

# Event-guided Video Transformer for End-to-end 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. However, 3D pose estimation for multi-person from a monocular video (3DMPPE) is particularly challenging. Recent transformer-based approaches focus on capturing the spatial-temporal information from sequential 2D poses, which unfortunately loses the visual feature relevant for 3D pose estimation. In this paper, we propose an end-to-end framework called Event Guided Video Transformer (EVT) which predicts 3D poses directly from video frames by learning spatial-temporal contextual information from visual features effectively. In addition, our design is the first that incorporates event features to help guide 3D pose estimation. EVT first decouples persons into different instance-aware feature maps from video frames. These features containing specific clues of body structure information are then fed together with event features into an attention based Event-Aware Embedding Module. Next, the fused features for each instance are then fed into an intra-human relation extraction module and subsequently to a temporal transformer to extract inter-frame relationship. Finally, the extracted features are fed into a decoder for 3D pose estimation. Experiments using three widely used 3D pose estimation benchmarks show that our proposed EVT achieves better performance than state-of-the-art models.*

## 1. Introduction

3D human pose estimation (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 [26, 47], behavior monitoring [12], and human-robot interaction [22]. However, directly estimating 3D human poses from monocular 2D images or videos is notably challenging due to the lack of depth information.

To address this problem, image-based approaches employ diverse strategies to learn depth information from im-

age. Some methods utilize depth map supervision [40] or 3D heatmap supervision [29] to extract depth information from image features, while other image-based approaches [49, 51] firstly estimate 2D pose from image, and then lift 2D pose to 3D pose. Additionally, temporal smoothing methods [16] integrate post-processing modules to align the shapes estimated by image-based methods [8, 36, 37]. However, these frameworks exhibit limitations in capturing spatial relationships among human instances in an image and cannot extract long-term global dependency effectively. Besides, the computational cost is expensive since it is proportional to the number of human instances in an image and it needs an extra tracker to identify each instance in a video sequence. These multi-stage methods split space and time dimensions and can not be end-to-end optimized. Thus, video-based approaches [1, 2, 6, 54] have been suggested in recent years.

Benefiting from the excellent performance of the SOTA 2D pose estimators, many video-based approaches follow the 2D-to-3D lifting paradigm which infers 3D human poses from intermediately estimated 2D poses. These approaches typically surpass direct estimation methods. However, lifting these 2D poses into 3D poses is non-trivial since the same 2D pose may yield various potential 3D poses due to depth ambiguity and occlusion. To overcome some of these challenges, many recent works have exploited spatial and temporal modeling methods to improve the performance of 3D pose estimation. [5, 24] leverage convolutional neural networks (CNN) and recurrent neural networks (RNN) to capture global dependencies across adjacent frames. However, constrained by the capabilities of CNN and RNN, such models can hardly capture the long-term dependency in both spatial and temporal dimensions. To mitigate this limitation, some works [2, 6, 43] build upon graph neural network (GNN) to exploit spatial-temporal information between keypoints, which can capture both short-term and long-term dependency. Recently, the transformer-based approaches [23, 54, 55] have been proposed to extract more effective representative features which can improve the performance of video 3D pose estimation significantly.

Although the above GNN or transformer-based methods

can achieve significant improvement for 3D HPE, they still did not break out of the 2D-to-3D lifting paradigm. This paradigm solely relies on the 2D pose structure for depth estimation, while disregarding the contextual depth information contained in the visual features of video frames, as these features have been lost during the 2D pose prediction stage. Current researchers [7, 34, 38] believe that the context depth information embedded in the visual feature is more effective than the 2D pose structure for 3D pose estimation. [7, 38] firstly conduct human detection and then predict 3D pose directly from the cropped video patches, which can be regarded as exploiting contextual depth information from the visual feature to some extent. But they apply RNN to extract temporal features between adjacent video frames, which is not effective compared with GNN or transformer-based spatial-temporal modeling methods mentioned above. Besides, the inputs of cropped patches introduce a new problem of keypoints feature alignment.

To handle the above problems, we propose an end-to-end framework for video 3D HPE which develops a transformer-based model to fully use the spatial-temporal visual features and predicts 3D pose directly from video frames. Our proposed framework has two unique design: First, inspired by [41], we effectively decouple instances while preserving rich context cues from visual features to estimate 3D pose. Second, our framework is the 1st approach to incorporate event stream information 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.

Our E2E design, Event-Guided Video Transformer (EVT) models the spatial-temporal information for each decoupled instance under the guidance of event stream. First, Event-Guided Video Transformer (EVT) utilizes Contextual Instance Decoupling (CID) to decouple the global deep visual features into a set of instance-aware feature maps, where each map preserves contextual cues to infer his/her 3D pose. These instance-aware feature maps are fused with event stream in our proposed Event-Aware Embedding Module (EEM) which utilizes a dual-branch (spatial and channel) attention mechanism. Such dual-branch mechanism enables each instance feature to be refined via the guidance of event features. Furthermore, we sequentially model a specific spatial relationship using the Intra-human Relation Extraction (IRE) module and the inter-frame temporal correlation within a decoupled instance feature map using a Temporal Transformer (TT). As a result, EVT enables effective spatial-temporal feature extraction and yields better video 3D HPE performance.

