Event-guided Fusion-Mamba for Context-aware 3D Human Pose Estimation

Bo Lang
Lehigh University
bol221@lehigh.edu

Mooi Choo Chuah
Lehigh University
mcc7@lehigh.edu

Abstract

3D human pose estimation (3D HPE) is an important computer vision task with various practical applications. Researchers have proposed various deep learning-based methods for 3D HPE. However, the majority of such methods rely on lifting 2D pose sequence to 3D which do not perform well in challenging scenarios and are often computationally expensive. Such methods typically rely on 2D joint coordinates which do not provide much spatial context to solve ambiguity problem. In addition, merely relying on information extracted from RGB frames may miss temporal information and structural context. Thus, in this paper, we propose a framework that incorporates event stream as an additional input since event features provide such useful information. Moreover, instead of using 2D joint coordinates in pose sequence, our framework uses intermediate visual representations produced by off-the-shelf 2D pose detectors to implicitly encode joint-centric spatial context. Our new framework is a novel state space model (SSM)-based solution called Event-Guided Context Aware MambaPose (CA-MambaPose). In CA-MambaPose framework, we design a novel cross modality fusion mamba module to skillfully fuse the RGB and Event features. CA-MambaPose has lower computational cost due to the efficiency of Mamba blocks. We conduct extensive experiments to evaluate CA-MambaPose using two existing datasets. Our experimental results show that CA-MambaPose achieves better performance than SOTA methods.

1. Introduction

3D human pose estimation (3D HPE), a fundamental and challenging computer vision task, has attracted much attention from computer vision researchers in recent years. 3D HPE aims to precisely localize the 3D joints of individuals given monocular images or videos, serving as a crucial component for various applications including action recognition [25, 45], behavior monitoring [14], and human-robot interaction [21].

Recently, benefiting from SOTA performance of the existing 2D pose estimators [5, 35, 47], lifting-based methods which lift 2D skeleton sequences to 3D space has become the dominated methods in the 3D HPE task. Compared to raw RGB images, 2D human poses (as an intermediate representation) have two essential advantages. First, 2D joint coordinates provide highly position-relevant information to such lifting-based methods for localizing joints in 3D space. In addition, 2D coordinate representation which only requires $J \times 2$ (J = number of joints) is exceptionally lightweight in terms of memory cost. Such properties enable SOTA lifting-based methods to take advantage of extremely long-term temporal clues to improve accuracy. Significant advancements in deep learning approaches have been made so far, consistently improving performance [4, 6, 23, 43]. Recently, the transformer-based approaches [22, 53, 55] have demonstrated further improvement in 3D HPE.

However, recovering accurate 3D pose from 2D keypoints is still challenging due to depth ambiguity and self-occlusion in monocular data [20, 26]. Most lifting-based methods generally consists of two stages. In Stage 1, an off-the-shelf 2D pose estimator detects 2D pose joints for each input video frame, with a set of intermediate representations as byproducts, e.g., feature maps of varying resolutions. In Stage 2, the detected 2D pose sequence is lifted to 3D space, while such feature representations are discarded. Some problems naturally arise here: the (multi-scale) joint-centric spatial context encoded by these image feature maps is lost. We claim that the spatial context carrying crucial visual clues (e.g., occlusions, shade) also provides necessary information for 3D HPE. For example, depth ambiguity and self-occlusion related problem can be mitigated by utilizing spatial information encoded in the image features. Since 2D keypoints alone are unable to encode the spatial contextual information, existing lifting-based approaches have to depend on long-term temporal clues to alleviate ambiguities, which bring non-trivial computational costs.

In this paper, we design a novel state space model (SSM) based framework which incorporates both RGB & Event stream, named Event-guided Context-Aware MambaPose (CA-MambaPose). Unlike existing lifting-based

methods, we engage the lost intermediate visual representations learned by 2D pose detectors. The proposed method leverages both multi-resolution RGB & Event feature maps produced by backbone network in a sparse manner. Specifically, we extract informative spatial or temporal contextual features from two modality feature maps using deformable operations [8, 56] where the detected 2D joints serve as reference points. This helps mitigate the noise brought from background while avoiding heavy computation. Since event stream includes important dynamic information of moving objects with clear edge structures, and captures motion changes in the scene at an extremely high dynamic range and high temporal resolution. Currently, event stream has been used for from low-level vision (feature detection and tracking, optic flow, etc.) to high-level vision (segmentation, recognition, etc.) tasks but not for 3D HPE. We aim to encode joint-centric context from both RGB image and event stream, that promotes reducing ambiguity in 3D human pose estimation. How to effectively fuse features from two modalities is a non-trivial question. Existing modern multi-modal fusion approaches generally employ Transformers [1, 39] to fuse the cross-modality features. Its self-attention mechanism enables it to efficiently capture spatio-temporal relationships for this domain. However, Transformer-based cross-modality fusion is compute-intensive with a quadratic time and space complexity.

To this end, we propose a Cross Modality Fusion Mamba method, aiming to fuse features in a hidden state space, which might open up a new paradigm for cross-modality feature fusion. We are inspired by Mamba [13, 24, 54] with a linear complexity to build a hidden state space, which can extracts key information from source features and explores relationships between different modalities. Furthermore, we design a Pose-Context Feature Exchange module to exchange information between the extracted multi-level contextual features and the 2D joint embedding that encodes positional clues. Specifically, we develop a Selective-Scan based modeling approach for feature exchange to help reduce the domain gaps. Finally, a spatial transformer module is adopted to model spatial dependencies between human joints. As a result, CA-MambaPose shows encouragingly strong performance in 3D HPE benchmarks.

