Optimizing Human Pose Estimation Through Focused Human and Joint Regions

Yingying Jiao[1,2], Zhigang Wang[3*], Zhenguang Liu[4,5*], Shaojing Fan[6], Sifan Wu[1,2*], Zheqi Wu[3], Zhuoyue Xu[3],

[1]College of Computer Science and Technology, Jilin University
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University
[3]College of Computer Science and Technology, Zhejiang Gongshang University
[4]The State Key Laboratory of Blockchain and Data Security, Zhejiang University
[5]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[6]School of Computing, National University of Singapore
jiaoyy21@mails.jlu.edu.cn, wangzhigang2024@gmail.com, liuzhenguang2008@gmail.com, fanshaojing@gmail.com, wusifan2021@gmail.com, chasewoo17@gmail.com, 1069516849xzyy@gmail.com

Abstract

Human pose estimation has given rise to a broad spectrum of novel and compelling applications, including *action recognition*, *sports analysis*, as well as *surveillance*. However, accurate video pose estimation remains an open challenge. One aspect that has been overlooked so far is that existing methods learn motion clues from all pixels rather than focusing on the target human body, making them easily misled and disrupted by unimportant information such as *background changes* or *movements of other people*. Additionally, while the current Transformer-based pose estimation methods has demonstrated impressive performance with global modeling, they struggle with local context perception and precise positional identification.

In this paper, we try to tackle these challenges from three aspects: (1) We propose a bilayer Human-Keypoint Mask module that performs coarse-to-fine visual token refinement, which gradually zooms in on the target human body and keypoints while masking out unimportant figure regions. (2) We further introduce a novel deformable cross attention mechanism and a bidirectional separation strategy to adaptively aggregate spatial and temporal motion clues from constrained surrounding contexts. (3) We mathematically formulate the deformable cross attention, constraining that the model focuses solely on the regions centered at the target person body. Empirically, our method achieves state-of-the-art performance on three large-scale benchmark datasets. A remarkable highlight is that our method achieves an 84.8 mean Average Precision (mAP) on the challenging *wrist* joint, which significantly outperforms the 81.5 mAP achieved by the current state-of-the-art method on the PoseTrack2017 dataset.

Introduction

Human pose estimation, as a fundamental problem in the realm of computer vision and artificial intelligence (Wang and Zhang 2022; Geng et al. 2023), involves accurately identifying the anatomical keypoints of human bodies. Precise pose estimation is the key for the success of a machine as it paves the way for machines to accurately interpret human movements and behaviors. Accordingly, human pose estimation spans a wide range of applications from *action recognization*, *movement tracking*, to *augmented reality* (Yang et al. 2023; Tse, De Martini, and Marchegiani 2019; Su et al. 2021; Wu et al. 2024b; Liu et al. 2022b).

A plethora of research has been dedicated to the field of pose estimation on still images, evolving from early methods employing tree-based and random forest models (Wang and Mori 2008; Sapp, Toshev, and Taskar 2010) to current methodologies utilizing convolutional neural networks (Sun et al. 2019) and Transformers (Li et al. 2021). Despite their excellent performance on still images, applying these methods directly to video pose estimation leads to significant performance degradation due to the exclusive characteristics in videos, such as *rapid movement* and *video defocus*, which are frequently encountered in videos but absent in static images (Zhao, Xiong, and Lin 2018).

To address this issue, substantial studies have emerged that leverage temporal continuity to extract rich semantic visual contexts for human pose estimation in videos. Current methods can be roughly categorized into two main branches. One line of research (Bertasius et al. 2019; Liu et al. 2021) aggregates temporal information from neighboring frames for video pose estimation, employing CNN-based architectures and pose calibration. Fueled by the development of Transformers (Dosovitskiy et al. 2020; Vaswani et al. 2017), another line of studies (Jin, Lee, and Lee 2022; He and Yang 2024) strive to integrate attention mechanisms into model construction, yielding impressive results and showcasing their immense potential. However, a limitation inherent in existing Transformer-based methods (Jin, Lee, and Lee 2022) lies in their inability to effectively manage local dependencies. This limitation poses a notable challenge for visual perception tasks such as pose estimation, which require precise local positioning.

Following thorough experimentation and empirical investigation, we uncover two insights: (1) Existing methods (Liu et al. 2022a; Feng et al. 2023; Wu et al. 2024a) struggle to handle subtle pose changes, particularly in challenging scenarios with occlusions or motion blur. This may stem from the fact that current methods tend to capture temporal dynamics pixel-by-pixel rather than focusing solely on target

*Corresponding authors.