In summary, EVT is a unified framework that is suitable for both single-person and multi-person video 3D HPE tasks. To the best of our knowledge, EVT is the first Event-guided E2E method that leverages transformer to directly capture spatial-temporal dependency for multi-person in video with the guidance from the event stream. Our contributions can be summarized as follows:

- We propose Event-Guided Video Transformer (EVT), a novel end-to-end transformer-based framework combining event stream and RGB image for both single-person and multi-person video 3D pose estimation.
- We design a novel Event-Aware Embedding Module (EEM) to enable effective fusion of both visual features and event features. EEM uses a dual-branch attention mechanism to learn helpful temporal and structural information from these two types of features.
- EVT achieves new state-of-the-art results on three widely used video 3D HPE benchmarks, Human3.6M, MPI-INF-3DHP, and CMU Panoptic.

## 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 [23, 29, 36] and bottom-up approaches [40, 50, 53].

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 by 2D pose estimator [33] and performing 2D-to-3D lifting [51]. The bottom-up paradigm [40, 50, 53] 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, they are not good at handling occlusion cases as the video-based approaches.

### 2.2. Video-based 3D Pose Estimation

Video-based 3D human pose estimation methods [5, 51, 53, 54] can extract more temporal context to ensure consistency for pose estimation across frames. Generally, the ways of extracting temporal information can be divided into two categories: based on image visual features [7, 17] and based on 2D pose sequence [1, 31, 54].

The methods [7, 17, 38] 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 sequences features. However, these methods essentially conduct single-person video 3D HPE, which results in feature alignment problem of the cropped image inputs. The methods [1, 31, 54] based on 2D joint coordinates usually estimate a sequence of 2D poses at first, then lift 2D coordinates sequence to 3D pose by a temporal lifting network. However, these methods cannot capture contextual depth information from visual features which have been dropped during the 2D HPE stage. Besides, these video-based methods are multi-stage which are limited to the human detector and cannot be optimized in an end-to-end manner.

### 2.3. Transformers in 3D Human Pose Estimation

Recently, the transformer-based approaches [27, 31, 34, 48, 54, 55] have been proposed to improve the long-term modeling capabilities of sequence for video 3D human pose estimation. PRTR [20] exploits the end-to-end transformer-based pose estimation network. PoseFormer [54] and MotionBERT [55] explore the spatial-temporal attention mechanism for 3D pose estimation. 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 on crowded scenarios. Although POTR-3D [31] 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.

## 3. Methodology

### 3.1. Overview

In this section, we elaborate on the details of the proposed Event-guided Video Transformer (EVT). The framework of EVT is shown in Fig 1. Given a sequence of video frames  $I = \{I_t \mid t \in [1, T]\}$ , EVT first extracts deep image features using a backbone network  $\emptyset$ . Next, the features are fed into the Instance Abstraction (IA) module to learn root joints heatmap and instance 2D tracking offsets (from  $t$  to  $t - 1$ ). Subsequently, for each frame, we apply Contextual Instance Decoupling (CID) (Sec. 3.2) to decouple the global deep features into a set of instance-aware feature maps, where each map represents context cues of a specific person to infer his/her 3D pose. To ensure the temporal consistency, we associate the instance feature maps belonging to the same instance across T frames using greedy matching

via predicted instance 2D tracking offsets.

After conducting CID, instance-aware feature maps and event features are sent into an Event-Aware Embedding Module (EEM) (Sec. 3.3) to help capture dynamic information in video sequence. Then, we feed the updated instance-aware feature maps containing event guidance into an Intra-human Relation Extraction (IRE) (Sec. 3.4) to further learn spatial relationships within a single person. In addition, we build a Temporal Transformer (TT) (Sec. 3.5) to effectively model the inter-frame temporal correlation for video 3D HPE. Finally, the output visual features are fed into a decoder for infer 3D poses of each person.

### 3.2. Contextual Instance Decoupling (CID)

Given each frame  $I_t$  of the video sequence, we first extract deep features  $\mathcal{F}_t \in \mathbb{R}^{C \times H \times W}$  from an image backbone network  $\emptyset$ , e.g. HRNet. We aim to locate each person and generate corresponding features  $\{\mathcal{F}_t^i\}_{i=1}^m$  which contain all the necessary cues required for single-person keypoint regression. Inspired by CID [41], our method differentiates each instance from global features  $\mathcal{F}_t$  on spatial and channel aspects, which isolates distractions from the background and allows better context cues exploration

Specifically, we first utilize the Instance Abstraction (IA) module to extract the root location and feature information for each individual. The input to the IA is the extracted global feature map  $\mathcal{F}_t$  and the output is an n-channel heatmap  $\mathcal{C}_t$  for all possible root keypoints. To associate the same instance through time, IA also predicts a 2D displacement map  $D_t \in \mathbb{R}^{2 \times H \times W}$ , representing instance 2D tracking offsets. The displacement can capture the difference in location of the object in the current frame and the previous frame. To generate an instance-aware feature map, we select the root coordinates of  $m$  persons based on the heatmap confidence scores. Features at the root joints on the feature map  $\mathcal{F}_t$  are regarded as representative features for those persons, which is used to identify and decouple each person from the background. To boost the discriminative power of person features, IA is trained with a contrastive loss.

