This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

High-Resolution Spatiotemporal Modeling with Global-Local State Space Models for Video-Based Human Pose Estimation

Runyang Feng[1,2], Hyung Jin Chang[3], Tze Ho Elden Tse[4], Boeun Kim[5], Yi Chang[1,2], Yixing Gao[1,2*]
[1] School of Artificial Intelligence, Jilin University,
[2] Engineering Research Center of Knowledge-Driven Human-Machine Intelligence,
Ministry of Education, China, [3]School of Computer Science, University of Birmingham,
[4]National University of Singapore, [5]Dankook University

Abstract

Modeling high-resolution spatiotemporal representations, including both global dynamic contexts (e.g., holistic human motion tendencies) and local motion details (e.g., high-frequency changes of keypoints), is essential for video-based human pose estimation (VHPE). Current state-of-the-art methods typically unify spatiotemporal learning within a single type of modeling structure (convolution or attention-based blocks), which inherently have difficulties in balancing global and local dynamic modeling and may bias the network to one of them, leading to suboptimal performance. Moreover, existing VHPE models suffer from quadratic complexity when capturing global dependencies, limiting their applicability especially for high-resolution sequences. Recently, the state space models (known as Mamba) have demonstrated significant potential in modeling long-range contexts with linear complexity; however, they are restricted to 1D sequential data. In this paper, we present a novel framework that extends Mamba from two aspects to separately learn global and local high-resolution spatiotemporal representations for VHPE. Specifically, we first propose a Global Spatiotemporal Mamba, which performs 6D selective space-time scan and spatial- and temporal-modulated scan merging to efficiently extract global representations from high-resolution sequences. We further introduce a windowed space-time scan-based Local Refinement Mamba to enhance the high-frequency details of localized keypoint motions. Extensive experiments on four benchmark datasets demonstrate that the proposed model outperforms state-of-the-art VHPE approaches while achieving better computational trade-offs.

1. Introduction

Human pose estimation is a fundamental task in computer vision that has attracted increasing attention in recent years.

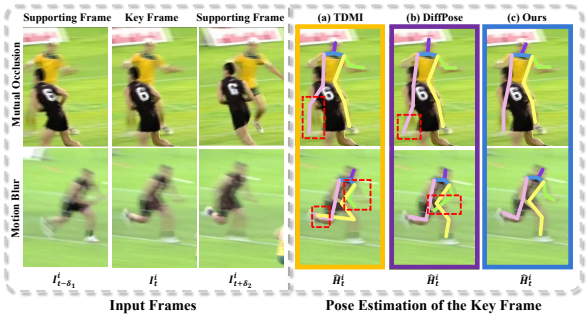


Figure 1. State-of-the-art methods such as (a) TDMI [10] and (b) DiffPose [11] focus either on global or local spatiotemporal contexts, which may fail for occlusion or blur cases. Our method (c) fully exploits both global and local high-resolution spatiotemporal representations, delivering more robust results.

The objective is to detect and localize anatomical human keypoints, such as elbows and wrists, from still images or video sequences. It finds enormous applications in diverse realistic scenes including human behavior understanding, augmented reality, and surveillance tracking [10, 38].

Accurately estimating human poses from videos requires dense spatiotemporal analysis, which significantly benefits from global and local high-resolution representations [40, 46]. The former typically characterizes holistic human motion patterns and contexts, while the later captures detailed high-frequency variations of local keypoints. With the surge in deep learning, numerous VHPE approaches using a single type of modeling structure such as convolutions [18] or Transformers [7] have been designed.

The CNN-based methods [3, 10, 28] usually design convolutional networks to integrate spatial and temporal information derived from HRNet (an off-the-shelf model for extracting high-resolution image features). For instance, [10] computes feature differences among frames to capture motion clues, and aggregates high-resolution appearance and motion features using convolutions to estimate pose heatmaps. [28] employs convolutions to align multi-

*Corresponding Author (gaoyixing@jlu.edu.cn)

ple supporting frames to the keyframe and fuses all aligned feature maps for pose estimation. However, the fixed receptive fields inherent in convolutions constrain the global inference capability of these approaches, which may result in large prediction deviations for degraded body parts in challenging cases. As illustrated in Fig. 1 (a), TDMI [10] produces inaccurate estimations for right leg in mutual occlusion scenes (or left arm for blur cases). In contrast, Transformer-based methods [11, 17] adopt a self-attention mechanism, allowing them to capture global dependencies of the input sequence. [11] concatenates the features of each frame and employs plain Vision Transformers to obtain global spatiotemporal representations. Nevertheless, the attention-based models often suffer from inferior local high-frequency details, and tend to yield inaccurate detections for localized ambiguous joints (right ankle in top of Fig. 1 (b) or left wrist in bottom). Moreover, the computation of self-attention involves quadratic complexity with respect to input tokens. Directly applying self-attention to high-resolution sequences (*e.g.*, $1/4 * T$) would result in excessive computation and memory overheads (Table 7).

Recently, state space models (SSMs) have gained significant attention for their strengths in capturing long-range dependencies [14, 51]. Notably, Mamba [12] incorporates parallelized selective scan and a hardware-aware algorithm, achieving remarkable performance in long language modeling with linear complexity. Despite these merits, Mamba's core operator, the *vanilla* selective scan, is specially designed for 1D sequential data. This presents substantial challenges when adapting to the spatiotemporal information in videos. Recent variants [24, 33] attempt to extend Mamba to video processing (*e.g.*, high-level video understanding) via frame-by-frame bidirectional scanning. They simply flatten the spatial tokens of each frame within the sequence to model global dependencies. However, such scanning schemes focus on sequential spatial processing and do not take into account of insights from other scanning directions, which is detrimental to the dense analysis of high-resolution spatiotemporal contexts in VHPE. For instance, they elongate the distance between temporally adjacent tokens, leading to insufficient capture of temporal-wise pixel dynamics. Moreover, these methods lack specific designs to guide the Mamba to learn fine-grained local details from sequences. Directly applying them to VHPE produces inferior performance.