In summary, CA-MambaPose is a novel Event-guided method that leverages Mamba to directly capture multi-modality dependency for accurate 3D pose estimation. Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to propose the Mamba-based cross-modality fusion, to combine event stream and RGB image for context-aware 3D HPE using lifting-based paradigm. Such context awareness which includes both spatial and temporal clues is achieved by leveraging readily available visual representations learned by 2D pose detectors.
- We introduce a novel Cross Modality Fusion Mamba which enables deep fusion between RGB & Event features in a hidden state space and reduces domain gaps between cross-modality features. Further, we use a selective-scan based module to efficiently fuse informative context features with 2D joint embedding that provides positional clues.
- CA-MambaPose achieves new state-of-the-art results on two widely used video 3D HPE benchmarks, Human3.6M and MPI-INF-3DHP. Our model demonstrates significant performance improvements over other temporal methods that use many video frames.

2. Related works

2.1. Image-based 3D Pose Estimation

The image-based multi-person 3D pose estimation methods can be mainly divided into two types of paradigms: top-down [22, 29, 36] and bottom-up approaches [40, 48, 52].

Similar to 2D HPE, the top-down paradigm first conducts human detection, followed by performing single-person 3D pose estimation. For single-person, they predict 3D poses by learning 3D heatmaps [29], or estimating 2D poses via 2D pose estimator [33] and performing 2D-to-3D lifting [49]. For example, PoseNet [29] predicts the root depths of each person during the human detection stage, then estimates the 3D coordinates from 3D heatmaps. The bottom-up paradigm [40, 48, 52] follows a pipeline of firstly estimating the 3D coordinates for each human joint in an image and then assigning them to different human instances. Although these methods achieve great improvement on 3D human HPE, the performances of these methods rely on the accuracy of human detection. In addition, image-based 3D HPE methods are not good at handling occlusion cases compared with the video-based approaches.

2.2. Video-based 3D Pose Estimation

Video-based 3D human pose estimation methods [4, 49, 52, 53] can extract more temporal context to achieve better consistency for pose estimation across frames. Generally, there are two categories methods of extracting temporal information: 1. based on image visual features [7, 17] and 2. based on 2D pose sequence [2, 30, 53].

The methods [7, 17, 37] based on image visual features usually crop the human features through predicted human bounding boxes, and then use 3D convolution or RNN to extract the temporal information from these cropped sequence features. For example, TCMR [7] uses ResNet to extract visual features from video frames, then captures temporal dependency on these deep features by the RNN. However, these methods rely on such cropped image inputs that face the feature alignment problem. The methods [2, 30, 53]

based on 2D pose sequence usually estimate 2D joints sequence at first, then lift 2D coordinate sequence to 3D pose using a temporal lifting network. However, these methods cannot capture contextual depth information from visual features which have been lost during the 2D HPE processing stage. Besides, these video-based methods are multi-stage which cannot be optimized in an end-to-end manner.

2.3. Transformers in 3D Human Pose Estimation

Recently, the transformer-based approaches [27, 30, 34, 46, 53, 55] have been proposed to improve the long-term modeling capabilities of video sequence for 3D human pose estimation. TransPose [46] formulates human joints as visual tokens and captures the relationship between human joints via self-attention. PRTR [19] exploits the end-to-end transformer-based pose estimation network. PoseFormer [53] and MotionBERT [55] explore the spatial-temporal attention mechanism for 3D pose estimation. MotionAGFormer [27] introduces a new GCNFormer module that harnesses the power of transformers to capture global information while simultaneously employing Graph Convolutional Networks (GCNs) to integrate local spatial and temporal relationships. However, these methods did not study the attention on real visual features from images since they lift 3D poses from a sequence of 2D poses. Moreover, the existing transformer-based pose estimation methods are designed for single-person pose estimation, which limits their applications in crowded scenarios. Although POTR-3D [30] proposes three types of transformer to model single-person, inter-person and inter-frame relationships, they still follow the 2D-to-3D lifting paradigm and lose the contextual information from visual features. In this paper, we study an E2E video 3D pose estimation framework for either single-person or multi-person. Our work explores extracting spatial and temporal relationships in both spatial and channel branches under the event stream guidance.

2.4. State Space Models

Recently, Mamba [13] has achieved a significant breakthrough with its linear-time inference and efficient training methodology. Building on the success of Mamba, MoE-Mamba [32] amalgamated Mixture of Experts with Mamba, unlocking the scalability potential of SSMs and achieving performance akin to Transformers. For vision applications, Vision Mamba [54] and VMamba [24] used bidirectional SSM blocks and the cross-scan module, respectively, to enhance data-dependent global visual context. For example, vision Mamba [54] expands the original 2D scan to different bidirectional 3D scans and designs a Mamba framework to use mamba in video understanding tasks. However, the exploration of Mamba's potential in 3D human pose estimation remains untapped. In this paper, we do not simply apply SSM for pose estimation. We incorporate event stream for 3D HPE, thus we focus on exploiting Mamba for multi-modality feature fusion. We introduce a carefully designed Mamba-based structure to integrate the cross-modality features.

