enhancement is achieved by establishing a more rational geometric scanning sequence, which is then seamlessly integrated with the



(a) Self-attention mechanism (b) Bidirectional Spatio-Temporal scan mechanism

(c) Bidirectional Global-Local Spatio-Temporal scan mechanism

Figure 3: Illustration of various spatio-temporal modeling mechanisms. (a) Self-attention (Vaswani et al. 2017; Dosovitskiy et al. 2020). (b) Bidirectional spatio-temporal scan (Liu et al. 2024). (c) Our proposed bidirectional global-local spatio-temporal scan mechanism, which leverages the geometry of the human skeleton to enhance detail.

global SSM framework. This integration facilitates a comprehensive global-local spatial scanning approach, as illustrated in Figure 3(c). Our proposed strategy not only refines the spatial scanning process but also ensures a harmonious fusion of local details with the broader spatial context, thereby significantly improving the precision of limb predictions. Specifically, we posit that scanning key points on the human skeleton from 0 to 16 enables the extraction of global spatial features. However, our experimental findings indicate that relying only on global scanning consistently led to inaccurate limb prediction. Therefore, exploiting the interactions between body joints, we propose a local scanning approach to capture local human skeleton details, as detailed in Figure 3(c). We design a global-local spatial scanning approach by merging these two scanning sequences. Additionally, by incorporating temporal scanning, we develop a bidirectional global-local spatio-temporal mamba block, advancing the modeling of spatio-temporal features for 3D HPE.

Bidirectional Global-Local Spatio-Temporal Mamba Block For each spatio-temporal Mamba block, layer normalization (LN), bidirectional spatio-temporal SSM, depth-wise convolution (Chollet 2017), and residual connections

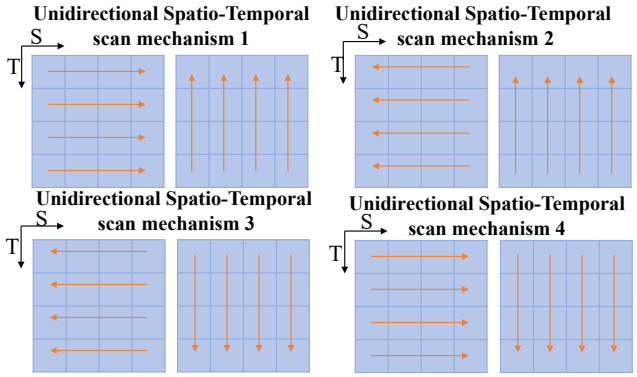


Figure 4: Illustration of different unidirectional spatio-temporal scan mechanisms.

are employed. A spatio-temporal Mamba block is shown in Figure 2, and the output can be summarized as follows:

$$\begin{aligned} Z'_l &= LN(SSM(\sigma(DW(LN(Z_{l-1}))))) + Z_{l-1}, \\ Z_l &= MLP(LN(Z'_l)) + Z'_l, \end{aligned} \tag{8}$$

where $Z_l \in \mathbb{R}^{T\times J\times C}$ is the output of the l-th block. *DW* means the depth-wise convolution. Following the DW, a SiLU (Hendrycks and Gimpel 2016) and SSM are adopted.

Spatio-Temporal Correlation Learning We employ the bidirectional global-local spatio-temporal Mamba blocks to learn spatio-temporal correlations among joints in over frames. Firstly, we take 2D keypoints sequence as input $C_{T,J} \in \mathbb{R}^{T\times J\times 2}$ and project each keypoint to a high-dimensional feature $P_{T,J} \in \mathbb{R}^{T\times J\times d_m}$ with the linear embedding layer. We then embed the spatial position information with a positional matrix $E_{spos} \in \mathbb{R}^{J\times d_m}$. Each joint token $p \in P_J$ is projected from joint c_i of the 2D coordinates $C_J \in \mathbb{R}^{J\times 2}$:

$$X = Norm(L_e(c_i) + E_{spos}),\ X \in \mathbb{R}^{J\times d_m}, \tag{9}$$

where $Norm$ denotes the layer normalization, and L_e indicates the linear embedding layer.