Inspired by the preceding analysis, we design a decoupled framework based on pure **Mamba** to explore **G**lobal and **L**ocal high-resolution **S**patiotemporal representations for VHPE (GLSMamba). The proposed GLSMamba extends Mamba in two aspects: **(i)** A Global Spatiotemporal Mamba (GSM) is designed for holistic contextual sequence modeling at high resolutions. Specifically, GSM engages a *6D selective Space-Time Scan (STS6D)* mechanism, which traverses along six tailored spatiotemporal scanning routes to fully resolve the high-resolution feature sequences from a global perspective. Then, GSM adaptively aggregates the scanning knowledge from different routes via a *Spatial- and Temporal-Modulated scan Merging (STMM)* strategy, thereby bridging the gap between 1D selective scan and high-resolution sequences. **(ii)** A Local Refinement Mamba (LRM) is further introduced to enhance the high-frequency details of local motion representations. LRM performs *frame-wise selective scan within windowed patch cubes*, processing localized pixels inside the same semantic 3D tubelet compactly together to effectively capture local spatiotemporal dependencies. This module significantly enhances fine-grained motion details while preserving sequence-size receptive field. Thanks to the Mamba-based decoupled structure design, our approach delivers more reliable high-resolution spatiotemporal representations that are globally consistent and locally enriched, and possesses better computational trade-offs.

From extensive evaluations on four widely-used benchmark datasets (PoseTrack2017, PoseTrack2018, PoseTrack21, and Sub-JHMDB), we show that GLSMamba surpasses state-of-the-art VHPE methods. We also provide ablation analysis on the effectiveness of each proposed component and design choice.

The key contributions of this work can be summarized as: (i) We propose to decouple the modeling of global dynamic contexts and local motion details for video-based human pose estimation. (ii) We present GLSMamba, the first pure Mamba-based framework for VHPE. GLSMamba extends the *vanilla* state space model in two ways, forming GSM and LRM to learn global and local high-resolution spatiotemporal representations, respectively. (iii) We demonstrate that GLSMamba achieves competitive state-of-the-art performance with fewer parameters on four benchmark datasets: PoseTrack2017, PoseTrack2018, PoseTrack21, and Sub-JHMDB.

2. Related Work

Human pose estimation in images. Estimating human joint locations from still images has been extensively studied which generally falls into two paradigms: bottom-up and top-down. *Bottom-up* approaches [4, 5, 22, 23] first detect individual body parts and then group them into an entire human skeleton. The main variation among these methods lies in the grouping algorithms, such as Part Affinity Field in [4] and Associative Embedding in [31]. Conversely, *top-down* approaches [25, 38, 42, 44] first extract human bounding boxes using an object detector, and then design models to estimate human poses within each bounding box region. [38] presents a high-resolution convolutional network that retains high-resolution feature maps throughout all stages and performs repeated multi-scale fusion to obtain rich hu-

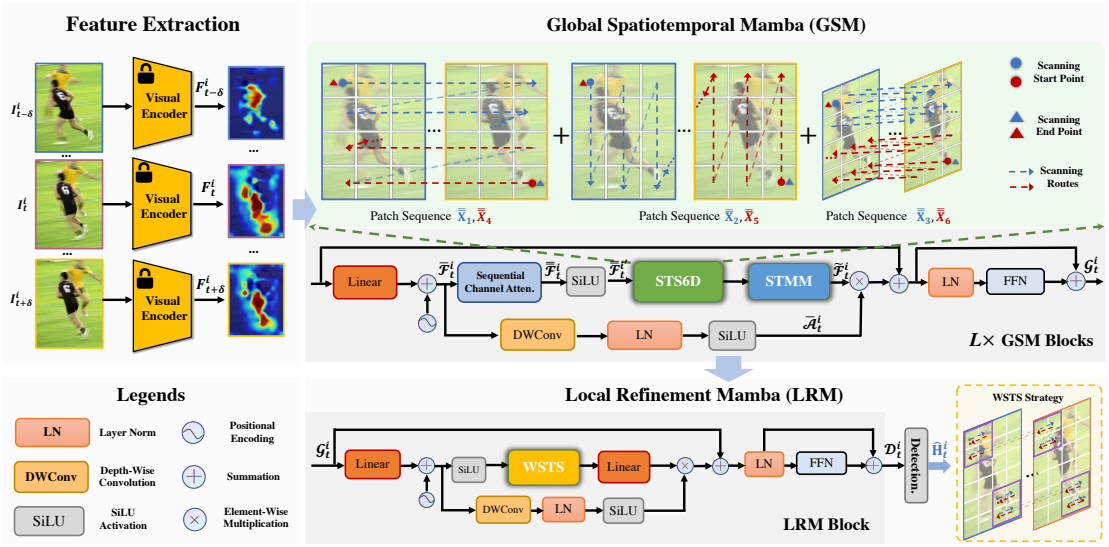


Figure 2. Overall pipeline of the proposed framework. Given an input sequence, we first extract high-resolution spatial features for each frame using a visual encoder. Then, these features are processed successively by GSM and LRM for global spatiotemporal modeling and local detail enhancement. Finally, a detection head is employed to yield the pose heatmap estimations.

man body information. [44] leverages cascaded plain vision transformers to learn generalizable image representations, achieving superior performance on multiple benchmarks.

Human pose estimation in videos. Directly applying existing image-based pose estimation models to videos often yields suboptimal results because they fail to capture the temporal dynamics among frames. To incorporate spatiotemporal contexts, several state-of-the-art approaches [3, 10, 27, 28] integrate HRNet [38] as the backbone network, and adopt a convolutional architecture to aggregate high-resolution spatial and temporal feature representations. [3, 27] compute joint motion offsets between frames and employ the motion information to guide accurate pose heatmap resampling. [10] introduces temporal feature differences as the motion clues, and employs convolutional blocks to aggregate appearance and motion features. A primary drawback of these methods lies in the restricted global spatiotemporal perception ability due to the limited receptive field, which may hinder their performance. Another line of work [11, 17] considers Transformers (self-attentions) for global spatiotemporal modeling. [11] extracts feature tokens for each frame, and then employs Vision Transformers to capture the global dependencies of the token sequence. However, these methods often neglect valuable high-frequency details of local keypoint motions, and incur quadratic computational complexity that is detrimental to high-resolution modeling. Different from the above methods, we aim to introduce a novel decoupled architecture to fully learn both global and local high-resolution spatiotemporal contexts for VHPE while maintaining an acceptable computational load.

State space models. State space models (SSMs) are a type of foundation models and have recently demonstrated great potential in capturing long range dependencies through HiPPO matrix initialization [13, 15]. To facilitate the practical applicability of SSMs, [14] proposes the structured SSM model (S4) which imposes a diagonalization structure on the parameter matrix, significantly reducing the computational overhead. The promising results from S4 have inspired the emergence of numerous SSM-based architectures. For example, S5 [36] proposes a multi-input and multi-output SSM, GSS [30] integrates a gated mechanism, and the recent advancement Mamba [12] introduces context-based reasoning and parallelized selective scanning. Due to the exceptional performance in long sequence modeling with linear complexity, Mamba has become a compelling alternative to Transformers, finding extensive applications in diverse fields ranging from language and audio [12] to vision tasks [26, 33, 51].