We use the Global Feature Decoupling (GFD) [41] to decouple person cues from  $\mathcal{F}_t$  based on instance features and corresponding root locations. It jointly considers spatial-wise and channel-wise decoupling as following:

$$\mathcal{F}_s^{(i)} = \mathcal{M}_t^{(i)} \cdot \mathcal{F}_t \quad (1)$$

where  $\mathcal{F}_s^{(i)}$  denotes the spatial-recalibrated feature map for the  $i$ -th person at each timestamp, and  $\mathcal{M}(i)_t$  represents the foreground for each person.

$$\mathcal{F}_c^{(i)} = \mathcal{F}_t \otimes f_t^i \quad (2)$$

where  $f_t^i$  is instance feature at the corresponding root location and  $\otimes$  denotes the element-wise manipulation.  $f_t^i$  is

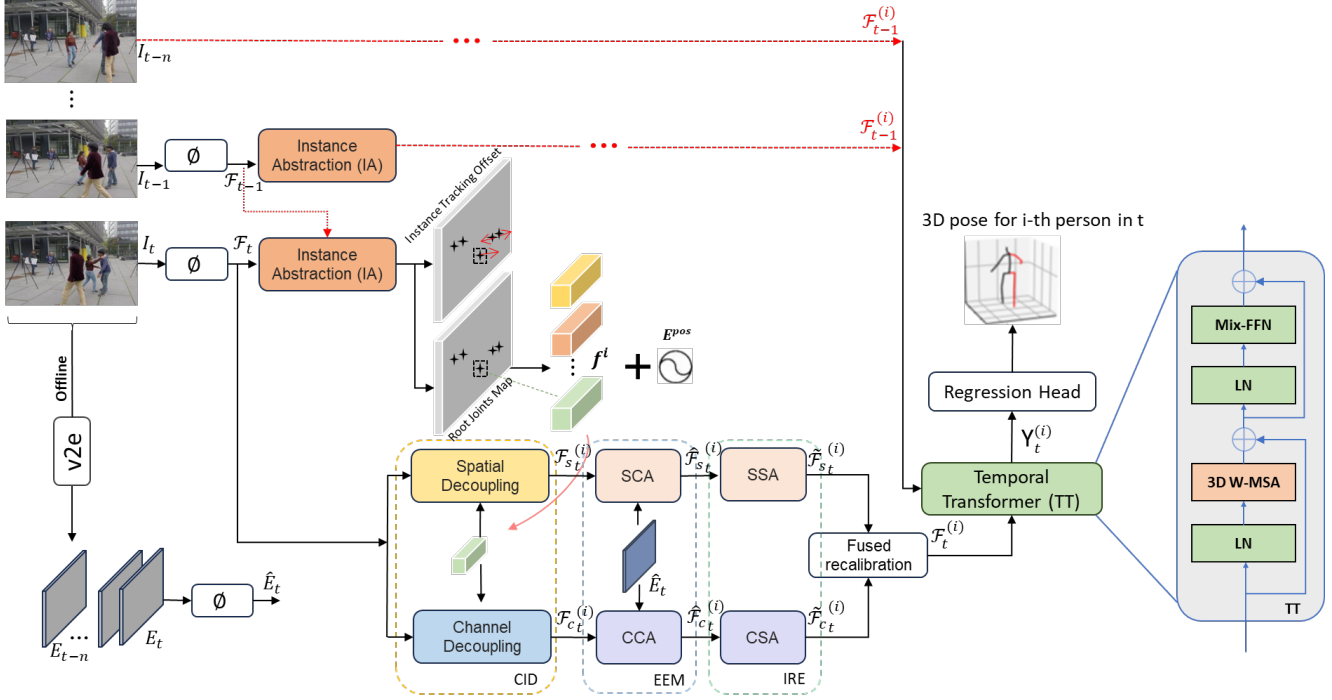


Figure 1. The overview of Event-guided Video Transformer (EVT), which includes Contextual Instance Decoupling (CID), Event-Aware Embedding Module (EEM), Intra-human Relation Extraction (IRE) and Temporal Modeling. Given video frames of height  $H$  and width  $W$ , deep features  $\mathcal{F}_t$  are extracted by a backbone network  $\emptyset$ , and further fed into Instance Abstraction (IA) module to extract individual features for each person. Based on individual features, instance-aware feature maps containing more important intra-human cues are decoupled in CID at both spatial and channel levels. The decoupled instance features are efficiently fused with the event features  $\hat{E}_t$  in EEM via two types of attention blocks (SCA, CCA Sec. 3.3). The updated instance-aware features are fed into SSA, CSA (Sec. 3.4) within IRE to further refine the features using self-attention. The refined feature maps from both branches are fused via Fused recalibration to produce the final instance-aware feature maps. Then, the instance-aware feature maps are fed into a Temporal Transformer (TT) to model inter-frame correlations. The output visual feature in the final layer of EVT outputs the 3D coordinates of a person.  $\oplus$  means element-wise addition. v2e [13] toolbox offline generate realistic synthetic event stream from video frames.

used to weight different channels, hence producing different channel-recalibrated feature maps for different instances.