Subsequently, the features are fed into a bidirectional spatio-temporal Mamba block to model dependencies across all joints. We also embed the temporal position information with a temporal positional matrix $E_{tpos} \in \mathbb{R}^{T\times d_m}$:

$$X = Norm(X + E_{tpos}),\ X \in \mathbb{R}^{T\times d_m}, \tag{10}$$

where $Norm$ denotes the layer normalization.

Then, it is fed into spatio-temporal Mamba block to model dependencies across all joints. Finally, we obtain spatio-temporal features through $N-2$ layers of bidirectional spatio-temporal mamba blocks. In the regression head, a linear layer is applied on the output Z to perform regression to produce the 3D pose sequence $Out \in \mathbb{R}^{T\times J\times 3}$.

Loss Function

Following the previous work (Zhu et al. 2023; Zhang et al. 2022), the network is trained in an end-to-end manner and

the final loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{3D} + \lambda_t \mathcal{L}_t + \lambda_m \mathcal{L}_m + \lambda_{2D} \mathcal{L}_{2D}, \tag{11}$$

where $\mathcal{L}_{3D}$ is the MPJPE loss, $\mathcal{L}_t$ is the TCLoss (Hossain and Little 2018) to generate smooth poses, $\mathcal{L}_m$ denotes the MPJVE loss (Pavllo et al. 2019) to improve the temporal coherence, and $\mathcal{L}_{2D}$ denotes the 2D re-projection loss (Zhu et al. 2023). During the training stage, different coefficients λ_t and λ_m are employed to $\mathcal{L}_t$ and $\mathcal{L}_m$ to avoid excessive smoothness in sequence. We merge the TCLoss and MPJVE as the temporal loss function (T-Loss) inspired by the previous work (Zhang et al. 2022). The MPJPE loss L_{3D} is computed as follows:

$$\mathcal{L}_{3D} = \sum_{t=1}^{T} \sum_{i=1}^{J} \left\| Y_i^t - \widetilde{X}_i^t \right\|_2, \tag{12}$$

where $\widetilde{X}_i^t$ and Y_i^t represent the predicted and ground truth 3D poses of joint i at frame t, respectively.

Experiment

We evaluate our proposed PoseMamba on two large-scale 3D human pose estimation datasets, i.e., Human3.6M (Ionescu et al. 2013) and MPI-INF-3DHP (Mehta et al. 2017a).

Datasets and Evaluation Metrics

Human3.6M is a commonly used indoor dataset for 3D human pose estimation. It contains 3.6 million video frames of 11 subjects performing 15 different daily activities. To ensure fair evaluation, we follow the standard approach and train the model using data from subjects 1, 5, 6, 7, and 8, and then test it on data from subjects 9 and 11. Following the previous work (Zhu et al. 2023), we use two protocols for evaluation. The first protocol (referred to as P1) uses Mean Per Joint Position Error (MPJPE) in millimeters between the estimated pose and the actual pose, after aligning their root joints (sacrum). The second protocol (referred to as P2) measures Procrustes-MPJPE, where the actual pose and the estimated pose are aligned through a rigid transformation. **MPI-INF-3DHP** is another large-scale dataset gathered in three different settings: green screen, non-green screen, and outdoor environments. This dataset has 1.3 million frames, containing a wider range of movements than Human3.6M. We utilize MPJPE as the evaluation metric.

Implementation Details

Model Variants We create three model configurations, detailed in Table 2. Our base model, PoseMamba-B, balances accuracy and computational cost. Other variants are named based on parameters and computational needs. The selection of each variant depends on specific application needs, like real-time processing or precise estimations. The MLP's expansion layer is $\alpha = 2$ for all experiments.