Despite several visual Mamba variants for action recognition [24, 33], these models have difficulties in adequately processing dense high-resolution sequence contexts. They also lack specific designs to capture local spatiotemporal details. In contrast, we purposefully extend Mamba for VHPE from two aspects, with a focus on learning global and local high-resolution spatiotemporal representations.

3. Our Approach

Problem formulation. Our work follows a top-down paradigm in which a human detector is first used to obtain the human bounding boxes in a frame I_t. Then, each of the bounding boxes is enlarged by 125% to crop the same individual i across a frame sequence $\mathcal{I}_t^i = \langle I_{t-\delta}^i, ..., I_t^i, ..., I_{t+\delta}^i \rangle$, where δ denotes the temporal span.

Given $\boldsymbol{\mathcal{I}_t^i}$, we seek to explore the spatiotemporal clues to foster the pose estimation in the current frame I_t^i .

Method overview. The overall pipeline of our proposed GLSMamba framework is shown in Fig. 2. Our objective is to extend Mamba to model global-local high-resolution spatiotemporal contexts effectively. There are two key components: Global Spatiotemporal Mamba (GSM) and Local Refinement Mamba (LRM). Specifically, we first extract high-resolution features for each frame $\boldsymbol{\mathcal{F}_t^i} = \langle F_{t-\delta}^i, ..., F_t^i, ..., F_{t+\delta}^i \rangle$ using a visual encoder. Then, these features are fed into GSM for global spatiotemporal modeling. The resulting tensor $\boldsymbol{\mathcal{G}_t^i}$ is passed to LRM to enhance the local spatiotemporal details and yield $\boldsymbol{\mathcal{D}_t^i}$. Finally, a detection head is used to estimate the pose heatmap $\hat{\mathbf{H}}_t^i$. In the following, we present the preliminaries of SSMs (Sec. 3.1), and detail the architectures of each component including GSM (Sec. 3.2) and LRM (Sec. 3.3).

3.1. Preliminaries

State space models. SSMs are inspired by the continuous system that maps a 1D input signal to output response $x(n) \in \mathbb{R} \mapsto y(n) \in \mathbb{R}$ via a hidden state $\mathbf{h}(n) \in \mathbb{R}^N$. Formally, SSMs can be expressed as the following ordinary differential equations:

$$\begin{aligned} \mathbf{h}'(n) &= \mathbf{A}\mathbf{h}(n) + \mathbf{B}x(n), \\ y(n) &= \mathbf{C}\mathbf{h}(n) + Dx(n), \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ denotes the evolution parameter, and $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$ are projection matrixes. The parameter $D \in \mathbb{R}^1$ can be ignored as a residual connection.

To apply in the deep learning context, the above continuous system has to be discretized. The commonly used technique for discretizing SSMs, known as zero-order hold (ZOH), incorporates a step size $\boldsymbol{\Delta}$ to convert the continuous parameters $\mathbf{A}$ and $\mathbf{B}$ into their discrete counterparts, $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$, respectively. This can be defined as:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp\left(\boldsymbol{\Delta}\mathbf{A}\right), \\ \overline{\mathbf{B}} &= \left(\boldsymbol{\Delta}\mathbf{A}\right)^{-1}\left(\exp\left(\boldsymbol{\Delta}\mathbf{A}\right) - \mathbf{I}\right) \cdot \boldsymbol{\Delta}\mathbf{B}. \end{aligned} \tag{2}$$

Consequently, the continuous-time SSMs in Eqs. 1 can be rewritten as:

$$\begin{aligned} \mathbf{h}_n &= \overline{\mathbf{A}}\mathbf{h}_{n-1} + \overline{\mathbf{B}}x_n, \\ y_n &= \mathbf{C}\mathbf{h}_n, \end{aligned} \tag{3}$$

which can be efficiently computed via global convolutions.

Selective SSM. A key property of the aforementioned SSM models is linear time invariance (LTI), implying that the parameters $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \boldsymbol{\Delta})$ remain constant and independent of the input across different time steps. To overcome this limitation, Mamba [12] introduces a selective scan mechanism (S6) as the core operator. Unlike LTI SSMs, Mamba dynamically generates model parameters based on the input, enabling context-based reasoning with linear complexity. Given these advantages, our work also adopts the S6 operation as a foundation.

3.2. Global Spatiotemporal Mamba

As illustrated in Fig. 2, we introduce the Global Spatiotemporal Mamba (GSM) to learn high-resolution spatiotemporal representations from a global perspective. To achieve this, we first construct the high-resolution feature sequence $\boldsymbol{\mathcal{F}_t^i}$ for the input clip $\boldsymbol{\mathcal{I}_t^i}$. The sequence $\boldsymbol{\mathcal{F}_t^i}$ is then passed through cascaded GSM blocks to produce the spatiotemporal features $\boldsymbol{\mathcal{G}_t^i}$.

High-resolution feature sequence extraction. Given $\boldsymbol{\mathcal{I}_t^i} = \langle I_{t-\delta}^i, ..., I_t^i \in \mathbb{R}^{\mathtt{C}\times\mathtt{H}\times\mathtt{W}}, ..., I_{t+\delta}^i \rangle$, a visual encoder pretrained on COCO is first leveraged to extract the features of each frame $\boldsymbol{\mathcal{F}_t^i} = \langle F_{t-\delta}^i, ..., F_t^i, ..., F_{t+\delta}^i \rangle$, where $(\mathtt{H}, \mathtt{W})$ indicates the image size and $\mathtt{C}$ is the number of channels. To ensure the high spatial resolution, we utilize ViTPose [44] to extract image features followed by deconvolution structures for spatially upsampling by $4\times$. Note that the parameters of the visual encoder are frozen during the model optimization.