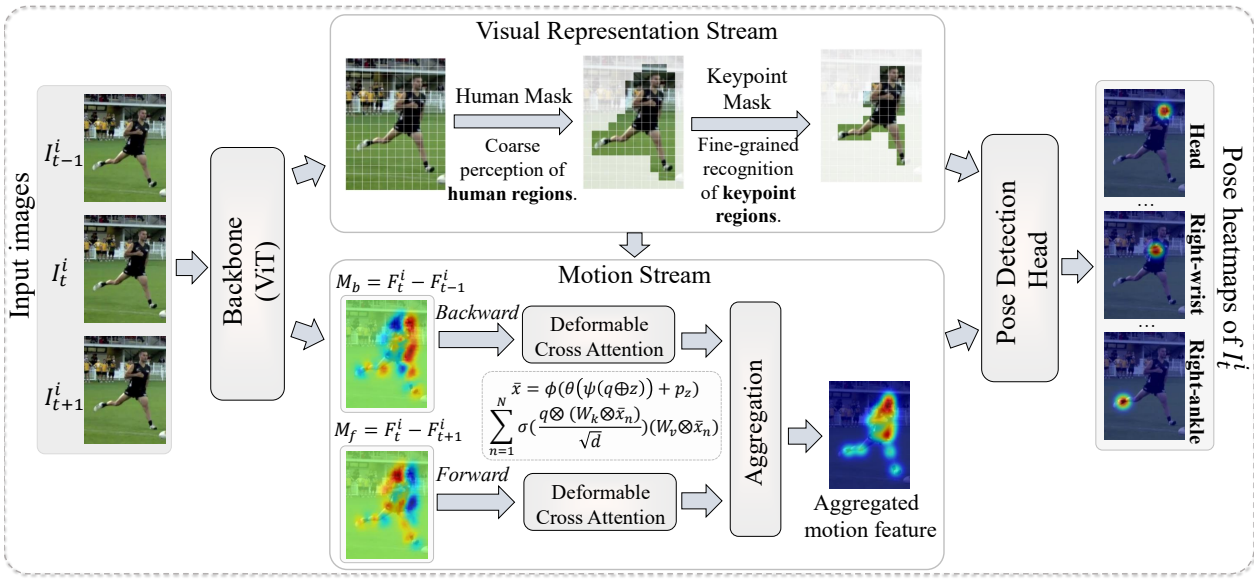

Figure 1: A high-level overview of our proposed VREMD, which utilizes a dual-stream architecture to collaboratively process and integrate complementary visual and motion features. The visual representation stream executes progressive enhancement of human keypoint-related features to achieve precise location recognition. The motion stream performs adaptive pose-related motion disentanglement through the novel deformable cross attention. $\{\boldsymbol{F}_{t-1}^i, \boldsymbol{F}_t^i, \boldsymbol{F}_{t+1}^i\}$ denote the visual features of three input frames $\{I_{t-1}^i, I_t^i, I_{t+1}^i\}$ output by backbone network.

human regions, leading to them being distracted by unuseful cues such as background changes or pixels far from the target person. (2) Additionally, previous studies (Liu et al. 2021, 2022a) adopting multiple sets of fixed deformable convolutions with varying dilation rates, which neglect the importance of adaptive scale selection.

Inspired by these, we propose a dual-stream framework, which executes **V**isual **R**epresentation **E**nhancement and **M**otion **D**isentanglement (VREMD) for human pose estimation in videos. Technically, we embrace three novel designs to tackle the challenge. (1) We propose a two-step human-keypoint mask module for coarse-to-fine visual enhancement, which progressively refines extracted representations from the human body and keypoints perspectives. (2) We further introduce a bidirectional decoupled module tailored for adaptively disentangling motion cues of the target person from unnecessary visual elements. (3) Furthermore, we mathematically formulate a deformable cross attention mechanism that constrains the model to focus exclusively on regions circumscribing the target human body.

Our framework exemplifies the collaborative advantage between local spatial focus and adaptive temporal clues extraction, opening up possibilities for rethinking the pose estimation task from emphasizing on the target human body and masking out the irrelevant spatio-temporal contexts. To evaluate the efficacy of our method, we conduct extensive experiments on three public benchmarks, achieving state-of-the-art performance. The key contributions of our method are summarized as follows:

- We present a dual-stream framework that integrates visual enhancement and motion disentanglement to highlight target human areas and fliter other non-essential regions for human pose estimation.
- We creatively introduce a deformable cross attention to disentangle pose-related motion cues, harnessing bidirectional temporal dynamics and enabling the model to robustly handle complex pose variations of the target human.
- Empirically, our method achieves state-of-the-art performance on three large-scale benchmarks, and overall provides insights into integrating Transformer-based methods with region-specific enhancement strategies to boost their local localization capabilities.

Our Method

Preliminaries. Our method follows the top-down paradigm, which first extracts each individual person from an image and then estimates their poses. Specifically, we first utilize an object detector to extract the bounding box for person i in a video frame I_t that is to be detected. Subsequently, we expand the bounding box by 25% and crop the same person in the adjacent frames (*i.e.*, I_{t-1} and I_{t+1}). As a result, we obtain a sequence of consecutive frames for person i: $\boldsymbol{\mathcal{I}_t^i} = \{I_{t-1}^i, I_t^i, I_{t+1}^i\}$. Given a sequence of video frames $\boldsymbol{\mathcal{I}_t^i}$ that includes the key frame I_t^i and the auxiliary frames I_{t-1}^i and I_{t+1}^i, our target is to detect the human pose within I_t^i. We aim to strengthen the utilization of supplementary temporal information in auxiliary frames by employing incremental visual representation enhancement and adaptively disentangling useful motion information, thus tackling the