Experimental settings Our model is developed utilizing PyTorch and deployed on one NVIDIA RTX 3090 GPU. Horizontal flipping augmentation is applied for both training and testing, as outlined in (Zhu et al. 2023; Zhao et al. 2023). During model training, the batch size is configured with 4 sequences. The optimization of network parameters is carried out using the AdamW (Loshchilov and Hutter 2017) optimizer across 120 epochs with a weight decay of 0.01. The initial learning rate is established at $2e^{-4}$ with an exponential learning rate decay schedule, utilizing a decay factor of 0.99. In our approach, we leverage the Stacked Hourglass (Newell, Yang, and Deng 2016) 2D pose detection outcomes and 2D ground truths sourced from the Human3.6M and MPI-INF-3DHP datasets, following MotionBERT (Zhu et al. 2023). In MPI-INF-3DHP, we employ ground truth 2D detection using a methodology following methods (Zhao et al. 2023; Tang et al. 2023).

Performance Comparison on Human3.6M

We present a comparative analysis of our PoseMamba model against other models using the Human3.6M dataset. To ensure a fair assessment, only the outcomes of models without additional pre-training on supplementary data are considered. The results, as detailed in Table 1, reveal that PoseMamba-L achieves a P1 error of 38.1 mm for estimated 2D pose and 15.6 mm for ground truth 2D pose. Notably, these results are accomplished with only 16% of the computational resources in comparison to the previous SOTA model, MotionBERT, while exhibiting an enhanced accuracy of 1.1 mm and 2.2 mm, respectively. Furthermore, our model achieves these results using only 36% of the computational resource compared to another previous SOTA model, MotionAGFormer (Mehraban, Adeli, and Taati 2024), while being 0.3 mm and 1.7 mm more accurate, respectively.

Performance Comparison on MPI-INF-3DHP

When assessing our approach to the MPI-INF-3DHP dataset, we adapted our small and base models to accommodate 27 and 81 frames to suit the shorter video sequences. Our method demonstrates superior performance across all model variants compared to others in terms of MPJPE, as illustrated in Table 3, showcasing the excellence of our model.

Ablation Studies

In this section, we evaluate the effectiveness of components.

Bidirectional Global-Local Spatio-Temporal Modeling We perform comprehensive experiments to verify the effectiveness of modifying the crucial bidirectional global-local spatio-temporal modeling in PoseMamba on Human3.6M using our small variant version, where feature dimensions are altered to ensure comparable architectural parameters and MACs for a fair evaluation. As shown in Table 4, employing unidirectional spatio-temporal modeling results in a model performance of MPJPE ranging from 43.0 to 43.8 mm, which is comparatively less efficient than the bidirectional spatio-temporal modeling yielding an MPJPE of 42.4 mm. Furthermore, integrated with the local spatial scan to enhance accurate limb prediction, our final model is 0.6 mm better than bidirectional spatial-temporal modeling, which indicates the efficacy of our global-local modeling.

Method	T	CE	Param	MACs	MACs/frame	P1↓ /P2↓	P1†↓
*MHFormer (Li et al. 2022b) CVPR'22	351	✓	30.9 M	7.0 G	20 M	43.0/34.4	30.5
MixSTE (Zhang et al. 2022) CVPR'22	243	×	33.6 M	139.0 G	572 M	40.9/32.6	21.6
P-STMO (Shan et al. 2022) ECCV'22	243	✓	6.2 M	0.7 G	3 M	42.8/34.4	29.3
Stridedformer (Li et al. 2022a) TMM'22	351	✓	4.0 M	0.8 G	2 M	43.7/35.2	28.5
Einfalt *et al.* (Einfalt, Ludwig, and Lienhart 2023) WACV'23	351	✓	10.4 M	0.5 G	1 M	44.2/35.7	-
STCFormer (Tang et al. 2023) CVPR'23	243	×	4.7 M	19.6 G	80 M	41.0/32.0	21.3
STCFormer-L (Tang et al. 2023) CVPR'23	243	×	18.9 M	78.2 G	321 M	40.5/<u>31.8</u>	-
PoseFormerV2 (Zhao et al. 2023) CVPR'23	243	✓	14.4 M	4.8 G	20 M	45.2/35.6	-
GLA-GCN (Yu et al. 2023) ICCV'23	243	✓	1.3 M	1.5 G	6 M	44.4/34.8	21.0
MotionBERT (Zhu et al. 2023) ICCV'23	243	×	42.3 M	174.8 G	719 M	39.2/32.9	17.8
HDFormer (Chen et al. 2023) IJCAI'23	96	×	3.7 M	0.6 G	6 M	42.6/33.1	21.6
HSTFormer (Qian et al. 2023) arXiv'23	81	×	22.7 M	1.0 G	12 M	42.7/33.7	27.8
DC-GCT (Kang et al. 2023) arXiv'23	81	✓	3.1 M	41 M	41 M	44.7/-	-
MotionAGFormer-L (Mehraban, Adeli, and Taati 2024) WACV'24	243	×	19.0 M	78.3 G	322 M	38.4/32.5	17.3
PoseMamba-S	243	×	0.9 M	3.6 G	15 M	41.8/35.0	20.0
PoseMamba-B	243	×	3.4 M	13.9 G	57 M	40.8/34.3	16.8
PoseMamba-L	243	×	6.7 M	27.9 G	115 M	<u>38.1</u>/32.5	<u>15.6</u>
PoseMamba-X	243	×	26.5 M	109.9 G	452 M	**37.1/31.5**	**14.8**