GSM block. After obtaining the feature sequence $\boldsymbol{\mathcal{F}_t^i}$, we design the GSM block with the novel 6D selective Space-Time Scan (STS6D) and Spatial- and Temporal-Modulated scan Merging (STMM) mechanisms to model holistic spatiotemporal contexts. The core parts, STS6D and STMM, will be detailed in the following section. Specifically, the feature of each frame within $\boldsymbol{\mathcal{F}_t^i}$ is first linearly projected to a tensor with size $\mathtt{D}$, and combined with a sine-cosine spatial embedding $\mathbf{E}_{spa}$ [39] as well as a learnable temporal embedding $\mathbf{E}_{tem}$ to preserve the spatial and temporal information, deriving $\bar{\boldsymbol{\mathcal{F}}}_{\boldsymbol{t}}^{\boldsymbol{i}} = \langle \bar{F}_{t-\delta}^i, ..., \bar{F}_t^i \in \mathbb{R}^{\mathtt{D}\times\mathtt{h}\times\mathtt{w}}, ..., \bar{F}_{t+\delta}^i \rangle$:

$$\bar{\boldsymbol{\mathcal{F}}}_{\boldsymbol{t}}^{\boldsymbol{i}} = \mathrm{Linear}\left(\boldsymbol{\mathcal{F}_t^i}\right) + \mathbf{E}_{spa} + \mathbf{E}_{tem}, \tag{4}$$

where $\mathtt{h} = \frac{1}{4}\mathtt{H}$ and $\mathtt{w} = \frac{1}{4}\mathtt{W}$. Subsequently, $\bar{\boldsymbol{\mathcal{F}}}_{\boldsymbol{t}}^{\boldsymbol{i}}$ is processed through two separate streams:

(1) Main Stream: To facilitate the global sequence modeling in STS6D and STMM, we first introduce a Sequential Channel Attention which adaptively activates significant spatiotemporal information within $\bar{\boldsymbol{\mathcal{F}}}_{\boldsymbol{t}}^{\boldsymbol{i}}$ at the channel level. Specifically, we concatenate the feature sequence and squeeze the global spatiotemporal information into sequential (frame-wise) channel descriptors via a global average pooling (GAP) layer. Next, several MLPs are leveraged to model channel interactions both spatially (intra-frame) and temporally (inter-frame), followed by a sigmoid function to obtain the sequential attention weights. The attention matrix ($\langle M_{t-\delta}^i, ..., M_t^i \in \mathbb{R}^{\mathtt{D}\times 1}, ..., M_{t+\delta}^i \rangle$) is used to rescale the input sequence $\bar{\boldsymbol{\mathcal{F}}}_{\boldsymbol{t}}^{\boldsymbol{i}}$ to obtain the modulated version $\bar{\bar{\boldsymbol{\mathcal{F}}}}_{\boldsymbol{t}}^{\boldsymbol{i}}$. The above process is formulated as:

$$\bar{\bar{\boldsymbol{\mathcal{F}}}}_{\boldsymbol{t}}^{\boldsymbol{i}} = \sigma\left(\mathrm{MLPs}\left(\mathrm{GAP}\left(\bar{\boldsymbol{\mathcal{F}}}_{\boldsymbol{t}}^{\boldsymbol{i}}\right)\right)\right) \otimes \bar{\boldsymbol{\mathcal{F}}}_{\boldsymbol{t}}^{\boldsymbol{i}}. \tag{5}$$

We then normalize and use SiLU [35] to transform $\bar{\bar{\mathcal{F}}}_t^i$, and feed the resulting tensor $\bar{\bar{\mathcal{F}}}_t^{i'}$ into STS6D and STMM to model global spatiotemporal dependencies and output $\tilde{\mathcal{F}}_t^i$.

(2) Another stream serves as a gated attention to further control the raw feature element propagation meticulously, which passes $\bar{\mathcal{F}}_t^i$ into a depth-wise convolution, followed by a LayerNorm and a SiLU activation to yield $\bar{\mathcal{A}}_t^i$.

Finally, the resulting features of these two branches are merged via multiplication, and fed into a feedforward neural network (FFN) to obtain the global spatiotemporal representations $\mathcal{G}_t^i$. In practice, we stack $L = 4$ GSM blocks for progressive information processing.

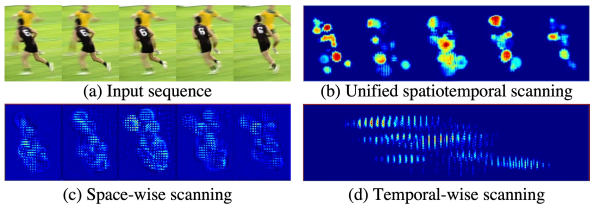
STS6D and STMM. Although the *vanilla* selective scan in S6 enjoys various advantages such as global modeling, context-aware inference, and linear complexity, it is designed for 1D sequential data that differs substantially from the video modality. To address this challenge, we design the 6D selective Space-Time Scan (STS6D) as well as the Spatial- and Temporal-Modulated scan Merging (STMM) modules, which adapt S6 to high-resolution spatiotemporal modeling while maintaining its strengths.

As illustrated in Fig. 2, we first flatten $\bar{\bar{\mathcal{F}}}_t^{i'}$ along six tailored space-time routes to obtain 1D patch sequences $\{\bar{\bar{\mathbf{x}}}_k\}_{k=1,2,...,6}$. Specifically, we stack the features of each frame within $\bar{\bar{\mathcal{F}}}_t^{i'}$ to form an image-like panoramic spatiotemporal representation, and traverse it horizontally and vertically to yield $\bar{\bar{\mathbf{x}}}_1$ and $\bar{\bar{\mathbf{x}}}_2$. We further perform pixel traversal along the depth (time) dimension across frames to attain $\bar{\bar{\mathbf{x}}}_3$. Reversing $\{\bar{\bar{\mathbf{x}}}_k\}_{k=1,2,3}$ produces the complete six-way scanning sequences. Then, each sequence is processed by a separate S6 block to capture the corresponding global dependencies:

$$\begin{aligned} &\mathbf{B}_k = f_B\left(\bar{\bar{\mathbf{x}}}_k\right), \mathbf{C}_k = f_C\left(\bar{\bar{\mathbf{x}}}_k\right), \boldsymbol{\Delta}_k = f_\Delta\left(\bar{\bar{\mathbf{x}}}_k\right), \\ &\overline{\mathbf{A}_k}, \overline{\mathbf{B}_k} = \mathrm{Dis}\left(\boldsymbol{\Delta}_k, \mathbf{A}_k, \mathbf{B}_k\right), \\ &\tilde{\mathbf{y}}_k = \mathrm{SSM}(\overline{\mathbf{A}_k}, \overline{\mathbf{B}_k}, \mathbf{C}_k)\left(\bar{\bar{\mathbf{x}}}_{\boldsymbol{k}}\right), \end{aligned} \tag{6}$$

where $k \in \{1, 2, ..., 6\}$, (f_B, f_C, f_Δ) refers to independent linear projections to generate parameters $(\mathbf{B}, \mathbf{C}, \boldsymbol{\Delta})$, and $\mathbf{A}$ is a learnable matrix with random initialization. The symbol $\mathrm{Dis}(\cdot)$ denotes the discretization progress in Eqs. 2, and $\mathrm{SSM}(\cdot)$ indicates the computations of state space model in Eqs. 3. *Intuitively, the selective scanning of different routes can characterize a video clip from diverse views.* For instance, the unified scanning $\{\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_4\}$ captures high-level spatiotemporal representations of salient global dynamic contexts, as shown in Fig. 3 (b). In contrast, the space-wise scanning $\{\tilde{\mathbf{y}}_2, \tilde{\mathbf{y}}_5\}$ provides complete human spatial contexts of each frame, while the time-wise scanning $\{\tilde{\mathbf{y}}_3, \tilde{\mathbf{y}}_6\}$ approximates the dense motion tendencies of human body.