3. Methodology

3.1. Preliminaries

State Space Model. State Space Models (SSMs) can be considered as a linear time-invariant (LTI) system that maps a one-dimensional input sequence $x(t) \in \mathbb{R}^L$ to an output $y(t) \in \mathbb{R}^L$ via intermediate hidden states $h(t) \in \mathbb{R}^N$. Mathematically, SSMs are often formulated as linear ordinary differential equations (ODEs):

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t) + Dx(t) \end{aligned} \tag{1}$$

where the system's behavior is defined by a set of parameters, including the state transition matrix $A \in \mathbb{R}^{N\times N}$, the projection parameters $B, C \in \mathbb{R}^{N\times 1}$, and the weighting parameter $D \in \mathbb{R}^1$.

Discretization of SSM. The continuous-time nature of SSMs in Eq. 1 poses significant challenges when applied in deep learning scenarios. To address this issue, it is necessary to discretize the ODEs through a discretization process, which includes a timescale parameter Δ to transform the continuous parameters A, B to discrete parameters $\overline{A}, \overline{B}$. The commonly used method for transformation is zero-order hold (ZOH), which is defined as follows:

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

After the discretization of $\overline{A}, \overline{B}$, the discretized version of Eq. 1 using a step size Δ can be rewritten as:

$$\begin{aligned} h_t &= \overline{A}h_{t-1} + \overline{B}x_t, \\ y_t &= Ch_t + Dx_t \end{aligned} \tag{3}$$

After discretization, SSMs are computed via a global convolution with a structured convolutional kernel $\overline{K} \in \mathbb{R}^M$.

$$\begin{aligned} \overline{K} &= (C\overline{B}, C\overline{AB}, \cdots, C\overline{A^{M-1}B}), \\ y &= x * \overline{K} \end{aligned} \tag{4}$$

where M is the length of the input sequence x.

3.2. Overview

In this section, we give an overview of the proposed ContextAware(CA)-Mambapose. The whole framework is shown in Fig 1. CA-Mambapose takes two types of input namely RGB image I and event e_I. A pretrained 2D pose

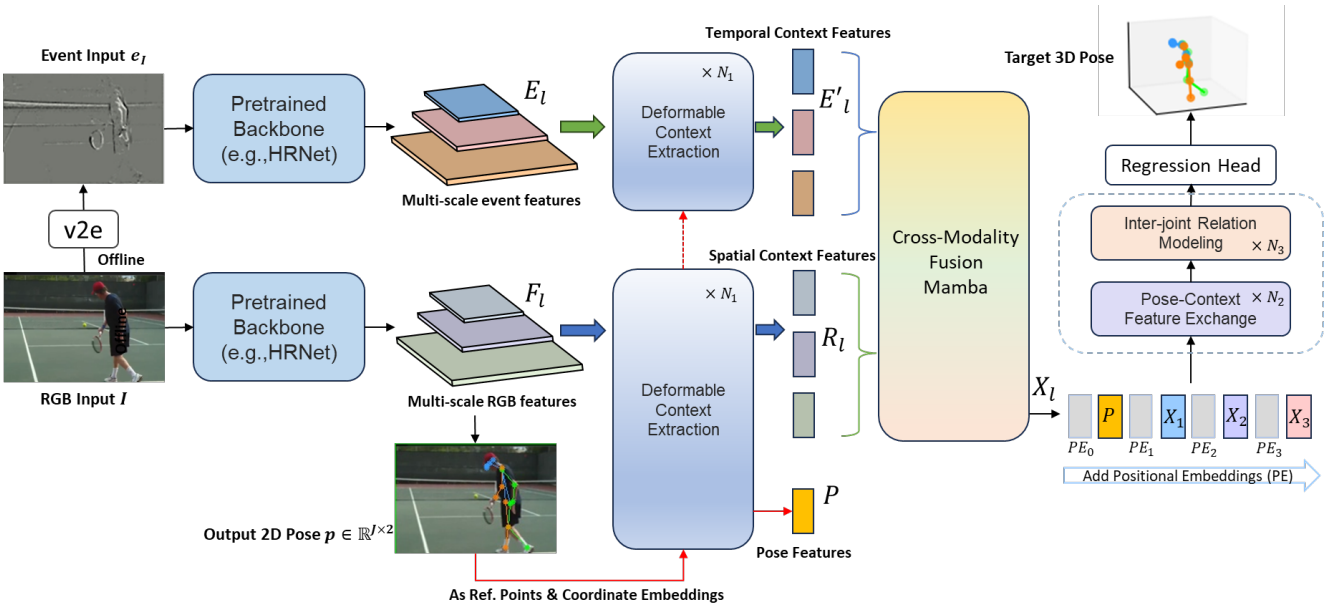

Figure 1. The overview of Event-guidend Context-Aware MambaPose (CA-MambaPose). CA-MambaPose takes two types of input which are RGB (I) and event (e_I) image. Event image is offline syntheticly generated using v2e [15] toolbox. A pretrained 2D pose detector estimates the 2D pose from RGB input with a set of multi-scale RGB feature maps F_l as byproducts. We use the same backbone to generate a set of multi-scale event feature maps E_l. Then, we extract informative join-context features from two types of feature maps via Deformable Context Extraction (Sec.3.3) and subsequently fuse them with a proposed Cross Modality Fusion Mamba (Sec.3.4) module. The fused features X_l with cross-modality information interact with 2D pose embeddings P via Pose-Context Feature Exchange (Sec.3.5) module. Finally, such joint-level context-aware representations are fed into Inter-joint Relation Modeling before inputing into the regression head for final 3D pose estimation.

estimator facilitates the extraction of multi-scale features from RGB and event images, denoted by F_l and E_l, respectively. It also estimates the 2D pose p from the multi-scale RGB features. After that, we feed both types of features with 2D pose as reference points into Deformable Context Extraction module to separately extract joint-context features. Subsequently, we input these contextual features into Cross Modality Fusion Mamba (CMFM), which reduces domain gaps between cross-modal features and enhances the representation consistency of fused features. After that, the fused features with cross-modality information interact with 2D pose embeddings via Pose-Context Feature Exchange module. Finally, we apply an Inter-joint Relation Modeling to extract inter-joint dependencies for each representation, before feeding into the regression head for final 3D pose estimation.