Table 1: Quantitative comparisons on Human3.6M. T: Number of input frames. CE: Estimating center frame only. P1: MPJPE error (mm). P2: P-MPJPE error (mm). P1†: P1 error on 2D ground truth. (*) denotes using HRNet (Sun et al. 2019) for 2D pose estimation. The best and second-best scores are in bold and underlined, respectively.

Method	N	d_m	T	Params	MACs
PoseMamba-S	20	64	243	0.860 M	3.587 G
PoseMamba-B	20	128	243	3.358 M	13.943 G
PoseMamba-L	40	128	243	6.714 M	27.881 G
PoseMamba-X	40	256	243	26.535 M	109.909 G

Table 2: PoseMamba model variants. N: Number of layers. d_m: Dimension of model. T: Number of input frames.

Method	T	MPJPE↓
MHFormer (Li et al. 2022b)	9	58.0
MixSTE (Zhang et al. 2022)	27	54.9
P-STMO (Shan et al. 2022)	81	32.2
Einfalt *et al.* (Einfalt, Ludwig, and Lienhart 2023)	81	46.9
STCFormer (Tang et al. 2023)	81	23.1
PoseFormerV2 (Zhao et al. 2023)	81	27.8
GLA-GCN (Yu et al. 2023)	81	27.7
HSTFormer (Qian et al. 2023)	81	41.4
HDFormer (Chen et al. 2023)	96	37.2
MotionAGFormer-XS	27	19.2
MotionAGFormer-S	81	17.1
MotionAGFormer-B	81	18.2
MotionAGFormer-L	81	16.2
PoseMamba-S	27	17.79
PoseMamba-S	81	<u>15.27</u>
PoseMamba-B	81	**14.51**

Table 3: Quantitative comparisons on MPI-INF-3DHP. T: Number of input frames. The best and second-best scores are in bold and underlined, respectively.

Effect of Loss Function We explore the contribution of our loss function using our small variant version in detail. As shown in Table 5, the MPJPE metric decreases from 43.7 to 43.5 mm after applying the 2D loss and decreases from 43.5 to 42.1 mm after applying the T-Loss. The result

Spatial-Temporal Modeling Strategy	Params	MACs	MPJPE
Unidirectional Spatial-Temporal 1	0.860 M	3.587 G	43.1
Unidirectional Spatial-Temporal 2	0.860 M	3.587 G	43.2
Unidirectional Spatial-Temporal 3	0.860 M	3.587 G	43.8
Unidirectional Spatial-Temporal 4	0.860 M	3.587 G	43.0
Bidirectional Spatial-Temporal	0.860 M	3.587 G	42.4
Bidirectional Global-Local Spatial-Temporal	0.860 M	3.587 G	**41.8**

Table 4: Ablation study for various spatial-temporal modeling with MPJPE on Human3.6M.

demonstrates that the T-Loss and 2D loss is an essential loss to improve accuracy. Finally, after applying the T-Loss, 2D-loss, and MPJPE loss to our method, the result achieves the best on the MPJPE metrics 41.8 mm. The results demonstrate that our loss function is comprehensive for the proposed model regarding accuracy and smoothness.