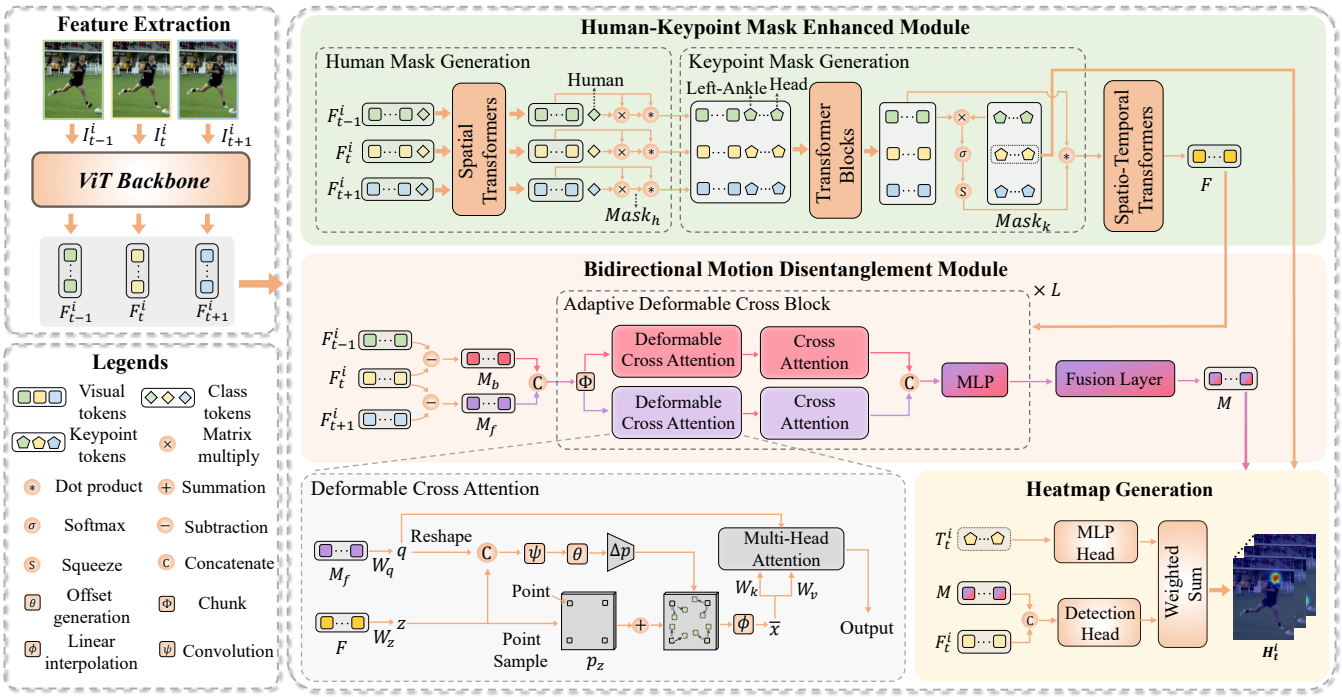

Figure 2: The overall pipeline of our VREMD framework. Given an input sequence $\{I^i_{t-1}, I^i_t, I^i_{t+1}\}$, our goal is to estimate the human pose of the key frame I^i_t. We initially extract the visual features via a ViT backbone, and then feed them into the Human-Keypoint Enhanced module and the Bidirectional Motion Disentanglement module to obtain $\boldsymbol{T}^i_t$ and $\boldsymbol{M}$. Finally, the outputs derived from different heads are combined through a weighted sum to arrive at the final predicted pose heatmap $\boldsymbol{H}^i_t$.

common issue of existing methods being interfered with by irrelevant information regarding the target human.

Method overview. The overview pipeline of our proposed VREMD is depicted in Figure 2. VREMD constructs a dual-stream architecture with inter-module communication that enhances both visual features and captures meaningful motion cues. Specifically, VREMD incorporates two distinct modules: a Human-Keypoint Mask Enhanced module (HKME) and a Bidirectional Motion Disentanglement module (BMD). First, we utilize a Vision Transformer backbone to extract visual features $\{\boldsymbol{F}^i_{t-1}, \boldsymbol{F}^i_t, \boldsymbol{F}^i_{t+1}\}$ from the input frame sequence $\boldsymbol{\mathcal{I}^i_t}$, which are then simultaneously fed into both the HKME and BMD modules. The HKME generates dual masks for a coarse-to-fine representation refinement, resulting in enhanced feature $\boldsymbol{F}$ and key frame keypoint tokens $\boldsymbol{T}^i_t$. The BMD computes the motion features and, utilizing $\boldsymbol{F}$ as a constraint, dynamically derives joint-related motion contexts to produce the filtered $\boldsymbol{M}$. Finally, the keypoint heatmaps $\boldsymbol{H}_k$ from key frame tokens $\boldsymbol{T}^i_t$ via an MLP and the heatmaps $\boldsymbol{H}_m$ decoded from $\boldsymbol{M}$ and key frame features $\boldsymbol{F}^i_t$ are weighted, summed, and combined to produce the final pose estimation $\boldsymbol{H}^i_t$. The following sections will elaborate on the two key components in detail.

Human-Keypoint Mask Enhanced Module

Despite the Transformers architecture achieving remarkable success in various fields (Dosovitskiy et al. 2020; Vaswani et al. 2017), its application in video pose estimation has been limited. Given the significant potential demonstrated by this architecture in other visual perception tasks (Zheng et al. 2021; Li et al. 2022), we seek to design a novel Transformer-based framework specially tailored for video pose detection. A naive approach to aggregate unique temporal cues from a video would be to concatenate features across multiple frames for full-token computation. Yet, such a straightforward treatment strategy faces two issues: excessive capture of redundant information between adjacent frames, and a lack of focus on task-relevant tokens.

Inspired by previous work (Strudel et al. 2021; Li et al. 2021), we propose a Human-Keypoint Mask Enhanced module with a progressive refinement architecture, addressing the aforementioned issues through three steps: (1) We generate a human mask to coarsely enhance the perception of the target human. (2) We produce a keypoint mask to achieve finer filtering of keypoint-related features. (3) We utilize spatio-temporal networks to aggregate the highlighted spatio-temporal cues of these visual features. This step-by-step optimization strategy can discern articular visual tokens, simulating the capability of localized identification, which promotes precise pose estimation.