The decoupled instance features  $\mathcal{F}_t^i = \{\mathcal{F}_s^{(i)}, \mathcal{F}_c^{(i)}\}_t$  with both spatial and contextual cues are regarded as intermediate features to be further fed into Event-Aware Embedding Module.

### 3.3. Event-Aware Embedding Module (EEM)

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 helps 3D HPE in video. We claim that human motion can be better estimated by jointly considering events and frames. Event stream capturing temporal coherence which can be used as guidance to intergrete multi-modality correlations. To this end, our approach is the first work to incorporate the event information into video 3D HPE. Since there is no existing 3D HPE dataset that contains hybrid inputs of events and RGB videos, we simply use the v2e toolbox [13] to gener-

ate realistic synthetic event stream from video frames. 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 [19] during data preparation.

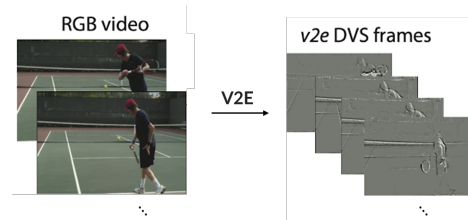


Figure 2. Event stream.

Given a sequence of video frames  $I$ , we offline generate a set of corresponding event stream  $E = \{E_t \mid t \in [1, T]\}$ .

$$E_t = v2e(I_t) \quad (3)$$

Then, we feed the  $E_t$  into an image backbone and extract the deep event features  $\hat{E}_t$  that contain meaningful tempo-

ral information. We aim to effectively incorporate the event features  $\hat{E}_t$  to the instance features at each timestamp. As we discussed in Sec. 3.2, we obtain the decoupled instance features  $\mathcal{F}_t^i = \left\{ \mathcal{F}_s^{(i)}, \mathcal{F}_c^{(i)} \right\}_t$  in both spatial and channel level respectively. To this end, we propose an Event-Aware Embedding Module (EEM) to take full advantage of event features from both spatial and channel branches. EEM concentrates on combining event feature representation with instance features through spatial and channel attention blocks. In spatial branch, we apply Spatial Cross Attention (SCA) to both visual features and event features so that we can identify important spatial-wise interaction between the feature maps. In channel branch, we use Channel Cross Attention (CCA) to identify important channel-wise interaction between the two types of feature maps.

### 3.3.1 Spatial Cross Attention (SCA)

The SCA module performs a spatial-wise interaction between  $\hat{E}_t$  and  $\mathcal{F}_s^{(i)}$ , then gives the final refined feature map  $\hat{\mathcal{F}}_s^{(i)}$ . The SCA module computes the event awareness of instance features through cross-modal similarity and produces an event-aware map. Inspired by LLFormer [44], we build our SCA blocks by integrating an Axis-based Attention and Feed-Forward Network (FN) with the plain transformer units. Such Axis-based Attention can compute the attention map on both height and width axis with a low computational cost. Since the mechanisms of height and width axis attention are similar, we thus only introduce the details of height axis attention for ease of illustration. Details of the learning process are provided below.

We first generate query ( $\mathbf{Q}$ ) by applying  $1 \times 1$  convolutions followed by  $3 \times 3$  depth-wise convolutions to encode spatial context as  $\mathbf{Q} = W_d^Q W_p^Q \hat{E}_t$ . Similarly, key ( $\mathbf{K}$ ) and value ( $\mathbf{V}$ ) are generated from the  $\mathcal{F}_s^{(i)}$ ,  $\mathbf{K} = W_d^K W_p^K \mathcal{F}_s^{(i)}$ ,  $\mathbf{V} = W_d^V W_p^V \mathcal{F}_s^{(i)}$ , where  $W_p^{(\cdot)}$  is the  $1 \times 1$  point-wise convolution and  $W_d^{(\cdot)}$  is the  $3 \times 3$  depth-wise convolution. After that, the query and key are reshaped for conducting dot-product to generate height axis attention map  $\mathbf{A}_S \in \mathbb{R}^{H \times H \times W}$ . The height axis attention (HA) can be formulated as

$$\begin{aligned} \mathbf{Q}, \mathbf{K}, \mathbf{V} &= W_d^Q W_p^Q \hat{E}_t, W_d^K W_p^K \mathcal{F}_s^{(i)}, W_d^V W_p^V \mathcal{F}_s^{(i)} \\ \mathbf{HA}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) &= \hat{\mathbf{V}} \text{Softmax}(\hat{\mathbf{Q}} \cdot \hat{\mathbf{K}} / \alpha) \end{aligned} \quad (4)$$

where  $\hat{\mathbf{Q}} \in \mathbb{R}^{H \times C \times W}$ ,  $\hat{\mathbf{K}} \in \mathbb{R}^{C \times H \times W}$  and  $\hat{\mathbf{V}} \in \mathbb{R}^{C \times H \times W}$  are obtained after reshaping tensors from the original size  $\mathbb{R}^{H \times W \times C}$ ,  $\alpha$  is a scale factor. The output feature  $\mathcal{F}'_s^{(i)}$  can be obtained by

$$\mathcal{F}'_s^{(i)} = W_{1 \times 1} \left( \mathbf{HA}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) \right) \quad (5)$$