Subsequently, given the processed feature sequences $\{\tilde{\mathbf{y}}_k\}_{k=1,2,...,6}$ with different semantics, a STMM mechanism is further proposed to adaptively aggregate them and



(a) Input sequence (b) Unified spatiotemporal scanning (c) Space-wise scanning (d) Temporal-wise scanning

Figure 3. Visualizations of activation maps of STS6D.

yield $\tilde{\mathcal{F}}_t^i$. To be specific, we first invert the backward scan sequences, and merge features belonging to the same type of scanning via an addition operation:

$$\tilde{\mathbf{y}}_u = \tilde{\mathbf{y}}_1 + \mathrm{Iv}(\tilde{\mathbf{y}}_4), \tilde{\mathbf{y}}_s = \tilde{\mathbf{y}}_2 + \mathrm{Iv}(\tilde{\mathbf{y}}_5), \tilde{\mathbf{y}}_t = \tilde{\mathbf{y}}_3 + \mathrm{Iv}(\tilde{\mathbf{y}}_6), \tag{7}$$

where $\mathrm{Iv}(\cdot)$ is the inverse transformation, tensors $\tilde{\mathbf{y}}_u$, $\tilde{\mathbf{y}}_s$, and $\tilde{\mathbf{y}}_t$ denote the high-level (unified), space-wise, and time-wise representations, respectively. Then, we perform spatial-modulated and and temporal-modulated feature compensation to progressively update the high-level spatiotemporal features $\tilde{\mathbf{y}}_u$. Given $\tilde{\mathbf{y}}_u$ and $\tilde{\mathbf{y}}_s$, we reshape them to 2D sequences and concatenate them in the channel dimension. We then leverage convolutions to adaptively generate the kernel sampling offsets $\boldsymbol{\mathcal{O}}_{u;s}$ for $\tilde{\mathbf{y}}_u$, facilitating the learning of spatial compensation from $\tilde{\mathbf{y}}_s$. We also estimate the modulated scalars $\boldsymbol{\mathcal{W}}_{u;s}$ to control the sampling intensity. Finally, we conduct feature modulation via a deformable convolution (DCN [52]) to update $\tilde{\mathbf{y}}_u$ as:

$$\tilde{\mathbf{y}}_{u;s} = \tilde{\mathbf{y}}_u + \mathrm{DCN}\left(\tilde{\mathbf{y}}_u, \boldsymbol{\mathcal{O}}_{u;s}, \boldsymbol{\mathcal{W}}_{u;s}\right). \tag{8}$$

Similarly, the temporal-guided feature modulation is further performed over $\tilde{\mathbf{y}}_{u;s}$ and $\tilde{\mathbf{y}}_t$ for dense temporal compensation, obtaining $\tilde{\mathcal{F}}_t^i$.

By thoroughly traversing the whole space-time domain and adaptively aggregating the multi-source scanning knowledge, GSM empowers each pixel to gather insights from all others across multiple directions. This facilitates the comprehensive processing and resolving of high-resolution sequences from a global perspective.

3.3. Local Refinement Mamba

The tensor $\mathcal{G}_t^i$ derived from GSM attends to the global understanding of human motion patterns, yet lacks rich local details of keypoints. To further enhance the fine-grained local spatiotemporal representations, we propose the Local Refinement Mamba (LRM) using a Windowed Space-Time Scan (WSTS) strategy. WSTS processes local pixels within a windowed 3D tubelet closely together to capture local spatiotemporal dependencies.

Specifically, WSTS first splits the input feature sequences into a series of local windowed temporal tubes (*e.g.* $8 \times 6 \times T$). Then, a frame-wise selective scan is performed within each localized 3D tubelet. Concretely, each windowed feature tubelet is unrolled frame-by-frame in both forward and reverse directions, and then fed into a S6 block

Method	Backbone	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
PoseFlow [43]	ResNet-152	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
FastPose [49]	ResNet-101	80.0	80.3	69.5	59.1	71.4	67.5	59.4	70.3
SimplePose [42]	ResNet-152	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
STEmbedding [21]	Hourglass	83.8	81.6	77.1	70.0	77.4	74.5	70.8	77.0
HRNet [38]	HRNet-W48	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
MDPN [16]	ResNet-152	85.2	88.5	83.9	77.5	79.0	77.0	71.4	80.7
CorrTrack [34]	CPN	86.1	87.0	83.4	76.4	77.3	79.2	73.3	80.8
Dynamic [45]	HRNet-W48	88.4	88.4	82.0	74.5	79.1	78.3	73.1	81.1
PoseWarper [3]	HRNet-W48	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
DCPose [27]	HRNet-W48	88.0	88.7	84.1	78.4	83.0	81.4	74.2	82.8
DetTrack [41]	HRNet-W48	89.4	89.7	85.5	79.5	82.4	80.8	76.4	83.8
FAMI-Pose [28]	HRNet-W48	89.6	90.1	86.3	80.0	84.6	83.4	77.0	84.8
TDMI [10]	HRNet-W48	90.0	91.1	87.1	81.4	85.2	84.5	78.5	85.7
DiffPose [11]	ViT-B	89.0	91.2	87.4	83.5	85.5	87.2	80.2	86.4
DSTA [17]	ViT-H	89.3	90.6	87.3	82.6	84.5	85.1	77.8	85.6
GLSMamba-B	ViT-B	90.6	91.3	88.2	83.8	85.4	87.1	80.5	**86.9**
GLSMamba-H	ViT-H	90.7	92.1	89.2	85.3	87.0	88.4	82.4	**88.0**

Table 1. Quantitative results on the **PoseTrack2017** validation set.