Human mask. Given a visual feature sequence $\{\boldsymbol{F}^i_{t-1}, \boldsymbol{F}^i_t, \boldsymbol{F}^i_{t+1}\} \in \mathbb{R}^{3\times N\times D}$ output by the ViT backbone, we concatenate a learnable class token $\boldsymbol{T}_h \in \mathbb{R}^{3\times 1\times D}$ with a category of human to each feature. These features

then individually pass through cascaded Transformer blocks for intra-frame spatial similarity computation. We separate the result into human token $\boldsymbol{T}_h$ and visual features $\overline{\boldsymbol{F}} \in \mathbb{R}^{3\times N\times D}$. After transposing the human token, we perform matrix multiplication to obtain the human mask $\boldsymbol{Mask}_h \in \mathbb{R}^{3\times N\times 1}$. Finally, we secure a coarsely selected feature $\boldsymbol{F}_c \in \mathbb{R}^{3\times N\times D}$ by executing element-wise dot product between the human mask $\boldsymbol{Mask}_h$ and the visual feature $\overline{\boldsymbol{F}}$, utilizing broadcasting. The above operations can be formulated as:

$$\boldsymbol{F}_c = \bigoplus_{\delta=t-1}^{t+1} \overline{\boldsymbol{F}}_\delta \odot \underbrace{(\overline{\boldsymbol{F}}_\delta \otimes \boldsymbol{T}_{h\delta}^T)}_{\boldsymbol{Mask}_{h\delta}}, \tag{1}$$

where $\bigoplus$, δ, $\odot$, $\otimes$, and $\boldsymbol{T}^T$ denote concatenation, temporal index of frames, dot product, matrix multiplication, the transpose of $\boldsymbol{T}$, respectively.

Keypoint mask. In pursuit of more precise keypoint-related feature enhancement, we employ additional auxiliary tokens to accurately localize spatial positions by integrating multi-frame representations in the spatio-temporal domain. We concatenate the learnable keypoint tokens $\boldsymbol{T}_k \in \mathbb{R}^{3\times J\times D}$ (Note that J is the number of keypoints) to the coarsely selected feature $\boldsymbol{F}_c$ and separate the multi-frame features, which are then linked along the token dimension and fed into Transformer blocks for spatio-temporal learning. Subsequently, we split the visual features and keypoint tokens from the output and gather them over multiple frames, resulting in multi-frame features $\widehat{\boldsymbol{F}} \in \mathbb{R}^{(3\cdot N)\times D}$ and multi-frame keypoint tokens $\widehat{\boldsymbol{T}}_k \in \mathbb{R}^{(3\cdot J)\times D}$. After transposing the multi-frame features, we perform matrix multiplication with the multi-frame keypoint tokens to produce the keypoint confidence map $\boldsymbol{Map} \in \mathbb{R}^{(3\cdot J)\times(3\cdot N)}$. We apply the softmax function to compute element-wise weights for the map $\boldsymbol{Map}$, and summing along the second-to-last dimension followed by transposition yields the keypoint mask $\boldsymbol{Mask}_k \in \mathbb{R}^{(3\cdot N)\times 1}$:

$$\boldsymbol{Mask}_k = \sum_{j=1}^{J} S(\sigma(\widehat{\boldsymbol{T}}_{kj} \otimes \widehat{\boldsymbol{F}}^T)), \tag{2}$$

where $j \in \{1, \ldots, J\}$, $S(\cdot)$, $\sigma(\cdot)$, and $\otimes$ denote the keypoint index, squeeze operation, softmax function, and matrix multiplication, respectively. The keypoint mask is element-wise multiplied with the multi-frame features $\widehat{\boldsymbol{F}}$ to create the refined filtered features $\boldsymbol{F}_f \in \mathbb{R}^{(3\cdot N)\times D}$.

Spatio-temporal aggregation. To fully leverage the refined representation information, we perform decoupled spatio-temporal feature aggregation through the spatio-temporal Transformers. Specifically, we first separate the refined filtered features $\boldsymbol{F}_f$ and undertake frame-level spatial modulation. Then, each token is concatenated with its corresponding token in the temporal domain to undergo temporal modulation, resulting in $\overline{\boldsymbol{F}}_f \in \mathbb{R}^{(3\cdot N)\times D}$. Finally, we adopt an MLP to execute token dimensionality reduction on $\overline{\boldsymbol{F}}_f$ to attain spatio-temporal aggregation of multi-frame features, leading to the enhanced feature $\boldsymbol{F} \in \mathbb{R}^{N\times D}$.

Bidirectional Motion Disentanglement Module

To extract useful complementary information from auxiliary frames, prior methods (Liu et al. 2021; Feng et al. 2023) implicitly model feature residuals to capture motion evidence. The common practice among these paradigms is to directly concatenate the computed multiple motion features for convolution after their calculation, which considers temporal continuity but overlooks insights from the temporal direction. We observe that, from the perspective centered around the key frame, the essential temporal details that need to be focused on actually originate from two different directions, namely forward and backward. Considering this intrinsic factor, we design a bidirectional separation strategy to decouple the continuous motion into parallel forward and backward motion trajectories. Furthermore, existing methods do not differentiate motion clues in the spatial dimension, which can lead to learning pose-irrelevant information (*e.g.*, background, other people, etc.) that can disrupt detection. Moreover, existing methods heavily rely on deformable convolutions for local motion calibration, potentially leading to models that are overly tailored and limiting their compatibility with Transformer-based architectures. To tackle these challenges, we introduce deformable cross attention (DCA) for the first time and create the Adaptive Deformable Cross block by employing it, which adaptively captures pose-related motion dynamics.

Adaptive Deformable Cross block. Given the features $\{\boldsymbol{F}_{t-1}^i, \boldsymbol{F}_t^i, \boldsymbol{F}_{t+1}^i\}$ from the backbone, we subtract $\boldsymbol{F}_t^i$ from both $\boldsymbol{F}_{t+1}^i$ and $\boldsymbol{F}_{t-1}^i$ to obtain $\{\boldsymbol{M}_f, \boldsymbol{M}_b\}$. Adaptive Deformable Cross blocks (ADC) take the concatenation of $\boldsymbol{M}_f$ and $\boldsymbol{M}_b$, along with the enhanced feature $\boldsymbol{F}$ from HKME. After entering the ADC block, $\boldsymbol{M}_f$ and $\boldsymbol{M}_b$ are first split, and then pass through a dual-branch structure that includes a deformable cross attention (DCA) and a cross attention. The results from the dual branches are concatenated and sent into an MLP for nonlinear transformation. After the final block, a fusion layer is applied to integrate the bidirectional motion features to obtain an aggregated motion representation $\boldsymbol{M}$.