Then, we reshape  $\mathcal{F}'_s^{(i)}$  to obtain the output instance feature  $\mathcal{F}_{out_s}^{(i)} \in \mathbb{R}^{H \times W \times C}$  of height axis attention.  $\mathcal{F}_{out_s}^{(i)}$  and  $E_t$  are forwarded to a similar width axis attention which computes cross-attention along the width axis. After doing both axis attention, we feed the update features into a normalization layer (LN) and Feed-Forward Network (FN) to generate the final refined feature map  $\hat{\mathcal{F}}_s^{(i)}$ .

$$\hat{\mathcal{F}}_s^{(i)} = \text{FN}(\text{LN}(\mathcal{F}_{out_s}^{(i)}) + \mathcal{F}_s^{(i)}) \quad (6)$$

### 3.3.2 Channel Cross Attention (CCA)

The CCA module performs a channel-wise interaction between  $\hat{E}_t$  and  $\mathcal{F}_c^{(i)}$ , then gives the final refined feature map  $\hat{\mathcal{F}}_c^{(i)}$ . We build our CCA blocks by integrating a transposed-attention mechanism and Feed-Forward Network (FN). Such transposed-attention can focus the event awareness of the instance features from channel level and still keep a low computational cost.

Similar to the SCA, we project  $\hat{E}_t$  and  $\mathcal{F}_c^{(i)}$  to query ( $\mathbf{Q}$ ), key ( $\mathbf{K}$ ) and value ( $\mathbf{V}$ ) respectively.

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = W_d^Q W_p^Q \hat{E}_t, W_d^K W_p^K \mathcal{F}_c^{(i)}, W_d^V W_p^V \mathcal{F}_c^{(i)} \quad (7)$$

Next, we reshape query and key so that their dot-product generates a transposed-attention map  $\mathbf{A}_C \in \mathbb{R}^{C \times C}$ . Then, we calculate channel cross attention (CA) as follows:

$$\mathbf{CA}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) = \hat{\mathbf{V}} \text{Softmax}(\hat{\mathbf{Q}} \cdot \hat{\mathbf{K}} / \beta) \quad (8)$$

where  $\hat{\mathbf{Q}} \in \mathbb{R}^{HW \times C}$ ,  $\hat{\mathbf{K}} \in \mathbb{R}^{C \times HW}$  and  $\hat{\mathbf{V}} \in \mathbb{R}^{HW \times C}$  are obtained after reshaping tensors from the original size  $\mathbb{R}^{H \times W \times C}$ ,  $\beta$  is a scale factor. The output feature  $\mathcal{F}'_c^{(i)}$  can be obtained by

$$\mathcal{F}'_c^{(i)} = W_{1 \times 1} \left( \mathbf{CA}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) \right) \quad (9)$$

Finally, we feed the update features into a normalization layer (LN) and Feed-Forward Network (FN) to generate the final refined feature map  $\hat{\mathcal{F}}_c^{(i)}$ .

$$\hat{\mathcal{F}}_c^{(i)} = \text{FN}(\text{LN}(\mathcal{F}'_c^{(i)}) + \mathcal{F}_c^{(i)}) \quad (10)$$

Compared with the standard attention which has quadratic complexity, both SCA and CCA result in linear complexity. The updated instance features  $\hat{\mathcal{F}}_t^i = \left\{ \hat{\mathcal{F}}_s^{(i)}, \hat{\mathcal{F}}_c^{(i)} \right\}_t$  with both spatial and contextual guidance from event features are further use to extract spatial-temporal relation via an Intra-human Relation Extraction (IRE) Module.

## 3.4. Intra-human Relation Extraction (IRE) Module

IRE is designed to implicitly model a specific relationship within a decoupled instance feature map. Given the

output event-aware instance features  $\hat{\mathcal{F}}_t^i = \left\{ \hat{\mathcal{F}}_s^{(i)}, \hat{\mathcal{F}}_c^{(i)} \right\}_t$ , IRE also learns intra-correlation of each individual in each frame on both spatial and channel branch. We achieve this by using the same attention mechanism discussed in Sec 3.3. Differently, we change the spatial cross attention and channel cross attention in SCA and CCA to self-attention which are SSA and CSA. In other words, the query (**Q**), key (**K**) and value (**V**) projections are generated from the same input, either  $\hat{\mathcal{F}}_s^{(i)}$  or  $\hat{\mathcal{F}}_c^{(i)}$ . The output  $\tilde{\mathcal{F}}_t^i = \left\{ \tilde{\mathcal{F}}_s^{(i)}, \tilde{\mathcal{F}}_c^{(i)} \right\}_t$  has the same shape, where pixel-wise interaction has been explored inside the attention operation. After obtaining the  $\tilde{\mathcal{F}}_t^i$ , we apply a Fused recalibration on these two elements to produce the final instance-aware feature map at time  $t$ . The instance-aware feature map  $\mathcal{F}_t^{(i)}$  for the  $i$ -th person can be computed as,

$$\mathcal{F}_t^{(i)} = \text{ReLU}(\text{Conv}(\left[ \tilde{\mathcal{F}}_s^{(i)}; \tilde{\mathcal{F}}_c^{(i)} \right])) \quad (11)$$

where  $\tilde{\mathcal{F}}_s^{(i)}$  and  $\tilde{\mathcal{F}}_c^{(i)}$  are fused to seek better discriminative power in decoupling persons. Across all the frames in the video sequence, we can extract all the instance-aware feature maps as  $\mathcal{F}_T^{(I)} = \left\{ \mathcal{F}_t^{(i)} \mid t \in [1, T], i \in [1, m] \right\}$ , where  $T$  is number of frames and  $m$  is the number of instances.