Method	Backbone	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
AlphaPose [9]	Hourglass	63.9	78.7	77.4	71.0	73.7	73.0	69.7	71.9
MDPN [16]	ResNet-152	75.4	81.2	79.0	74.1	72.4	73.0	69.9	75.0
PGPT [2]	ResNet-152	-	-	-	72.3	-	-	72.2	76.8
Dynamic [45]	HRNet-W48	80.6	84.5	80.6	74.4	75.0	76.7	71.8	77.9
PoseWarper [3]	HRNet-W48	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
PT-CPN++ [47]	CPN	82.4	88.8	86.2	79.4	72.0	80.6	76.2	80.9
DCPose [27]	HRNet-W48	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
DetTrack [41]	HRNet-W48	84.9	87.4	84.8	79.2	77.6	79.7	75.3	81.5
FAMI-Pose [28]	HRNet-W48	85.5	87.7	84.2	79.2	81.4	81.1	74.9	82.2
TDMI [10]	HRNet-W48	86.2	88.7	85.4	80.6	82.4	82.1	77.5	83.5
DiffPose [11]	ViT-B	85.0	87.7	84.3	81.5	81.4	82.9	77.6	83.0
DSTA [17]	ViT-H	85.9	88.8	85.0	81.1	81.5	83.0	77.4	83.4
GLSMamba-B	ViT-B	85.0	88.2	85.6	82.9	82.5	84.9	79.7	**84.2**
GLSMamba-H	ViT-H	85.6	88.9	86.5	83.6	82.9	85.7	81.4	**84.9**

Table 2. Quantitative results on the **PoseTrack2018** validation set.

separately. We reshape the processed features and sum them to obtain locally enhanced spatiotemporal representations. Notably, WSTS leverages a non-overlapping window partition scheme to maintain the computational efficiency.

In our implementation, we remove the Sequential Channel Attention from the Global Spatiotemporal Mamba (GSM) block, and replace core operators *i.e.* STS6D and STMM with the proposed WSTS strategy to construct the LRM block. We employ 2 cascade LRM blocks to process the input tensor $\boldsymbol{\mathcal{G}_t^i}$, and obtain refined representations $\boldsymbol{\mathcal{D}_t^i}$ with abundant local details. Finally, we aggregate the features of each frame within $\boldsymbol{\mathcal{D}_t^i}$ via an element-wise addition, and feed the resulting tensor into a detection head (3×3 convolution) to yield the predicted pose heatmaps $\hat{\mathbf{H}}_t^i$.

3.4. Loss Function

We employ the standard heatmap estimation loss [28, 38] $\mathcal{L}_H$ to optimize the GLSMamba framework.:

$$\mathcal{L}_H = \left\| \hat{\mathbf{H}}_t^i - \mathbf{H}_t^i \right\|_2^2, \tag{9}$$

where $\hat{\mathbf{H}}_t^i$ and $\mathbf{H}_t^i$ denote the predicted and corresponding ground truth heatmaps, respectively.

4. Experiments

4.1. Experimental Settings

Datasets and evaluation. We evaluate our approach on four challenging VHPE benchmarks, including PoseTrack2017 [19], PoseTrack2018 [1], PoseTrack21 [6], and Sub-JHMDB [20]. Specifically, **PoseTrack2017** provides $80,144$ human pose annotations which are divided into *train*/*val* sets, consisting of 250 and 50 video clips, respectively. **PoseTrack2018** significantly expands the amount of data and contains 593 videos for training and 170 for validation, with a total of $153,615$ manually labeled poses. Both datasets are annotated with 15 keypoints along the same criteria, and include an extra flag for visibility. **PoseTrack21** further enriches the annotations of PoseTrack2018 especially for complex small persons, providing $177,164$ human pose labels. **Sub-JHMDB** contains 316 video sequences with $11,200$ frames. Following [28], we adopt three data splits for training and testing, and report the average performance. We benchmark the model over visible joints using the metric of average precision (AP) [27, 38].

Method	Backbone	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
Simple [42]	ResNet-152	80.5	81.2	73.2	64.8	73.9	72.7	67.7	73.9
HRNet [38]	HRNet-W48	81.5	83.2	81.1	75.4	79.2	77.8	71.9	78.8
PoseWarper [3]	HRNet-W48	82.3	84.0	82.2	75.5	80.7	78.7	71.6	79.5
DCPose [27]	HRNet-W48	83.2	84.7	82.3	78.1	80.3	79.2	73.5	80.5
FAMI-Pose [28]	HRNet-W48	83.3	85.4	82.9	78.6	81.3	80.5	75.3	81.2
TDMI [10]	HRNet-W48	85.8	87.5	85.1	81.2	83.5	82.4	77.9	83.5
DiffPose [11]	ViT-B	84.7	85.6	83.6	80.8	81.4	83.5	80.0	82.9
DSTA [17]	ViT-H	87.5	87.0	84.2	81.4	82.3	82.5	77.7	83.5
GLSMamba-B	ViT-B	86.3	86.7	85.1	82.1	83.0	84.3	79.4	**84.1**
GLSMamba-H	ViT-H	87.0	86.9	85.4	83.2	83.4	84.8	80.8	**84.7**

Table 3. Quantitative results on the **PoseTrack21** dataset.

Method	Backbone	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Avg**
Thin-slicing [37]	–	97.1	95.7	87.5	81.6	98.0	92.7	89.8	92.1
LSTM PM [29]	–	98.2	96.5	89.6	86.0	98.7	95.6	90.0	93.6
DKD [32]	ResNet-50	98.3	96.6	90.4	87.1	99.1	96.0	92.9	94.0
K-FPN [50]	ResNet-18	94.7	96.3	95.2	90.2	96.4	95.5	93.2	94.5
K-FPN [50]	ResNet-50	95.1	96.4	95.3	91.3	96.3	95.6	92.6	94.7
MAPN [8]	ResNet-18	98.2	97.4	91.7	85.2	99.2	96.7	92.2	94.7
FAMI-Pose [28]	HRNet-W48	99.3	98.6	94.5	91.7	99.2	91.8	95.4	96.0
DeciWatch [48]‡	SimplePose	99.8	99.5	99.7	99.7	98.7	99.4	96.5	**98.8**
GLSMamba-B	ViT-B	99.2	98.3	98.1	97.1	99.3	98.0	95.9	97.9

Table 4. Quantitative results on the **Sub-JHMDB** dataset.

Implementation details. The proposed GLSMamba framework is implemented by PyTorch. We incorporate data augmentations such as random rotation/scaling, truncation, and flipping in the training phase. We take ViTPose pretrained on COCO as the backbone, and freeze its parameters during training. The temporal span δ is set to 2. We employ AdamW optimizer with a base learning rate of $1e-4$, which decays to $1e-5$ at 6-th epoch and $1e-6$ for 12-th epoch. All training process is performed on one TITAN RTX GPU and terminated within 20 epochs.