Loss	MPJPE↓ /PMPJPE↓
MPJPE Loss	43.7/36.5
MPJPE Loss + 2D-Loss	43.5/36.2
MPJPE Loss + T-Loss	42.1/35.1
Ours (MPJPE Loss + T-Loss + 2D-Loss)	**41.8/35.0**

Table 5: Ablation study for loss function with MPJPE and PMPJPE on Human3.6M.

Parameter Setting Analysis Table 6 shows how the setting of different hyper-parameters in our method impacts the performance under Protocol 1 with MPJPE. There are three main hyper-parameters for the network: the depth of PoseMamba (N), the dimension of model (d_m), and the input sequence length (T). We divide the configurations into 2 groups row-wise, and different values are assigned for one hyper-parameters while keeping the other two hyper-

Depth	d_m	Param	MACs	P1
12	64	0.516 M	2.2 G	43.0
16	64	0.688 M	2.9 G	42.0
20	64	0.860 M	3.6 G	41.8
24	64	1.031 M	4.3 G	41.5
32	64	1.375 M	5.7 G	41.5
40	64	1.719 M	7.2 G	41.1
48	64	2.062 M	8.6 G	41.1
40	32	0.450 M	1.9 G	43.1
40	64	1.719 M	7.2 G	41.1
40	128	6.714 M	27.9 G	38.1
40	256	26.535 M	109.9 G	37.1
12	256	7.963 M	33.0 G	39.1
20	128	3.358 M	13.9 G	40.8

Table 6: The P1 error comparison by varying number of PoseMamba blocks and number of channels on Human3.6M. d_m: Number of channels in each PoseMamba block. T is kept 243 in all experiments.

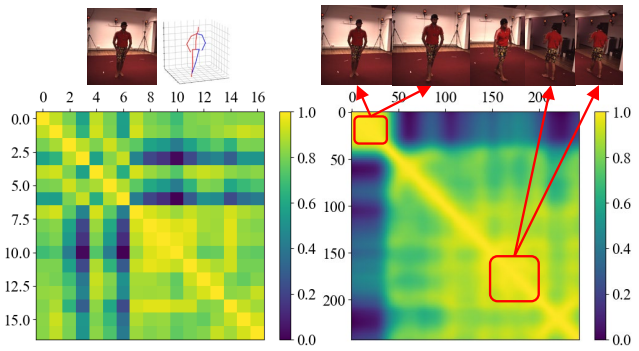


Figure 5: Visualization of SSM map among body joints and frames.

parameters fixed to evaluate the impact and choice of each configuration. In addition to these two sets of experiments, we have also conducted additional hyperparameter experiments. Based on the results in the table, considering performance and efficiency, we choose three variants in Table 2.

Qualitative Analysis

Figure 5 visualizes last spatio-temporal SSM block map of action (*Walking* of testset *S9*). It can be easily observed from spatial map (left of Figure 5) that our model learns distinct dependencies between joints. Furthermore, we also visualize the temporal map (right of Figure 5). The two light-colored parts have similar poses in adjacent frames, while dark-colored frame (the middle image in the frame sequence) has a more distinct pose in adjacent frames. Figure 6 compares PoseMamba-L with recent approaches, which shows that our PoseMamba achieves more accurate poses than MotionBERT and MotionAGFormer. Moreover, Figure 7 shows the qualitative comparison on some wild videos. It is evident that our method can produce more accurate 3D poses, particularly in cases the human action is complex and rare.