Deformable cross attention. Our deformable cross attention (DCA) predicts multiple offsets at a single point, rather than predicting offsets at each point of the kernel as in the case of deformable convolution. This endows it with a stronger ability to characterize the relationships between elements and to flexibly handle different scales. The concept of our cross mechanism is realized by incorporating the enhanced feature $\boldsymbol{F}$ as a constraint to control the generation of offsets in the spatial domain, ensuring that only a subset of motion features are selected as keys and values for attention computation. Specifically, the DCA can be represented by the following formulas:

$$\begin{aligned} q &= W_q \otimes x, \qquad z = W_z \otimes \boldsymbol{F}, \\ \triangle p &= \theta(\psi(q \oplus z)), \quad \overline{x} = \phi(\triangle p + p_z), \end{aligned} \tag{3}$$

$$\mathrm{DCA}(x, z, p_z) = \sum_{n=1}^{N} \sigma\left(\frac{q \otimes (W_k \otimes \overline{x}_n)^T}{\sqrt{d}}\right)(W_v \otimes \overline{x}_n), \tag{4}$$

where x, q, $\triangle p$, p_z, $\overline{x}$, N, and d are motion features $\boldsymbol{M}_f$ or $\boldsymbol{M}_b$, query, point offset, reference points from z, sample

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
PoseTracker (Girdhar et al. 2018)	67.5	70.2	62.0	51.7	60.7	58.7	49.8	60.6
PoseFlow (Xiu et al. 2018)	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
HRNet (Sun et al. 2019)	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
CorrTrack (Rafi et al. 2020)	86.1	87.0	83.4	76.4	77.3	79.2	73.3	80.8
PoseWarper (Bertasius et al. 2019)	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
DCPose (Liu et al. 2021)	88.0	88.7	84.1	78.4	83.0	81.4	74.2	82.8
SLT-Pose (Gai et al. 2023)	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2
KPM (Fu et al. 2023)	89.5	90.0	87.6	81.8	81.1	82.6	76.1	84.6
M-HANet (Jin et al. 2024)	90.3	90.7	85.3	79.2	83.4	82.6	77.8	84.8
FAMI-Pose (Liu et al. 2022a)	89.6	90.1	86.3	80.0	84.6	83.4	77.0	84.8
DSTA (He and Yang 2024)	89.3	90.6	87.3	82.6	84.5	85.1	77.8	85.6
TDMI-ST (Feng et al. 2023)	**90.6**	91.0	87.2	81.5	85.2	84.5	78.7	85.9
VREMD (Ours)	89.9	**91.4**	**88.8**	**84.8**	**88.5**	**87.8**	**81.0**	**87.6**

Table 1: Comparisons with the state-of-the-art methods for video pose estimation on the validation sets of the **PoseTrack2017** (Iqbal, Milan, and Gall 2017) dataset. Note that we aggregate temporal information from neighboring frames (*i.e.*, one frame to the left and one to the right).

features, number of sampling points, and embedding dimension, respectively. $\otimes$, $\oplus$, $\psi(\cdot)$, $\theta(\cdot)$, $\phi(\cdot)$, and $\sigma(\cdot)$ denote the operations of matrix multiplication, concatenation, convolution, offset generation, bilinear interpolation, softmax, respectively. W_q, W_k, W_v, and W_z are all learnable mapping matrices. The offset $\triangle p$ generated under the constraint of $\boldsymbol{F}$, ensures the filtering of spatial regions related to the human joints within the global domain, thereby facilitating adaptive motion cue extraction from motion features.

Heatmap generation. We first split the key frame keypoint tokens $\boldsymbol{T}_t^i$ from $\widehat{\boldsymbol{T}}_k$ and then transform them into $\boldsymbol{H}_k$ through an MLP and reshaping. By aggregating $\boldsymbol{M}$ and $\boldsymbol{F}_t^i$ and up-sampling, we obtain $\boldsymbol{H}_m$. The final pose heatmaps $\boldsymbol{H}_t^i$ are derived by adding $\boldsymbol{H}_k$ and $\boldsymbol{H}_m$ with equal weights.

Loss function. We adopt the established pose heatmap loss $\mathcal{L}_{\mathrm{H}}$ to supervise the final predicted pose heatmaps $\boldsymbol{H}_t^i$ to converge to the ground truth pose heatmaps $\boldsymbol{G}_t^i$:

$$\mathcal{L}_{\mathrm{H}} = \left\| \boldsymbol{H}_t^i - \boldsymbol{G}_t^i \right\|_2^2. \tag{5}$$

Experiments

Experimental Settings

Datasets. PoseTrack has become a crucial dataset in video-based human pose estimation benchmarks. **PoseTrack2017**(Iqbal, Milan, and Gall 2017) introduces 250 training videos and 50 validation videos, with 80,144 pose annotations across 15 key points. **PoseTrack2018**(Andriluka et al. 2018) expands to 593 training and 170 validation videos, totaling 153,615 annotations. **PoseTrack2021** (Doering et al. 2022) further enriches the dataset, particularly improving the representation of smaller figures and crowded scenes, reaching 177,164 pose annotations, with recalibrated joint visibility flags to better address occlusions.

Evaluation metric. To evaluate the efficacy of our proposed model in pose estimation, we calculate the average precision (AP) for each joint and then aggregate these values to obtain the mean average precision (mAP).