4.2. Comparison with State-of-the-art Approaches

We first compare GLSMamba with state-of-the-art (SOTA) methods on the PoseTrack2017 validation set, and report the results in Table 1. We comprehensively evaluate GLSMamba under two widely-used backbones namely ViT-B and ViT-H, and provide the computational cost in Table 7. We observe that GLSMamba, the first pure Mamba-based VHPE framework with only 9.8 M trainable parameters,



Figure 4. Visual results of our method on benchmarks. Challenging scenes such as occlusion and motion blur are involved.

delivers SOTA pose estimation performance against existing well-established CNN- and Transformer-based models across various backbones. **(i)** Compared to the impressive convolution-based PoseWarper [3], GLSMamba-B attains a remarkable performance gain of 5.7 mAP with drastically reduced FLOPs ($\downarrow$ 34%). GLSMamba-B also improves the pose estimation performance by 1.2 mAP over the SOTA method TDMI [10]. Compared to the Transformer-based DiffPose [11] that operates on low-resolution sequences, GLSMamba-B improves the mAP by 0.5 points. Such compelling results demonstrate the importance of explicitly embracing both global and local high-resolution spatiotemporal contexts, reflecting the great potential of the novel Mamba-based architecture for VHPE. Noticeably, in contrast to existing SOTA methods [3, 11, 27] that often additionally fine-tune the backbone on VHPE datasets to improve performance, we directly leverage the pre-trained backbone weights on COCO from [44]. This simplifies the training pipeline, and remarkably diminishes the trainable parameters † by $\downarrow$ 86.2%. **(ii)** When adopting the larger backbone ViT-H, GLSMamba-H further pushes forward the performance boundary and achieves 88.0 mAP ($\uparrow$ 1.6).

Table 2 and Table 3 provide the experimental comparisons of various approaches on the PoseTrack2018 and PoseTrack21 datasets, respectively. With the base backbone ViT-B, our GLSMamba-B has already surpassed all other methods in both datasets. Our large model, GLSMamba-H, further obtains new state-of-the-art performance of 84.9 mAP and 84.7 mAP. We also illustrate in Fig. 4 the example visualizations of pose estimates in complex scenarios, which attest to the effectiveness of the proposed method.

Furthermore, we benchmark the proposed model on Sub-JHMDB and tabulate the results in Table 4. Compared to the SOTA representation learning approach FAMI-Pose [28], GLSMamba-B can provide a significant performance improvement of 1.9 mAP. On the other hand, in contrast to the best-performed post-processing method [48] ‡ that operates in the pose coordinate space, our GLSMamba-B still achieves a competitive performance of 97.9 mAP.

Qualitative analyses. In addition to the quantitative comparisons, we also qualitatively examine the ability of GLS-Mamba to cope with challenging scenes. As illustrated in Fig. 5, we present the side-by-side comparisons of the proposed method (a) against SOTA models TDMI [10] (b) and DiffPose [11] (c). Remarkably, our approach achieves more robust and accurate results across various scenarios. TDMI is built upon convolutions that suffer limited receptive fields, leading to suboptimal performance. On the other hand, DiffPose leverages self-attentions and overlooks rich keypoint motional details. Through the principled design of GSM and LRM, our method can capture reliable global-local high-resolution spatiotemporal representations and is more adept at handling complex cases.

Method	Global Spat. Mamba (GSM)	Local Ref. Mamba (LRM)	mAP
(a) Backbone			74.2
(b) GSM	✓		86.0
(c) GLSMamba-B	✓	✓	**86.9**

Table 5. Ablation study of different components.

Methods	#Params.	GFLOPs	mAP
(a) unified scanning	9.1 M	137.4	85.8
(b) unified + space-wise scanning	9.4 M	138.1	86.5
(c) unified + space-wise + time-wise scanning	9.8 M	138.9	**86.9**
(d) w/o STMM	9.1 M	137.4	86.2

Table 6. Ablation study of STS6D and STMM.

4.3. Ablation Study

In this section, we investigate the impact of each proposed component and design choice in GLSMamba-B. All experiments are performed on the PoseTrack2017 validation set.

Study on components. We first study the contribution of each individual component including Global Spatiotemporal Mamba (GSM) and Local Refinement Mamba (LRM), and provide the empirical results in Table 5. **(a)** For the first setting, we remove both proposed GSM and LRM modules, and estimate human poses employing only the backbone (ViT-B). This baseline obtains a 74.2 mAP. **(b)** Subsequently, we incorporate the GSM module on top of the backbone (a) for global dynamic modeling, which significantly improves upon the baseline by a large margin of 11.8 mAP and is on par with the SOTA approach DiffPose [11]. This corroborates the effectiveness of our method in introducing global spatiotemporal knowledge to facilitate VHPE. **(c)** For the final setting, we further introduce LRM which corresponds to the complete GLSMamba-B model.



Figure 5. Qualitative comparisons of pose predictions of (a) GLSMamba-B, (b) TDMI, and (c) DiffPose on the PoseTrack dataset. Inaccurate results are highlighted by red circles.

The performance improvement of 0.9 mAP suggests the importance of capturing enriched high-frequency details of local keypoint motions for accurate pose estimation.

Study on GSM designs. Then, we validate the efficacy of the core GSM designs, including the 6D selective Space-Time Scan (STS6D) and Spatial- and Temporal-Modulated scan Merging (STMM). **(1)** As presented in Table 6 (a)-(c), we gradually introduce diverse scanning directions containing unified, space-wise, and time-wise scanning routes. The results in mAP reflect a progressive and remarkable performance improvement, from $85.8 \rightarrow 86.5 \rightarrow 86.9$, with negligible extra computations. This is in line with our expectations, *i.e.*, an adequate space-time traversal allows for effective mining of dense high-resolution sequence knowledge, thereby contributing to enhanced accuracy. **(2)** We also examine the impact of the proposed STMM strategy by removing it and merging diverse scans via a simple addition. The significant performance reduction of 0.7 mAP (d) highlights the importance of STMM in adaptively aggregating distinct scanning knowledge.

Comparison with VideoMamba. We notice that VideoMamba [24, 33] has proposed the latest Mamba-based framework for high-level video understanding. However, our method differs notably from VideoMamba: (1) For the global modeling, we introduce a Sequential Channel Attention to filter unnecessary information, and design STS6D and STMM for adequate spatiotemporal scanning and adaptive fusion. (2) Unlike VideoMamba that lacks of the local modeling capability, we also propose a Windowed Space-Time Scan (WSTS) to enhance local details.