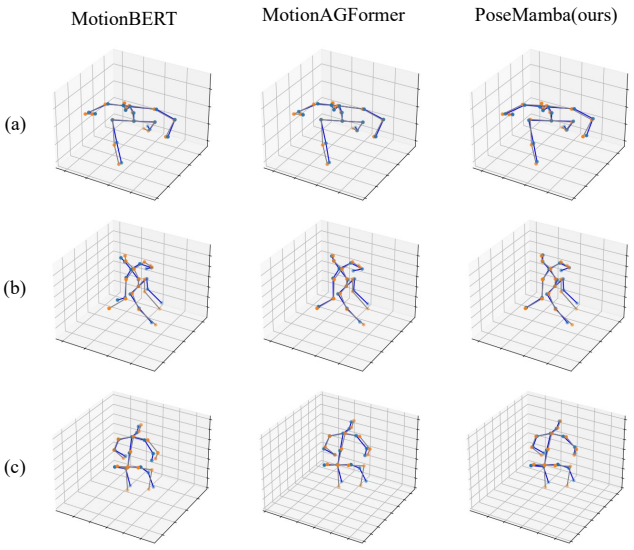


Figure 6: Qualitative comparisons with MotionBERT and MotionAGFormer. The gray skeleton is the ground-truth 3D pose and the blue skeleton is the estimated body.

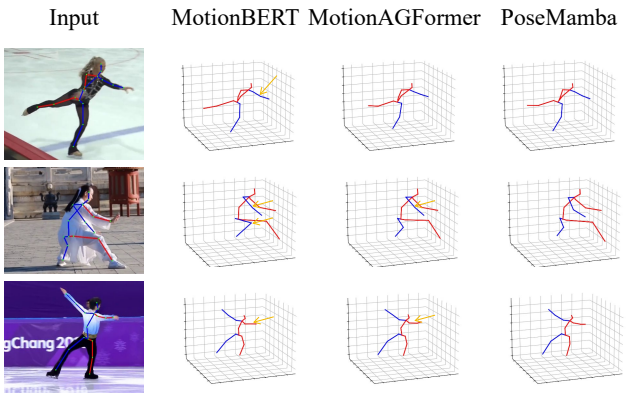


Figure 7: Qualitative comparisons with MotionBERT and MotionAGFormer on challenging wild videos. Wrong estimations are highlighted by yellow arrows.

Conclusion

We present PoseMamba, a novel SSM-based approach for 3D human pose estimation, which has a bidirectional global-local spatio-temporal mamba block to comprehensively model the human joint relations within each frame as well as the temporal correlations across frames. In the bidirectional global-local spatio-temporal mamba block, we propose a reordering strategy to enhance SSM's local modeling ability by providing a more logical geometric scanning order and fusing it with global SSM to get global-local spatial scan. Experimental results demonstrate that PoseMamba outperforms the existing counterparts on both datasets while significantly reducing parameters and MACs. PoseMamba is a promising new option for 3D human pose estimation, potentially providing a new perspective for the field.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62406120, the Guangxi Science and Technology Project (GuiKe-AB21196034), and the CCF-Zhipu Large Model Innovation Fund (NO.CCF-Zhipu202411).

References

Bauer, P.; Bouazizi, A.; Kressel, U.; and Flohr, F. B. 2023. Weakly Supervised Multi-Modal 3D Human Body Pose Estimation for Autonomous Driving. In *IEEE Intelligent Vehicles Symposium*, 1–7.

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. 2020. Anatomy-aware 3D Human Pose Estimation in Videos. *arXiv preprint arXiv:2002.10322*.

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/CVF Conference on Computer Vision and Pattern Recognition*, 7103–7112.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1251–1258.

Chun, S.; Park, S.; and Chang, J. Y. 2023. Learnable human mesh triangulation for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2850–2859.

Czech, P.; Braun, M.; Kreßel, U.; and Yang, B. 2022. On-Board Pedestrian Trajectory Prediction Using Behavioral Features. In *IEEE International Conference on Machine Learning and Applications*, 437–443.

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.

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

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

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

Holmquist, K.; and Wandt, B. 2023. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15977–15987.