Implementation details. Our VREMD framework is realized utilizing PyTorch. For feature extraction on single frames, we adopt the most primitive Vision Transformer (ViT-L) architecture (Dosovitskiy et al. 2020; Xu et al. 2022), pre-trained on the COCO dataset (Lin et al. 2014), as our backbone. The input image size is fixed at 256×192. We integrate a series of data augmentation techniques, consistent with methodologies employed in previous works (Bertasius et al. 2019; Liu et al. 2021), comprising random rotation $[-45°, 45°]$, random scale [0.65, 1.35], truncation (half body), and flipping during training. The number of input frames is set to 3, consisting of one key frame accompanied by two auxiliary frames sourced from preceding and succeeding neighbors, respectively. This configuration mirrors that of DCPose (Liu et al. 2021), rather than employing the five frame input as seen in TDMI (Feng et al. 2023) and FAMI-Pose (Liu et al. 2022a). Our model is trained on a single RTX 4090 GPU for 20 epochs with the backbone frozen. We utilize the AdamW optimizer with an initial learning rate of 2e-3, which is then reduced by a factor of ten at the 16th epoch.

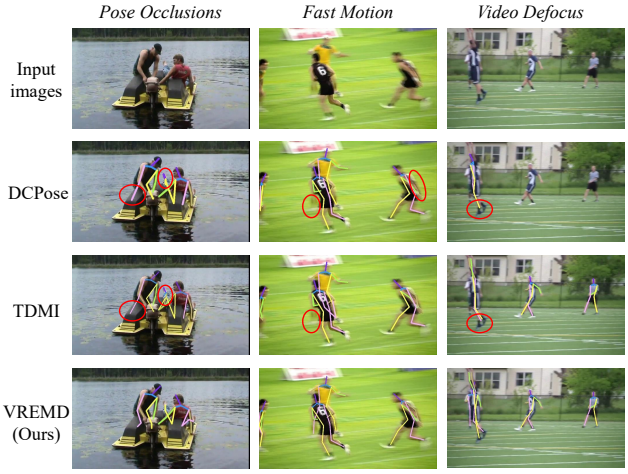

Figure 3: Qualitative comparison of our VREMD, DCPose (Liu et al. 2021), and TDMI (Feng et al. 2023) on the PoseTrack2017 dataset, featuring challenges such as pose occlusions, fast motion, and video defocus. Red solid circles denote the inaccurate pose predictions.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
AlphaPose (Fang et al. 2017)	63.9	78.7	77.4	71.0	73.7	73.0	69.7	71.9
PoseWarper (Bertasius et al. 2019)	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
DCPose (Liu et al. 2021)	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
FAMI-Pose (Liu et al. 2022a)	85.5	87.7	84.2	79.2	81.4	81.1	74.9	82.2
M-HANet (Jin et al. 2023)	86.7	88.9	84.6	79.2	79.7	81.3	78.7	82.7
DSTA (He and Yang 2024)	85.9	88.8	85.0	81.1	81.5	83.0	77.4	83.4
TDMI-ST (Feng et al. 2023)	**86.7**	88.9	85.4	80.6	82.4	82.1	77.6	83.6
VREMD (Ours)	**86.7**	**89.3**	**85.6**	**82.1**	**85.0**	**83.9**	**79.3**	**84.6**

Table 2: Comparisons with the state-of-the-art methods for video pose estimation on the validation sets of the **PoseTrack2018** (Andriluka et al. 2018) dataset.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
DCPose (Liu et al. 2021)	83.2	84.7	82.3	78.1	80.3	79.2	73.5	80.5
FAMI-Pose (Liu et al. 2022a)	83.3	85.4	82.9	78.6	81.3	80.5	75.3	81.2
DSTA (He and Yang 2024)	**87.5**	87.0	84.2	81.4	82.3	82.5	77.7	83.5
TDMI-ST (Feng et al. 2023)	86.8	87.4	85.1	81.4	83.8	82.7	78.0	83.8
VREMD (Ours)	87.2	**89.1**	**85.2**	**82.4**	**85.1**	**83.4**	**79.2**	**84.5**

Table 3: Comparisons with the state-of-the-art methods for video pose estimation on the validation sets of the **PoseTrack2021** (Doering et al. 2022) dataset.

Method	HKME	BMD	Mean
Baseline			80.2
(a)	✓		85.3
(b)		✓	85.6
(c)	✓	✓	**87.6**

Table 4: Ablation of different components in our **VREMD**.

Method	Human mask	Keypoint mask	Mean
(a)			85.9
(b)	✓		86.5
(c)		✓	86.8
(d)	✓	✓	**87.6**

Table 5: Ablation of various designs in the **HKME** module.

Method	DC	DA	DCA (Ours)	BS (Ours)	Mean
(a)	✓				84.7
(b)		✓			85.8
(c)			✓		87.1
(d)			✓	✓	**87.6**

Table 6: Ablation of various designs in the **BMD** module.

Comparison with State-of-the-art Approaches

Results on the PoseTrack2017 Dataset. We first benchmark our method on the PoseTrack2017 (Iqbal, Milan, and Gall 2017) dataset. A total of 12 methods are compared and their performances on the PoseTrack2017 validation set are summarized in Table 1. Our proposed VREMD consistently outperforms existing state-of-the-art methods, reaching an mAP of 87.6. Compared to the latest top-performing method TDMI-ST (Feng et al. 2023), our VREMD obtain a 1.7 mAP gain. The performance boost for challenging joints (*i.e.*, wrist, ankle) is also promising: we attain an mAP of 84.8 (↑ 3.3) for wrists and an mAP of 81.0 (↑ 2.3) for ankles. It is noteworthy that our VREMD operates effectively with fewer input video frames than the most recent works (Liu et al. 2022a; Feng et al. 2023), requiring just three frames as opposed to five. These consistent and substantial improvements in effectiveness indicate the importance of reinforcing the positional attributes of visual representations and integrating joint-related motion dynamics. In addition, we present the visualized results, which include a comparison with existing methods, for scenarios involving complex spatio-temporal interactions (*e.g.*, pose occlusion, blur) in Fig 3, demonstrating our method's robustness.