Spatiotemporal representation resolution. Finally, we examine the influence of feature resolutions on the pose estimation performance. As reported in Table 7, the following baselines are constructed: **a)** We directly adapt GLSMamba-B to low-resolution sequences ($\frac{1}{16}\mathtt{H} \times \frac{1}{16}\mathtt{W} \times \mathtt{T}$) which forms *GLSMamba-BLR**. **b)** We employ ViT-B as backbone and stack six standard ViT-B blocks to learn spatiotemporal features at low (*TransLR**), normal (*TransNR**), and high (*TransHR**) resolutions, respectively. It is observed that spatiotemporal representations with higher resolutions indeed result in better performance, across both Mamba (↑ 1.2 mAP) and Transformer (↑ 0.6 mAP) architectures. This is in line with our intuitions that high-resolution sequence representations can capture inter-frame temporal dynamics and intra-frame spatial details more precisely, which facilitate accurate pose heatmaps. Another observation is that high-resolution settings lead to significantly increased computational overhead, especially for Transformer structures (Out Of Memory (OOM) vs 138.9G FLOPs at $15,360$ input tokens). This highlights that Mamba can achieve better computational trade-offs in handling high-resolution feature sequences.

Method	Resolution	Token Num.	#Params.	GFLOPs	Mean
GLSMamba-B	$1/4 \times T$	$15,360$	**9.8** M†	**138.9**	86.9(↑ 1.2)
GLSMamba-BLR*	$1/16 \times T$	960	9.8 M†	85.1	85.7
TransLR*	$1/16 \times T$	960	46.3 M†	125.7	84.2
TransNR*	$1/8 \times T$	$3,840$	47 M†	315.2	84.8(↑ 0.6)
TransHR*	$1/4 \times T$	$15,360$	–	–	**OOM**
PoseWarper [3]	$1/4 \times T$	–	71.1 M†	210.5	81.2

Table 7. Impact of feature sequence resolutions. "†" denotes trainable parameters and "*" indicates manually-constructed baselines.

5. Conclusion and Future Works

This paper introduces GLSMamba, a novel framework that leverages State Space Models to learn decoupled global and local high-resolution spatiotemporal representations for VHPE. We design a Global Spatiotemporal Mamba with 6D selective space-time scan and spatial- and temporal-modulated scan merging mechanisms, to fully analyze holistic human dynamics embedded in dense high-resolution spatiotemporal contexts from a global perspective. A Local Refinement Mamba based on windowed space-time scan is further introduced for enhancing localized keypoint motion details. Extensive experiments on four benchmarks demonstrate the superiority of GLSMamba in both performance and computational trade-offs. Future works include applications for other vision tasks such as 3D human pose estimation and video segmentation.

6. Acknowledgements

This research was supported by the National Natural Science Foundation of China under Grant Nos. 62203184 and W2421093, and the International Cooperation Project of Jilin Province under Grant No. 20250205079GH. This research was also supported by the National Key R&D Program of China under Grant No. 2023YFF0905400 and the National Natural Science Foundation of China through Grant No. U2341229. This research was also supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-RS-2020-II201789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[2] Qian Bao, Wu Liu, Yuhao Cheng, Boyan Zhou, and Tao Mei. Pose-guided tracking-by-detection: Robust multi-person pose tracking. *IEEE Transactions on Multimedia*, 23: 161–175, 2020. 6

[3] Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, and Lorenzo Torresani. Learning temporal pose estimation from sparsely-labeled videos. In *Advances in Neural Information Processing Systems*, pages 3027–3038, 2019. 1, 3, 6, 7, 8

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[5] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020. 2

[6] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20963–20972, 2022. 6

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[8] Zhipeng Fan, Jun Liu, and Yao Wang. Motion adaptive pose estimation from compressed videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11719–11728, 2021. 6

[9] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 6

[10] Runyang Feng, Yixing Gao, Xueqing Ma, Tze Ho Elden Tse, and Hyung Jin Chang. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17131–17141, 2023. 1, 2, 3, 6, 7

[11] Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14861–14872, 2023. 1, 2, 3, 6, 7

[12] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2, 3, 4

[13] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33:1474–1487, 2020. 3

[14] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2, 3

[15] 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. 3

[16] Hengkai Guo, Tang Tang, Guozhong Luo, Riwei Chen, Yongchen Lu, and Linfu Wen. Multi-domain pose network for multi-person pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 6

[17] Jijie He and Wenwu Yang. Video-based human pose regression via decoupled space-time aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1022–1031, 2024. 2, 3, 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[19] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[20] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 6

[21] Sheng Jin, Wentao Liu, Wanli Ouyang, and Chen Qian. Multi-person articulated tracking with spatial and temporal embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5664–5673, 2019. 6

[22] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–433, 2018. 2

[23] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019. 2

[24] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024. 2, 3, 8

[25] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11313–11322, 2021. 2

[26] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model, 2024. 3

[27] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 525–534, 2021. 3, 6, 7

[28] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang Wang. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11016, 2022. 1, 3, 6, 7

[29] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. Lstm pose machines. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5207–5215, 2018. 6

[30] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long range language modeling via gated state spaces. *arXiv preprint arXiv:2206.13947*, 2022. 3

[31] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017. 2

[32] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6942–6950, 2019. 6

[33] Jinyoung Park, Hee-Seon Kim, Kangwook Ko, Minbeom Kim, and Changick Kim. Videomamba: Spatio-temporal selective state space model. *arXiv preprint arXiv:2407.08476*, 2024. 2, 3, 8

[34] Umer Rafi, Andreas Doering, Bastian Leibe, and Juergen Gall. Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In *European Conference on Computer Vision*, pages 36–52. Springer, 2020. 6

[35] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 5

[36] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022. 3

[37] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4220–4229, 2017. 6

[38] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 1, 2, 3, 6

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[40] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[41] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11088–11096, 2020. 6

[42] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 2, 6

[43] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977*, 2018. 6

[44] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 2, 3, 4, 7

[45] Yiding Yang, Zhou Ren, Haoxiang Li, Chunluan Zhou, Xinchao Wang, and Gang Hua. Learning dynamics via graph neural networks for human pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8074–8084, 2021. 6

[46] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-hrnet: A lightweight high-resolution network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10440–10450, 2021. 1

[47] Dongdong Yu, Kai Su, Jia Sun, and Changhu Wang. Multi-person pose estimation for pose tracking with enhanced cascaded pyramid network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 6

[48] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for 10× efficient 2d and 3d pose estimation. In *European Conference on Computer Vision*, pages 607–624. Springer, 2022. 6, 7

[49] Jiabin Zhang, Zheng Zhu, Wei Zou, Peng Li, Yanwei Li, Hu Su, and Guan Huang. Fastpose: Towards real-time pose estimation and tracking via scale-normalized multi-task networks. *arXiv preprint arXiv:1908.05593*, 2019. 6

[50] Yuexi Zhang, Yin Wang, Octavia Camps, and Mario Sznaier. Key frame proposal network for efficient pose estimation in videos. In *European Conference on Computer Vision*, pages 609–625. Springer, 2020. 6

[51] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 2, 3

[52] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 5