Hossain, M. R. I.; and Little, J. J. 2018. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision*, 68–84.

Huang, Z.; Shi, M.; Liu, C.; Xian, K.; and Cao, Z. 2023. SimHMR: A Simple Query-based Framework for Parameterized Human Mesh Reconstruction. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6918–6927.

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.

Islam, M. M.; and Bertasius, G. 2022. Long movie clip classification with state-space video models. In *Proceedings of the European Conference on Computer Vision*, 87–104.

Kang, H.; Wang, Y.; Liu, M.; Wu, D.; Liu, P.; and Yang, W. 2023. Double-chain constraints for 3d human pose estimation in images and videos. *arXiv preprint arXiv:2308.05298*.

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.; 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.

Liu, Q.; Zhang, Y.; Bai, S.; and Yuille, A. 2022. Explicit occlusion reasoning for multi-person 3d human pose estimation. In *European Conference on Computer Vision*, 497–517. Springer.

Liu, R.; Shen, J.; Wang, H.; Chen, C.; Cheung, S.-c.; and Asari, V. 2020. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5064–5073.

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*.

Martin, E.; and Cundy, C. 2017. Parallelizing linear recurrent neural nets over sequence length. *arXiv preprint arXiv:1709.04057*.

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. 2017a. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision*, 506–516.

Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.-P.; Xu, W.; Casas, D.; and Theobalt, C. 2017b. VNect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4): 1–14.

Mondal, A.; Alletto, S.; and Tome, D. 2024. HumMUSS: Human Motion Understanding using State Space Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2318–2330.

Munea, T. L.; Jembre, Y. Z.; Weldegebriel, H. T.; Chen, L.; Huang, C.; and Yang, C. 2020. The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation. *IEEE Access*, 8: 133330–133348.

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

Pavlakos, G.; Zhou, X.; and Daniilidis, K. 2018. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7307–7316.

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.

Pióro, M.; Ciebiera, K.; Król, K.; Ludziejewski, J.; and Jaszczur, S. 2024. Moe-mamba: Efficient selective state space models with mixture of experts. *arXiv preprint arXiv:2401.04081*.

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*.

Reddy, N. D.; Guigues, L.; Pishchulin, L.; Eledath, J.; and Narasimhan, S. G. 2021. TesseTrack: End-to-end learnable multi-person articulated 3d pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15190–15200.

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 *Proceedings of the 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*.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5693–5703.

Sun, X.; Xiao, B.; Wei, F.; Liang, S.; and Wei, Y. 2018. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision*, 529–545.

Tang, Z.; Qiu, Z.; Hao, Y.; Hong, R.; and Yao, T. 2023. 3D Human Pose Estimation With Spatio-Temporal Criss-Cross Attention. In *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, S.; Xiong, Y.; and Lin, D. 2020. Motion guided 3d pose estimation from videos. In *Proceedings of the European Conference on Computer Vision*, 764–780.

Wang, J.; Zhu, W.; Wang, P.; Yu, X.; Liu, L.; Omar, M.; and Hamid, R. 2023. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6387–6397.

Wiederer, J.; Bouazizi, A.; Kressel, U.; and Belagiannis, V. 2020. Traffic control gesture recognition for autonomous vehicles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 10676–10683.

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.

Zeng, A.; Sun, X.; Huang, F.; Liu, M.; Xu, Q.; and Lin, S. 2020. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *Proceedings of the European Conference on Computer Vision*, 507–523.

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, X.; Bao, Q.; Cui, Q.; Yang, W.; and Liao, Q. 2024. Pose Magic: Efficient and Temporally Consistent Human Pose Estimation with a Hybrid Mamba-GCN Network. *arXiv preprint arXiv:2408.02922*.

Zhang, Z.; Wang, C.; Qiu, W.; Qin, W.; and Zeng, W. 2021. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision*, 129: 703–718.

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 International Conference on Computer Vision*, 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.

Zhou, K.; Han, X.; Jiang, N.; Jia, K.; and Lu, J. 2019. HEMlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2344–2353.

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.