Results on the PoseTrack2018 Dataset. We further evaluate our VREMD on the PoseTrack2018 dataset, and the detailed validation set results are showcased in Table 2. Once again, as illustrated in this table, our VREMD surpasses all prior state-of-the-art methods, achieving the most exceptional outcomes. We obtain the final performance of 84.6 mAP. The precision for wrists and ankles also shows a noticeable improvement compared to TDMI-ST, scoring 82.1 (↑ 1.5) and 79.3 (↑ 1.7) respectively.

Results on the PoseTrack2021 Dataset. Performance comparisons of our model and previous state-of-the-art methods on the PoseTrack21 dataset are provided in Table 3. When evaluated on the PoseTrack2021 validation dataset, the results highlight the outstanding performance of our model. Achieving new state-of-the-art results, our model records an overall mAP of 84.5, outperforming TDMI-ST by a margin of 0.7 mAP. Encouragingly, our method yields a 1.0 mAP improvement over the previous best, attaining 82.4 at the wrist, and shows a 1.2 mAP advance, achieving 79.2 at the ankle, which are recognized as difficult joints to accurately predict. These results, once again, underscore the robustness and superiority of our method in this domain.

Ablation Study

We carry out extensive ablation studies centered on assessing the impact of individual components within our VREMD architecture, encompassing the Human-Keypoint Mask Enhancement module (HKME) and the Bidirectional Motion Disentanglement module (BMD). We additionally probe

into the efficacy of diverse micro-designs incorporated in each module. All experiments are performed on the PoseTrack2017 validation set.

Study on components of VREMD. We experimentally evaluate the effectiveness of each component in our VREMD framework, detailing the quantitative results in Table 4. Firstly, we establish a baseline for this experiment by coupling a Vision Transformer (ViT) Backbone with a pose detection head. (a) Integrating the Human-Keypoint Mask Enhanced module (HKME) into the baseline yields a substantial gain of 5.1 mAP. This substantial progress indicates that the dual-mask mechanism, offering a coarse-to-fine representation refinement, facilitates improvements in human pose estimation. (b) In the next setup, we exclusively incorporate the Bidirectional Motion Disentanglement module (BMD) into the baseline system. Notably, the Adaptive Deformable Cross (ADC) block, which originally utilized enhanced features from the HKME, now receives backbone output features instead. The outcome achieves an mAP of 85.6, marking an increase of 5.4 mAP. Such a significant boost in performance unequivocally validates the BMD module's proficiency in adaptively excavating bidirectional temporal information, guiding accurate pose estimation. (c) Finally, we incorporate both the HKME and BMD modules into our framework, attaining a culminating performance of 87.6 mAP, which indicates that the synergy of these two components can lead to further enhancements.

Study on Human-Keypoint Mask Enhanced module. We then investigate the impact of the two mask generation techniques in HKME on overall performance. We conduct four experiments and presented them in Table 5. (a) Generating visual representations using only the spatio-temporal Transformers network. (b) Producing a human mask for coarse filtering of human-related tokens. (c) Calculating a keypoint mask for basic joint token screening. (d) Utilizing dual masks, derived from methods (b) and (c), for the progressive refinement and enhancement of visual tokens, transitioning from coarse to fine detail. This table illustrates that method (a), which does not generate any masks, offers a slight improvement of 0.3 mAP over the setting that removes HKME. Subsequently, applying the human mask alone (b) and the keypoint mask alone (c) achieves respective performances of 86.5 mAP and 86.8 mAP. Although utilizing these masks individually can yield certain accuracy gains, simultaneously employing both for coarse-to-fine representation refinement (d) leads to the optimal results. This promising outcome attests to the superiority of our dual-mask paradigm, which provides a prompt of human joints to the framework, enabling more accurate keypoint localization.

Study on Bidirectional Motion Disentanglement module. Additionally, we explore the influence of our deformable cross attention (DCA) and bidirectional separation strategy. Four experiments are performed and displayed in Table 6. (a) We first replace our Adaptive Deformable Cross (ADC) block with the deformable conv (DC) (Dai et al. 2017), as adopted in previous works (Liu et al. 2021, 2022a; Feng et al. 2023). We observe a slight performance decline, that is, a 0.6 mAP decrease. We speculate that the reason might be the feature map obtained through the attention mechanism is more spatially dispersed and structurally diverse, which is incompatible with the local adaptive variation characteristics of deformable conv. (b) We further employ plain deformable attention (DA) (Zhu et al. 2020) and achieve an 85.8 mAP, which proves that deformable attention is more suitable for our frameworks based on attention mechanisms. (c) We propose a novel deformable cross attention (DCA), which integrates the advantages of adaptive receptive field of deformable attention and selective feature highlighting of cross attention, achieving an 87.1 mAP. (d) Finally, we apply a bidirectional separation (BS) strategy to independently capture bidirectional motion dynamics, resulting in a 0.5 mAP improvement, unlike previous methods that concatenate and jointly process bidirectional motion features. These results strongly demonstrate that our method can more effectively capture task-relevant motion cues to facilitate pose estimation.

Related Work

Video-based human pose estimation. In the early stages, substantial approaches involve utilizing optical flow to establish motion-based assumptions (Pfister, Charles, and Zisserman 2015). These approaches commonly generate dense optical flow across frames to improve pose heatmap predictions, yet the technique is computationally demanding and prone to errors when faced with marked deterioration in image quality. Recent methods (He and Yang 2024) have shifted towards attempting to implicitly capture motion evidence from temporal information by employing deformable convolutions. DCPose (Liu et al. 2021) and PoseWarper (Bertasius et al. 2019) model and process pose temporal residuals and re-refine keypoint detection via multi-scale deformable convolutions for accurate pose estimation. TDMI (Feng et al. 2023) introduces a multi-stage framework that encodes temporal differences for dynamic context modeling, leveraging mutual information to uncover useful temporal clues. Contrary to prior approaches that directly execute feature difference learning in the global space, we strive to enhance visual representations through the aggregation of joint positions, and to dissect representative joint-associated motion dynamics for more robust human pose estimation.