Next, we provide detailed descriptions of our approach. Since the spatial context carries crucial visual clues (e.g., occlusions, shade) to help solve the ambiguity problem in 3D HPE task, we aim to utilize the intermediate visual representations learned via 2D pose detector in our approach. Besides the informative spatial context encoded in RGB feature maps, DVS event stream includes important dynamic (temporal) information, reacting to changes in the scene with microsecond precision. This can be advantageous for capturing accurate temporal motion information which can be considered as temporal context to help 3D HPE.

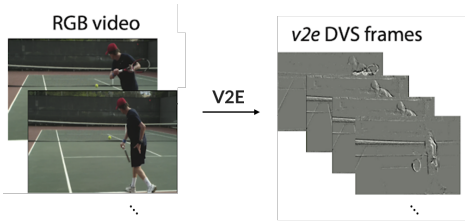

Figure 2. Event stream.

For a given RGB image I of size $H \times W \times 3$, we first simply use the v2e toolbox [15] to generate realistic synthetic event stream e_I for each RGB image I (shown in Fig 2). The events and frames of a hybrid camera system are hard to be perfectly aligned in practice. To take this into consideration, we apply random perspective transforms between them as in [18] during data preparation. An off-the-shelf (pretrained) 2D pose detector (e.g., HRNet) produces two sets of intermediate feature maps with varying resolutions for both inputs, $F_l \in \left\{\mathbb{R}^{H_l \times W_l \times C_l}\right\}_{l=1}^{L}$ from RGB image I (where L is the total number of feature maps) and $E_l \in \left\{\mathbb{R}^{H_l \times W_l \times C_l}\right\}_{l=1}^{L}$ from event e_I. Since high-

resolution feature maps encode fine-grained visual cues (e.g., joint position) while low-resolution ones tend to keep high-level semantics (e.g., human skeleton structures), we aim to take full advantage of these multi-scale feature maps to implicitly encode information about human joints. In addition, the 2D pose detector also estimates the corresponding 2D human pose $p \in \mathbb{R}^{J\times 2}$ merely based on multi-scale RGB features. Inspired by [50], we apply the Deformable Context Extraction module to further extract the spatial and temporal contextual features from both types of multi-scale feature maps using the 2D detected joints as reference points. This helps mitigate the noise brought from background while avoiding heavy computation.

How to effectively fuse the spatial contextual features and temporal contextual features from both input streams is a non-trivial question. Blindly processing them in a global way like e.g., convolutions, or vision transformers, may bring unnecessary computational overhead. Since Mamba [13] was proposed for linear-time sequence modeling in the NLP field, it has been rapidly extended in various computer vision tasks. We propose a Cross Modality Fusion Mamba (CMFM) module which incorporates bidirectional SSM to fuse the spatial contextual features and temporal contextual features in a hidden state space. Moreover, we design a Pose-Context Feature Exchange module to fuse the contextual features and 2D joint embeddings that encode positional clues about human joints. Specifically, we propose a Selective-Scan based modeling approach for feature exchange to help reduce the domain gaps. The key components will be explained in the following sub-sections.

3.3. Deformable Context Extraction

This module uses deformable attention to extract informative spatial/temporal contextual cues from RGB/event feature maps. Specifically, for each detected 2D joint, we produce a set of sampling points on multi-scale feature maps whose offsets and weights are learned based on the features of reference points (i.e., the detected joint of interest). We also add the position embedding of 2D joint coordinates P_j to the source features to preserve position detail. In this way, we can sample feature vectors not only at the detected joints but also the regions around them. Let l index a feature level and j index a human joint, and Deformable Context Extraction is formulated as:

$$
\begin{aligned}
&R'^{n}_{lj} = \text{DeformAttn}(R^{n-1}_{lj} + P_j) + R^{n-1}_{lj} \\
&R^{n}_{lj} = MLP(R'^{n}_{lj}) + R'^{n}_{lj} \\
&\text{DeformAttn}(R^{n-1}_{lj} + P_j) = \\
&\sum_{m=1}^{M}\left[\sum_{k=1}^{K} A_{lmk}\cdot W_{lm}F_l(p_j + \Delta p_{lmk})\right]
\end{aligned}
\tag{5}
$$

where m iterates over the attention heads, $n \in (1\cdots N_1)$ represents each layer, k over the sampled points around the detected joint p_j, and K is the total sampling point number. For the l^{th} feature map, Δp_{lmk} represents the sampling offset of the k^{th} sampling point in the m^{th} attention head, while A_{lmk} denotes its corresponding attention weight. Given different source features F_l or E_l from RGB image and event stream, we can extract the contextual features $\{R\}_{l=1}^{L}$ and $\{E'\}_{l=1}^{L}$ containing either spatial or temporal clues of joints.