### 3.5. Temporal Modeling

Compared with 3D HPE on image, the key of video 3D HPE is to leverage the inter-frame temporal correlations. Here, we follow the design paradigm of video Swin-Transformer [25] and propose Temporal Transformer to facilitate temporal modeling. Specifically, for each query frame and its previous frames, we feed  $\left( \mathcal{F}_t^{(i)}; \mathcal{F}_1^{(i)}, \sim, \mathcal{F}_{t-1}^{(i)} \right)$  to the Temporal Transformer block. As Equation 12, we call the Temporal Transformer as  $f_{TT}$  and symbol  $\circ$  is composition operator.

$$\mathcal{Y}_t^{(i)} = f_{TT} \circ \left( \mathcal{F}_t^{(i)}; \mathcal{F}_1^{(i)}, \sim, \mathcal{F}_{t-1}^{(i)} \right) \quad (12)$$

where  $\mathcal{Y}_t^{(i)}$  is the final output instance-aware feature map encoding the meaningful temporal information from adjacent frames and is used to infer 3D pose. The architecture of our Temporal Transformer base unit is illustrated at the right of Fig. 1. Each unit consists of a 3D Windows Multi-head Self Attention (3D W-MSA) [25], a Mix-FeedForward Network (Mix-FFN) [46] and two Layer Norm (LN) layers. 3D W-MSA evenly partitions the 3D input feature map into a set of non-overlapping cubes and applies MSA on them. Mix-FFN introduces a depth-wise 3x3 convolution between the two MLPs to connect non-overlapping cubes.

### 3.6. Loss Function

For  $i$ -th person in each video frame  $t$ , its feature map  $\mathcal{Y}_t^{(i)}$  is used to learn 3D offset  $M_o$  of size  $J \times H \times W$

using a convolutional layer. Since we already know the root keypoint for this instance from the heatmap  $C_t$  predicted in the Instance Abstraction (IA) module (Sec. 3.2). The corresponding 3D offsets at the root coordinates in  $M_o$  are extracted to decode a whole 3D pose of size  $J \times 3$  for a person,  $J$  represents number of 3D joints. The decoding process is same as previous works [3, 30].

During training, we use L1 loss for 3D offsets regression and instance 2D tracking offsets regression.

$$\mathcal{L}_{3D} = \sum_{t=1}^T \left( M_o, \hat{M}_o \right)_t, \mathcal{L}_{tracking} = \sum_{t=1}^T \left( D, \hat{D} \right)_t \quad (13)$$

where  $\hat{M}_o$  and  $\hat{D}$  mean the ground-truth of 3D offset map and instance 2D tracking offsets map.

$\mathcal{L}_{IA}$  is computed with ground-truth heatmap  $\mathcal{H}_t^*$  of root joints and predicted heatmap  $C_t$  at each timestamp. It also combines the contrastive loss  $l(f_t^i)$  [41] to ensure the discriminative power of each decoupled instance feature.

$$\mathcal{L}_{IA} = \sum_{t=1}^T \left( \text{FL}(|\mathcal{H}_t^*; C_t|) + \frac{1}{m} \sum_{i=1}^m l(f_t^i) \right) \quad (14)$$

where  $\text{FL}(\cdot)$  computes the Focal Loss [3, 18].

The total loss function  $\mathcal{L}$  is the weighted sum of all these loss components,

$$\mathcal{L} = \mathcal{L}_{3D} + \mathcal{L}_{tracking} + \alpha \mathcal{L}_{IA} \quad (15)$$

where  $\alpha$  represents a loss weight.

## 4. Experiments

In this section, we elaborate the experiment results of EVT. We first introduce the implementation details of EVT, and then report results and compare with SOTA methods using two widely-used single-person datasets: Human3.6M [14], MPI-INF-3DHP [28] and one multi-person dataset: CMU Panoptic [15]. Next, we conduct ablation studies for EVT. 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  $\emptyset$  of EVT for all experiments and follow the most configuration of [41]. In our experiments, EVT is trained on 4 A100 GPUs with a batch size of 4 sequences/GPU, while the sequence length is 7 frames and the input size is 512x512. The total training epochs is 60. Adam optimizer is adopted and the initial learning rate is 5e-4, which decreases 10x at 40 and 50 epochs. The loss weight  $\alpha$  equals 10 during training.