Conclusion

In this paper, we investigate the video-based human pose estimation task from the perspective of local spatial perception and temporal cues disentanglement. A dual-stream architecture is designed to effectively capture spatio-temporal dependencies by collaboratively executing gradual human joint focus and adaptive motion decoupling. Specifically, we present a Human-Keypoint Mask Enhanced module that performs a coarse-to-fine selective representation enhancement to assist the framework in exploring human and joint regions. Additionally, we create a Bidirectional Motion Disentanglement module to adaptively mine pose-related motion evidence. Our method significantly and consistently outperforms state-of-the-art performances on three benchmark datasets: PoseTrack2017, PoseTrack2018, and PoseTrack2021.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62372402), and the Key R&D Program of Zhejiang Province (No. 2023C01217).

References

Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; and Schiele, B. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5167–5176.

Bertasius, G.; Feichtenhofer, C.; Tran, D.; Shi, J.; and Torresani, L. 2019. Learning temporal pose estimation from sparsely-labeled videos. *Advances in neural information processing systems*, 32.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Doering, A.; Chen, D.; Zhang, S.; Schiele, B.; and Gall, J. 2022. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20963–20972.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, 2334–2343.

Feng, R.; Gao, Y.; Ma, X.; Tse, T. H. E.; and Chang, H. J. 2023. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17131–17141.

Fu, Z.; Zuo, W.; Hu, Z.; Liu, Q.; and Wang, Y. 2023. Improving Multi-Person Pose Tracking with A Confidence Network. *IEEE Transactions on Multimedia*.

Gai, D.; Feng, R.; Min, W.; Yang, X.; Su, P.; Wang, Q.; and Han, Q. 2023. Spatiotemporal learning transformer for video-based human pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9): 4564–4576.

Geng, Z.; Wang, C.; Wei, Y.; Liu, Z.; Li, H.; and Hu, H. 2023. Human pose as compositional tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 660–671.

Girdhar, R.; Gkioxari, G.; Torresani, L.; Paluri, M.; and Tran, D. 2018. Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 350–359.

He, J.; and Yang, W. 2024. Video-Based Human Pose Regression via Decoupled Space-Time Aggregation. *arXiv preprint arXiv:2403.19926*.

Iqbal, U.; Milan, A.; and Gall, J. 2017. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011–2020.

Jin, K.-M.; Lee, G.-H.; and Lee, S.-W. 2022. OTPose: occlusion-aware transformer for pose estimation in sparsely-labeled videos. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3255–3260. IEEE.

Jin, K.-M.; Lee, G.-H.; Nam, W.-J.; Kang, T.-K.; Kim, H.-W.; and Lee, S.-W. 2024. Masked Kinematic Continuity-aware Hierarchical Attention Network for pose estimation in videos. *Neural Networks*, 169: 282–292.

Jin, K.-M.; Lim, B.-S.; Lee, G.-H.; Kang, T.-K.; and Lee, S.-W. 2023. Kinematic-aware hierarchical attention network for human pose estimation in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5725–5734.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, 280–296. Springer.

Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.-T.; and Zhou, E. 2021. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, 11313–11322.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, Z.; Chen, H.; Feng, R.; Wu, S.; Ji, S.; Yang, B.; and Wang, X. 2021. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 525–534.

Liu, Z.; Feng, R.; Chen, H.; Wu, S.; Gao, Y.; Gao, Y.; and Wang, X. 2022a. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11006–11016.

Liu, Z.; Wu, S.; Xu, C.; Wang, X.; Zhu, L.; Wu, S.; and Feng, F. 2022b. Copy motion from one to another: Fake motion video generation. *arXiv preprint arXiv:2205.01373*.

Pfister, T.; Charles, J.; and Zisserman, A. 2015. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE international conference on computer vision*, 1913–1921.

Rafi, U.; Doering, A.; Leibe, B.; and Gall, J. 2020. Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, 36–52. Springer.

Sapp, B.; Toshev, A.; and Taskar, B. 2010. Cascaded models for articulated pose estimation. In *Computer Vision–ECCV*

2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11, 406–420. Springer.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.

Su, P.; Liu, Z.; Wu, S.; Zhu, L.; Yin, Y.; and Shen, X. 2021. Motion prediction via joint dependency modeling in phase space. In *Proceedings of the 29th ACM international conference on multimedia*, 713–721.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.

Tse, T. H. E.; De Martini, D.; and Marchegiani, L. 2019. No need to scream: Robust sound-based speaker localisation in challenging scenarios. In *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*, 176–185. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, D.; and Zhang, S. 2022. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11060–11068.

Wang, Y.; and Mori, G. 2008. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*, 710–724. Springer.

Wu, S.; Chen, H.; Yin, Y.; Hu, S.; Feng, R.; Jiao, Y.; Yang, Z.; and Liu, Z. 2024a. Joint-Motion Mutual Learning for Pose Estimation in Video. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8962–8971.

Wu, S.; Liu, Z.; Zhang, B.; Zimmermann, R.; Ba, Z.; Zhang, X.; and Ren, K. 2024b. Do as I Do: Pose Guided Human Motion Copy. *IEEE Transactions on Dependable and Secure Computing*.

Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; and Lu, C. 2018. Pose Flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977*.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35: 38571–38584.

Yang, Y.; Chen, H.; Liu, Z.; Lyu, Y.; Zhang, B.; Wu, S.; Wang, Z.; and Ren, K. 2023. Action recognition with multi-stream motion modeling and mutual information maximization. *arXiv preprint arXiv:2306.07576*.

Zhao, Y.; Xiong, Y.; and Lin, D. 2018. Recognize actions by disentangling components of dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6566–6575.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.