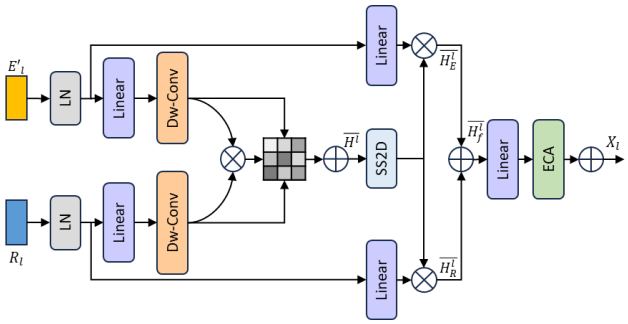

Figure 3. Cross Modality Fusion Mamba Module (CMFM)

3.4. Cross Modality Fusion Mamba

In contrast to prior methods using attention mechanisms with quadratic computational complexity, we propose a state space model to capture comprehensive spatial-temporal relationship. The features output by the Deformable Context Extraction are further fed into the Cross Modality Fusion Mamba (CMFM) module (Fig 3) for fine-grained fusion and exploration of information correlation between different modalities. The contextual features $\{R\}_{l=1}^{L}$ and $\{E'\}_{l=1}^{L}$ from both RGB and event stream are first mixed to generate the mixed features $\overline{H}^n$:

$$
\overline{H^l} = \text{Dwc}\left(\text{Linear}(R_l)\right)\otimes\text{Dwc}(\text{Linear}(E'_l))\oplus R_l \oplus E'_l \tag{6}
$$

where Dwc$(\cdot)$ is the Depthwise convolution operation. $\otimes$ and $\oplus$ are the Element-wise multiplication and addition operation. These hybrid contextual features are then input into a **bidirectional** SS2D (2D-Selective-Scan) layer to capture the spatial long-term dependencies.

$$
\begin{aligned}
H^l_R &= \text{LN}\left(\text{SS2D}(\text{SiLU}(\overline{H^l}))\right)\otimes\text{SiLU}(\text{Linear}(R_l)), \\
H^l_E &= \text{LN}\left(\text{SS2D}(\text{SiLU}(\overline{H^l}))\right)\otimes\text{SiLU}(\text{Linear}(E'_l)), \\
H^l_f &= H^l_R \oplus H^l_E
\end{aligned}
\tag{7}
$$

To enhance the expressive power of different channels, we integrate Efficient Channel Attention (ECA) [44] into our CMFM module. This allows SS2D to concentrate on learning diverse channel representations, with subsequent channel attention selecting critical channels to prevent redun-

dancy. The output contextual features H_l^n pass through the ECA module, resulting in the final fused feature map X_l:

$$X_l = \text{ECA}\left(\text{LN}(H_f^l)\right) \oplus R_l \oplus E_l' \tag{8}$$

3.5. Pose-Context Feature Exchange

This module aims at exchanging information between 2D pose embeddings and the fused multi-level contextual features simultaneously. Instead of using a unified transformer encoder to model interactions between pose embedding and multi-level context features, we apply SSM to learn a joint representation for both types of features. In contrast to the self-attention mechanism in transformer, SSM ensures that each feature token gains contextual knowledge exclusively through a compressed hidden state computed along the corresponding scanning path, thereby reducing the computational complexity from quadratic to linear. The implementation of this module is:

$$\begin{aligned} Y_j^0 &= \text{Concat}([P_j, X_{1j}, \cdots X_{Lj}], \text{dim} = 0) \\ Y_j^{'n} &= \text{SSM}(Y_j^{n-1}) + Y_j^{n-1} \\ Y_j^n &= \text{MLP}(Y_j^{'n}) + Y_j^{'n} \end{aligned} \tag{9}$$

where $j \in (1 \cdots J)$ indicates that Pose-Context feature exchange is performed for each joint and $n \in (1 \cdots N_2)$ represents each layer. Mamba layers reduce domain gaps for both types of features in shared hidden space and promote message-passing where joint-position information and multi-level contextual cues complement each other.

3.6. Inter-joint Relation Modeling

With the three modules above, elegant local representations are learned for each joint individually. To understand the human skeleton system and its corresponding spatial context in a global view, inter-joint dependencies are modeled based on the learned per-joint features using a spatial transformer encoder as in PoseFormer [53]. The spatial encoder takes our joint-level context-aware representation $\left\{Y_j^{N_2}\right\}_{j=1}^{J}$ as input, where each joint token (J in total) is of $(L+1) \times C$ dimensions, encoding both positional and contextual information for the related joint.

3.7. Output and Loss Function

Since we already have each joint-level representation, a simple MLP layer is adopted to obtain the final 3D pose $y \in \mathbb{R}^{j \times 3}$. We use L2 loss to minimize the error between the predicted and ground truth pose as:

$$\mathcal{L} = \frac{1}{J} \sum_{j=1}^{J} \|y_j - \hat{y_j}\|_2 \tag{10}$$

where $\hat{y_j}$ and y_j are the ground truth and estimated 3D joint locations of the j_{th} joint, respectively.

4. Experiments

In this section, we present the experimental results of ContextAware-Mambapose. We first introduce the implementation details of Mambapose, and then report results and compare with SOTA methods using two widely-used single-person datasets: Human3.6M [16], MPI-INF-3DHP [28]. All ablation studies are based on Human3.6 dataset.