Protocol 1		T	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg. ↓
Zheng et al. [54]	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. [21]	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. [39]	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
MehrabanI 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	<u>38.4</u>
Qiu et al. [34] *	ACMMM'22	5	36.5	40.1	38.4	40.7	42.6	42.8	30.1	43.4	46.1	58.0	40.2	37.1	40.8	32.1	33.5	40.2
Ours(EVT) w/o Event *		7	37.3	39.1	38.6	35.5	42.7	45.9	31.6	39.4	47.5	56.3	40.4	36.9	37.5	28.3	27.6	39.0
Ours(EVT) w Event *		7	36.1	38.2	37.9	32.1	41.6	45.1	30.5	38.2	47.2	54.7	40.1	36.3	35.6	26.7	27.2	<b>37.8</b>
Protocol 2		T	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg. ↓
Pavilo et al. [32]	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. [54]	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. [39]	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	32.0
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
MehrabanI 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
Qiu et al. [34] *	ACMMM'22	5	27.0	24.8	32.2	30.1	27.8	32.1	22.3	28.7	30.7	24.4	32.7	37.8	21.9	31.1	24.7	<u>28.5</u>
Ours(EVT) w/o Event *		7	27.9	25.6	31.8	32.4	28.3	33.0	23.5	29.1	32.4	25.8	32.2	35.3	23.7	30.4	23.6	29.0
Ours(EVT) w Event *		7	27.5	24.9	32.0	31.6	27.7	32.5	23.3	28.2	32.1	24.9	31.8	34.6	23.4	28.2	22.7	<b>28.3</b>

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, and \* represents an end-to-end model. Bold indicates the best and underline indicates the second best.

## 4.2. Datasets and Evaluation Metric

**Human3.6M dataset.** Human3.6 [14] is the largest indoor benchmark for single-person 3D pose estimation, which includes 7 subjects that 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, 39], MPJPE, Percentage of Correct Keypoint (PC) within 150 mm range, and Area Under the Curve (AUC) are reported as evaluation metrics.

**CMU Panoptic dataset.** CMU Panoptic [15] is a larger-scale multi-person dataset, captured by multiple cameras. Following the settings of previous works, we use 160K images from different videos as the training set and the videos from two cameras (16, 30) as the testing set. For comparison, MPJPE is used for evaluation.

## 4.3. Compare 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 Table.1. Our EVT with  $T = 7$  achieves new state-of-the-art results with an MPJPE of 37.8mm and a PA-MPJPE of 28.3mm in Protocol 1 and Protocol 2, respectively. The results demonstrate the effectiveness of the proposed EVT. Compared with other transformer-based methods [27, 55], EVT outperforms them. Even these models are based on a larger frame number above 81, EVT with only  $T = 7$  (window size) obtains better results since most of the approaches in Table.1 follow 2D-to-3D lifting paradigm which loses the visual depth feature in the process of temporal modeling.

Compared with EVT w/o Event guidance, the complete one gains 3% improvement in both Protocol 1 and Protocol 2, which prove that event stream will provides implicit structure and temporal information to help infer more accurate 3D pose.

Methods	T	PCK ↑	AUC ↑	MPJPE ↓
Einfalt et al. [10]	81	95.4	67.6	46.9
Zhao et al. [52]	81	97.9	78.8	27.8
Tang et al. [39]	81	<b>98.7</b>	83.9	23.1
Chen et al. [4]	96	<b>98.7</b>	72.9	37.2
MehrabanI et al. [27]	81	98.3	<u>84.2</u>	<u>18.2</u>
Ours(EVT) w/o Event *	9	98.2	84.1	18.3
Ours(EVT) w Event *	9	<u>98.5</u>	<b>84.7</b>	<b>17.9</b>

Table 2. Quantitative comparison with state-of-the-art methods on MPI-INF-3DHP. T: Number of input frames. \* represents an end-to-end model. 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 modified EVT to use  $T = 9$  frames. As shown in Table.2, our method consistently outperforms others in terms of MPJPE. Notably, our EVT w Event achieves remarkable results with an 84.7% AUC and a 17.9 mm P1 error. This outperforms the previous 2nd-best STCFormer [39] by a significant margin of 1% in AUC and 5.1 mm in P1 error. Besides, it achieves 98.5% PCK, which is 0.2% lower than the PCK performance of the compared models.

Methods		MPJPE(mm)↓
DAS [45]	CVPR 22	53.8
VirtualPose [35]	ECCV 22	58.9
IVT * [34]	ACMMM 22	48.4
POTR-3D [31]	ICCV 23	57.8
<b>Ours(EVT) w/o Event *</b>		<b>47.6</b>
<b>Ours(EVT) w Event *</b>		<b>45.7</b>

Table 3. Comparison with SOTA methods on multi-person 3D human pose estimation dataset (CMU Panoptic) in MPJPE. \* represents an end-to-end model.

### 4.3.3 Results on CMU Panoptic dataset

As seen in Table.3, EVT also achieves the state-of-the-art performance on CMU-Panoptic, 2.7mm or 5.6% leading the other E2E model IVT [34]. EVT achieves a relative gain of 21% compared with the POTR-3D [31]. CMU-Panoptic contains videos with a denser crowd of 3–8 people, making the estimation more challenging. The result indicates that EVT operates well even in this challenging situation. It demonstrates that having the discriminative power in decoupling persons allows EVT to handle both single-person and multi-person 3D HPE tasks.