4.1. Implemental Details

We use HRNet-32 [42] pre-trained on ImageNet [9] as the backbone network $\varnothing$ of CA-Mambapose for all experiments and follow the most configuration of [41]. In our experiments, CA-Mambapose is trained on 4 A100 GPUs with a batch size of 10 frames/GPU, while the input size is 512×512. The total number of training epochs is 60. Adam optimizer is adopted and the initial learning rate is 5e-4, which decreases 10× at 40 and 50 epochs.

4.2. Datasets and Evaluation Metric

Human3.6M dataset. Human3.6 [16] is the largest indoor benchmark for single-person 3D pose estimation, which includes 7 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 previous works [27, 34, 55], we use two protocols for evaluation. The first protocol (referred to as P1) uses Mean Per Joint Position Error (MPJPE) in millimeters that measures the error 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. MPI-INF-3DHP [28] is another large-scale dataset gathered in three different settings: green screen, non-green screen, and outdoor environments. Following previous works [34, 38], MPJPE, Percentage of Correct Keypoint (PC) within the 150 mm range, and Area Under the Curve (AUC) are reported as evaluation metrics.

4.3. Comparison with the State-of-the-art Methods

4.3.1 Results on Human3.6M

The comparisons with state-of-the-art methods on the Human3.6M dataset are shown in Table1. Our single-frame CA-Mambapose achieves new state-of-the-art results with an MPJPE of 36.5mm and a PA-MPJPE of 28.6mm in Protocol 1 and Protocol 2, respectively. The results demonstrate the effectiveness of the proposed CA-Mambapose. Compared with other transformer-based methods [27, 55], CA-Mambapose outperforms them by a large margin. Even these models are based on a larger frame number above 81,

Protocol 1		T	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.↓
Pavllo et al. [31]	CVPR'19	243	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Liu et al. [23]	CVPR'20	243	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
Zheng et al. [53]	ICCV'21	81	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
Li et al. [20]	CVPR'22	351	39.2	43.1	40.1	42.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
Einfalt et al. [10]	WACV'23	351	39.6	43.8	40.2	42.4	46.5	53.9	42.3	42.5	55.7	62.3	45.1	43.0	44.7	30.1	30.8	44.2
Tang et al. [38]	CVPR'23	243	39.6	41.6	37.4	38.8	43.1	51.1	39.1	39.7	51.4	57.4	41.8	38.5	40.7	27.1	28.6	41.0
Foo et al. [11]	CVPR'23	243	37.5	39.2	36.9	40.6	39.3	46.8	39.0	41.7	50.6	63.5	40.4	37.8	44.2	26.7	29.1	40.8
Zhu et al. [55]	ICCV'23	243	36.3	38.7	38.6	33.6	42.1	50.1	36.2	35.7	50.1	56.6	41.3	37.4	37.7	25.6	26.5	39.2
Gong et al. [12]	CVPR'23	243	33.2	36.6	33.0	35.6	37.6	45.1	35.7	35.5	46.4	49.9	37.3	35.6	36.5	24.4	24.1	<u>36.9</u>
Mehraban1 et al. [27]	WACV'24	243	36.4	38.4	36.8	32.9	40.9	48.5	36.6	34.6	51.7	52.8	41.0	36.4	36.5	26.7	27.0	38.4
Ours(CA-Mambapose)		1	33.1	36.0	33.3	34.5	37.8	44.7	35.2	34.9	46.9	49.7	40.1	35.3	36.1	25.1	25.7	**36.5**
Protocol 2		**T**	**Dir.**	**Disc.**	**Eat.**	**Greet**	**Phone**	**Photo**	**Pose**	**Purch.**	**Sit**	**SitD.**	**Smoke**	**Wait**	**WalkD.**	**Walk**	**WalkT.**	**Avg. ↓**
Pavllo et al. [31]	CVPR'19	243	34.1	36.1	34.4	37.2	36.4	42.4	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Wang et al. [43]	ECCV'20	96	32.9	35.2	35.6	34.4	36.4	42.7	31.2	32.5	45.6	50.2	37.3	32.8	36.3	26.0	23.9	35.5
Zheng et al. [53]	ICCV'21	81	32.5	34.8	32.6	34.6	35.3	39.5	32.1	32.0	42.8	48.5	34.8	32.4	35.3	24.5	26.0	34.6
Tang et al. [38]	CVPR'23	243	29.5	33.2	30.6	31.0	33.0	38.0	30.4	29.4	41.8	45.5	33.6	29.5	31.6	21.3	22.6	<u>32.0</u>
Foo et al. [11]	CVPR'23	243	30.3	32.2	30.8	33.1	31.1	35.5	30.3	32.1	39.4	49.6	32.9	29.2	33.9	21.6	24.5	32.5
Zhu et al. [55]	ICCV'23	243	30.8	32.8	32.4	28.7	34.3	38.9	30.1	30.0	42.5	49.7	36.0	30.8	22.0	31.7	23.0	32.9
Mehraban1 et al. [27]	WACV'24	243	30.6	32.6	32.2	28.2	33.8	38.6	30.5	29.9	43.3	47.0	35.2	29.8	31.4	22.7	23.5	32.6
Ours(CA-Mambapose)		1	28.1	30.9	28.2	30.3	29.7	32.5	29.3	28.4	32.1	34.9	31.8	27.6	23.4	20.2	21.4	**28.6**

Table 1. Quantitative comparison with state-of-the-art methods on Human3.6M under Protocol 1 (MPJPE) and Protocol 2 (PA-MPJPE). T denotes the number of input frames used in each method. Bold indicates the best and underline indicates the second best.

CA-Mambapose with only single frame obtains better results since most of the approaches in Table 1 follow the 2D-to-3D lifting paradigm which loses the visual contextual feature extracted from the 2D pose estimation.

Methods	T	PCK ↑	AUC ↑	MPJPE ↓
Li et al. [20]	9	93.8	66.3	58.0
Einfalt et al. [10]	81	95.4	67.6	46.9
Zhao et al. [51]	81	97.9	78.8	27.8
Tang et al. [38]	81	<u>98.7</u>	83.9	23.1
Chen et al. [3]	96	<u>98.7</u>	72.9	37.2
Gong et al. [12]	81	98.0	75.9	29.1
Ours(CA-Mambapose)	1	**99.1**	**84.0**	**20.2**

Table 2. Quantitative comparison with state-of-the-art methods on MPI-INF-3DHP. T: Number of input frames. Bold indicates the best and underline indicates the second best.

4.3.2 Results on MPI-INF-3DHP

In evaluating our method on the MPI-INF-3DHP dataset, we also use HRNet-32 as the backbone network to generate multi-resolution feature maps. As shown in Table 2, across all metrics, our method consistently outperforms others in terms of MPJPE. Notably, our CA-Mambapose achieves remarkable results with an 84.0% AUC and a 20.2 mm P1 error. This outperforms the previous 2nd-best STCFormer [38] by a significant margin of 1% in AUC and 2.9 mm in P1 error. Besides, it achieves 99.1% PCK, which is 0.4% better than the PCK performance of the 2nd-best models. Our approach has the access to temporal information from event stream and does not need an extra pre-training stage. The results verify the generalization ability of our method to different datasets, particularly in challenging outdoor environments.

4.4. Ablation Studies

In this section, we verify the effectiveness of the proposed Cross Modality Fusion Mamba, Pose-Context Feature Exchange, in CA-Mambapose.

4.4.1 Effectiveness of proposed sub-modules

We conduct the ablation study on Human3.6m dataset to verify the effectiveness of each component in our method. First, we show why context-awareness is important. Next, we show the effectiveness of each individual module in CA-MambaPose (the results are shown in Table 3)).

- **Base, Context-Agnostic:** The baseline follows PoseFormer [53] where 2D joint coordinates estimated by a 2D pose detector are then further projected to a high dimension C as joint embeddings, and Inter-joint Modeling (Transformer encoder) is subsequently performed to determine correlations across the joint embeddings. Such plain 2D coordinates contain no spatial context for joints, is thus referred to as "context-agnostic". This serves as the baseline of our method.

- **Exps 0-1, Context-Aware with Pose-Context Feature Exchange:** The key of "context-aware" is to incorporate joint-context features into each per-joint representation. We apply the Deformable Context Extraction to simply sample feature vectors on the detected joint locations from multi-scale feature maps, and project them to a high dimension C. For each joint, we first promote pose-context feature exchange by applying transformers to joint embeddings and the sampled multi-scale context features before concatenation. As shown in the 2nd row, merely using pose-

Exp	Deform. Cont. Extraction	Cross Modality Feature Fusion	Pose-Context Feature Exchange	Inter-joint Relation Modeling	FLOPs(M)	MPJPE↓
Base	✗	✗ (w/o Event)	✗	✓	446.0	51.2
(0)	✓	✗ (w/o Event)	self-attention	✓	609.9	41.4
(1)	✓	✗ (w/o Event)	SSM	✓	557.5	40.8
(2)	✓	concat	self-attention	✓	632.7	39.7
(3)	✓	cross-attention	self-attention	✓	702.3	38.4
(4)	✓	CMFM	self-attention	✓	644.5	36.9
(5)	✓	CMFM	SSM	✓	615.1	36.5

Table 3. Ablation study on each component of our method. Experiments are conducted on Human3.6M with HRNet-32 as the backbone. MPJPE is reported in millimeters.

context feature exchange provides a 19.1% error reduction (from 51.2 to 41.4mm), which demonstrates that leveraging the readily available visual representations is effective and promising. Since the quadratic complexity of self-attention presents a significant challenge when dealing with long input tokens, we replace it to SSM using selective Scan Mechanism. In Exp 1, using SSM decreases MPJPE by 0.6mm and saves huge computational cost.

- **Exps 2-4, Incorporating event features:** Multi-scale features from RGB input provide spatial context with visual cues. Event streams also provide visual information at a high-dynamic range and with strong robustness against motion blur. These unique properties offer great potential for motion analysis. We claim that incorporating event stream containing temporal context and clear structural information can be beneficial. We investigate three different cross-modality fusion methods to incorporate event features and show the results in the 4-6th rows. Even naive involvement (as simple as concatenation) of both features brings an obvious improvement (from 41.4 to 39.7, ↓4.1%). Next, we apply transformers (cross-attention) to sampled multi-scale RGB features and event features, which brings another 3.3% error reduction. Consider the quadratic computational complexity of attention mechanism, we develop our own Cross Modality Fusion Mamba (CMFM) module to capture the spatial-temporal relationship. As shown in the 6th row, our CMFM further significantly reduces the MPJPE by 1.5mm and also improves the efficiency, which shows the effective guidance provided by event stream. These results demonstrate that CMFM of CA-MambaPose is effective for cross-modality feature fusions.