## 4.4. Ablation Studies

In this section, we verify the effectiveness of the proposed EEM, IRE, in event-guided video transformer (EVT). Next, we compare the parameters and computational costs of variants in EVT design.

Methods	Submodule	Param (M)	Flops (T)	MPJPE (mm) ↓
MHFormer [21]		30.92	-	43.1
IVT [34]		40.85	-	40.2
MotionBERT [55]		42.50	-	39.2
EVT-base (w/o Event)	CID + TT	36.54	0.127	41.8
EVT (w/o Event)	CID + TT + IRE	39.27	0.135	39.0
EVT (w Event)	CID + TT + IRE + EEM	44.71	0.146	37.8

Table 4. Ablation study of EVT on Human3.6. CID + TT means using Contextual Instance Decoupling and Temporal Transformer for baseline setting. IRE means using Intra-human Relation Extraction Module in EVT. EEM means using Event-Aware Embedding Module in EVT.

#### 4.4.1 Effectiveness of proposed sub-modules

We conduct the ablation study on Human3.6m dataset to verify the effectiveness of each proposed submodule in the

EVT. First of all, we build an EVT-base for 3D human pose estimation baseline where we do not input event stream. We only keep the CID to generate instance-aware feature maps and Temporal Transformer to capture the essential temporal dependency in video sequence. As shown in Table.4, EVT-base achieves 41.8mm in MPJPE. Combined with the IRE, EVT with both SSA and CSA obtains 39.0mm in MPJPE and achieves a relative gain of 6.7%. This indicates that extracting the intra-human relation within each instance provide good clues to infer 3D pose. Compared with the first two rows without event guidance, the complete EVT adding EEM further improves MPJPE by 1.2mm (3%). EEM can effectively fuse visual features and event features and enhance each instance-aware feature maps for better 3D pose estimation. These results show that the proposed IRE, EEM significantly improve the performance of video 3D HPE.

#### 4.4.2 Parameters and computational costs

The comparisons of including each submodule on parameters and computational costs are shown in Table. 4. Compared with the EVT-base, IRE obtains a relative gain of 6.7% but adds 2.73MB parameters, while the flops increase slightly. Even adding EEM to incorporate event features, the increasing parameters and the computational costs are acceptable while improving MPJPE to 37.8mm. Compared with lifting methods which take extremely long 2D pose sequences as input, EVT still achieves better accuracy despite using shorter lengths of both modalities. By incorporating an event stream, our model can capture temporal dependencies without using longer input frames. The model parameters do not increase much since we have solved the quadratic time and space complexity problem of vanilla transformer in our design (refer to Sec.3.3 for details).

## 5. Conclusion

In this paper, we propose a novel end-to-end event-guided video transformer (EVT) for video 3D human pose estimation. To capture spatial-temporal dependency for multi-person in video with the guidance from event stream, we propose an Event-Aware Embedding Module (EEM) to enable effective fusion of both visual features and event using a dual-branch attention mechanism. It learns helpful temporal and structure information in both spatial and channel branch simultaneously. To further model the single-person and inter-frame relationships human, we propose Intra-human Relation Extraction (IRE) Module and Temporal Transformer (TT). Combined with all these components, EVT outperforms the state-of-the-art methods on both single person and multi-person 3D human pose estimation benchmarks.

## References

- [1] 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. **1, 2, 3**
- [2] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2272–2281, 2019. **1**
- [3] 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. **6**
- [4] 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**
- [5] 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**
- [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] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1475–1484, 2022. **1**
- [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] 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**
- [13] 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**
- [14] 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, 7**
- [15] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. **6, 7**
- [16] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. **1**
- [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] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. **6**
- [19] 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**
- [20] 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**
- [21] 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. **7, 8**
- [22] 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**
- [23] 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**
- [24] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheng, 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**

- [25] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 6
- [26] 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
- [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, 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, 7
- [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. 1, 2
- [30] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6951–6960, 2019. 6
- [31] 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, 8
- [32] 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
- [33] Zhongwei Qiu, Kai Qiu, Jianlong Fu, and Dongmei Fu. Dgcnet: 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. 2, 3, 7, 8
- [35] Jiajun Su, Chunyu Wang, Xiaoxuan Ma, Wenjun Zeng, and Yizhou Wang. Virtualpose: Learning generalizable 3d human pose models from virtual data. In *European Conference on Computer Vision*, pages 55–71. Springer, 2022. 8
- [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. 1, 2
- [37] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 1
- [38] 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
- [39] 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. 7, 8
- [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. 1, 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. 2, 3, 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] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: a benchmark and transformer-based method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2654–2662, 2023. 5
- [45] Zitian Wang, Xuecheng Nie, Xiaochao Qu, Yunpeng Chen, and Si Liu. Distribution-aware single-stage models for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13096–13105, 2022. 8
- [46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 6
- [47] 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
- [48] 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

- [49] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5255–5264, 2018. [1](#)
- [50] 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](#)
- [51] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Sernet: 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. [1](#), [2](#)
- [52] 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](#)
- [53] 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, August 23–28, 2020, Proceedings, Part XV 16*, pages 550–566. Springer, 2020. [2](#)
- [54] 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](#), [7](#)
- [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](#), [7](#), [8](#)