- **Exp 5, Final version of our Context-Aware MambaPose:** Our Context-Aware MambaPose includes all the sub-modules. CA-Mambapose further reduce MPJPE to 36.5mm without a large increase of FLOPs. Thus, the usage of event with a novel mamba fusion module does not have a significant impact on model parameters and runtime, while it significantly improves the pose estimation performance.

5. Conclusion and Discussion

In this paper, we propose a novel Event-guidend Context-Aware MambaPose (CA-MambaPose) for 3D human pose estimation. We leverages readily available visual representations from both RGB images and event streams, which are learned by off-the-shelf 2D pose detectors. The spatial context and temporal context from two modalities provide crucial visual clues to improve the pose estimation ability. We propose a novel Cross Modality Fusion Mamba method for multi-modal feature fusion, which effectively extracts correlation between modalities while suppressing redundant information. To further exchange information within 2D joint embedding and contextual features, we introduce a selective-scan based Pose-Context Feature Exchange module which avoids quadratic complexity of attention mechanism. Combined with all these components, CA-MambaPose outperforms the state-of-the-art methods on two 3D human pose estimation benchmarks with higher accuracy and lower computational costs.

Limitations. We observed in Sec.4.3 that incorporating the temporal context from event stream improves temporal stability, enhancing consistency and smoothness in the estimated results, even without access to long-term video input. However, we acknowledge for all single-frame methods, including ours, mitigating jitters remains a challenge compared to multi-frame methods that leverage more temporal clues. A potential solution is to extend our method to model short-term temporal dependencies, which should not introduce unacceptably high costs. We could achieve this by incorporating a Temporal Transformer to our method, where Temporal Transformer models temporal correlations of each joint across frames independently.

References

[1] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 2

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 2

[3] Hanyuan Chen, Jun-Yan He, Wangmeng Xiang, Zhi-Qi Cheng, Wei Liu, Hanbing Liu, Bin Luo, Yifeng Geng,

and Xuansong Xie. Hdformer: High-order directed transformer for 3d human pose estimation. *arXiv preprint arXiv:2302.01825*, 2023. 7

[4] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):198–209, 2021. 1, 2

[5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 1

[6] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10631–10638, 2020. 1

[7] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1964–1973, 2021. 2

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2903–2913, 2023. 7

[11] Lin Geng Foo, Tianjiao Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. Unified pose sequence modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13030, 2023. 7

[12] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13041–13051, 2023. 7

[13] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2, 3, 5

[14] Lorna Herda, Pascal Fua, Ralf Plänkers, Ronan Boulic, and Daniel Thalmann. Using skeleton-based tracking to increase the reliability of optical motion capture. *Human movement science*, 20(3):313–341, 2001. 1

[15] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1312–1321, 2021. 4

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 6

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 2

[18] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7652–7661, 2020. 4

[19] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1944–1953, 2021. 3

[20] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 1, 7

[21] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10166–10175, 2020. 1

[22] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 1, 2

[23] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5064–5073, 2020. 1, 7

[24] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 2, 3

[25] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020. 1

[26] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Hai Ci, and Yizhou Wang. Context modeling in 3d human pose estimation: A unified perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6238–6247, 2021. 1

[27] Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6920–6930, 2024. 3, 6, 7

[28] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian

Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 6

[29] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10133–10142, 2019. 2

[30] Sungchan Park, Eunyi You, Inhoe Lee, and Joonseok Lee. Towards robust and smooth 3d multi-person pose estimation from monocular videos in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14772–14782, 2023. 2, 3

[31] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 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*, pages 7753–7762, 2019. 7

[32] Maciej Pióro, Kamil Ciebiera, Krystian Król, Jan Ludziejewski, and Sebastian Jaszczur. Moe-mamba: Efficient selective state space models with mixture of experts. *arXiv preprint arXiv:2401.04081*, 2024. 3

[33] Zhongwei Qiu, Kai Qiu, Jianlong Fu, and Dongmei Fu. Dgcn: Dynamic graph convolutional network for efficient multi-person pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11924–11931, 2020. 2

[34] Zhongwei Qiu, Qiansheng Yang, Jian Wang, and Dongmei Fu. Ivt: An end-to-end instance-guided video transformer for 3d pose estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6174–6182, 2022. 3, 6

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

[36] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021. 2

[37] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5349–5358, 2019. 2

[38] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4790–4799, 2023. 6, 7

[39] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[40] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 242–259. Springer, 2020. 2

[41] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11060–11068, 2022. 6

[42] 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*, 43(10):3349–3364, 2020. 6

[43] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European conference on computer vision*, pages 764–780. Springer, 2020. 1, 7

[44] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020. 5

[45] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1

[46] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11802–11812, 2021. 3

[47] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 1

[48] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. *Advances in neural information processing systems*, 31, 2018. 2

[49] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 507–523. Springer, 2020. 2

[50] Qitao Zhao, Ce Zheng, Mengyuan Liu, and Chen Chen. A single 2d pose with context is worth hundreds for 3d human pose estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[51] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8877–8886, 2023. 7

[52] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, Au-*

gust 23–28, 2020, Proceedings, Part XV 16, pages 550–566. Springer, 2020. 2

[53] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11656–11665, 2021. 1, 2, 3, 6, 7

[54] 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

[55] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. 1, 3, 6, 7

[56] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2