Mutual Information-Based Temporal Difference Learning for Human Pose Estimation in Video

Runyang Feng[1,2], Yixing Gao[1,2,*], Xueqing Ma[1,2], Tze Ho Elden Tse[3], Hyung Jin Chang[3]

[1] School of Artificial Intelligence, Jilin University,
[2] Engineering Research Center of Knowledge-Driven Human-Machine Intelligence,
Ministry of Education, China, [3]School of Computer Science, University of Birmingham

{fengry22, maxq21}@mails.jlu.edu.cn, gaoyixing@jlu.edu.cn,
txt994@student.bham.ac.uk, h.j.chang@bham.ac.uk

Abstract

Temporal modeling is crucial for multi-frame human pose estimation. Most existing methods directly employ optical flow or deformable convolution to predict full-spectrum motion fields, which might incur numerous irrelevant cues, such as a nearby person or background. Without further efforts to excavate meaningful motion priors, their results are suboptimal, especially in complicated spatio-temporal interactions. On the other hand, the temporal difference has the ability to encode representative motion information which can potentially be valuable for pose estimation but has not been fully exploited. In this paper, we present a novel multi-frame human pose estimation framework, which employs temporal differences across frames to model dynamic contexts and engages mutual information objectively to facilitate useful motion information disentanglement. To be specific, we design a multi-stage Temporal Difference Encoder that performs incremental cascaded learning conditioned on multi-stage feature difference sequences to derive informative motion representation. We further propose a Representation Disentanglement module from the mutual information perspective, which can grasp discriminative task-relevant motion signals by explicitly defining useful and noisy constituents of the raw motion features and minimizing their mutual information. These place us to rank No.1 in the Crowd Pose Estimation in Complex Events Challenge on benchmark dataset HiEve, and achieve state-of-the-art performance on three benchmarks PoseTrack2017, PoseTrack2018, and PoseTrack21.

1. Introduction

Human pose estimation has long been a nontrivial and fundamental problem in the computer vision community.
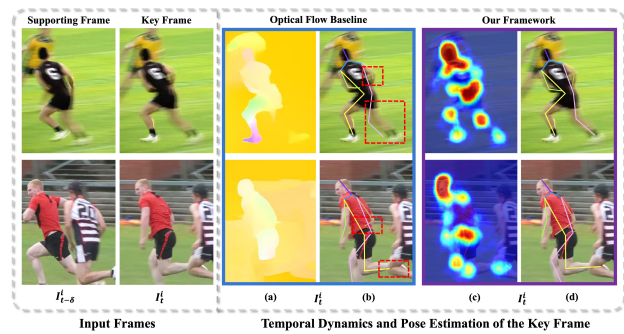
---

*Corresponding Author

Figure 1. Directly leveraging optical flow can be distracted by irrelevant clues such as background and blur (a), and sometimes fails in scenarios with fast motion and mutual occlusion (b). Our proposed framework proceeds with temporal difference encoding and useful information disentanglement to capture more tailored temporal dynamics (c), yielding more robust pose estimations (d).

The goal is to localize anatomical keypoints (*e.g.*, nose, ankle, etc.) of human bodies from images or videos. Nowadays, as more and more videos are recorded endlessly, video-based human pose estimation has been extremely desired in enormous applications including live streaming, augmented reality, surveillance, and movement tracking [14, 15, 24, 35, 44].

An extensive body of literature focuses on estimating human poses in *static images*, ranging from earlier methods employing pictorial structure models [43, 51, 55, 69] to recent attempts leveraging deep convolutional neural networks [35, 49, 56, 58] or Vision Transformers [32, 61, 65]. Despite the impressive performance in still images, the extension of such methods to video-based human pose estimation still remains challenging due to the additional temporal dimension in videos [34, 57]. By nature, the video presents distinctive and valuable *dynamic contexts* (*i.e.*, the temporal evolution in the visual content) [71]. Therefore, being able to effectively utilize the temporal dynamics (motion information) is fundamentally important for accurate pose

estimation in videos [35].

One line of work [35, 37, 54] attempts to derive a unified spatial-temporal representation through *implicit* motion compensation. [54] presents a 3DHRNet which utilizes 3D convolutions to extract spatiotemporal features of a video tracklet to estimate pose sequences. [35] adopts deformable convolutions to align multi-frame features and aggregates aligned feature maps to predict human poses. On the other hand, [40, 45, 67] *explicitly* model motion contexts with optical flow. [40, 45] propose to compute dense optical flow between every two frames and leverage the flow features for refining pose heatmaps temporally across multiple frames.

Upon studying the previous methods [34, 35, 40, 45], we empirically observe that the pose estimation performance is boosted with the implicit or explicit imposition of motion priors. However, the movement of any visual evidence is usually attended to in these paradigms, resulting in cluttered motion features that include numerous irrelevant information (*e.g.*, nearby person, background), as illustrated in Fig. 1. Directly exploiting such vanilla motion features delivers inferior results, especially in complex scenarios of mutual occlusion and fast motion. More specifically, not all pixel movements are equally important in video-based human pose estimation [66]. For example, background variations and pixel changes caused by image quality degradation (*e.g.*, blur and occlusion) are usually useless and distracting, whereas the salient pixel movements driven by human body motions play a more important role in understanding motion patterns [21]. Therefore, discovering meaningful motion dynamics is crucial to fully recovering human poses across a video. On the other hand, investigating temporal differences across video frames allows one to discover representative motion cues [25, 52, 59]. Although it has already shown success in various video-related tasks (action recognition [52], video super-resolution [24]), its application on video-based human pose estimation remains under-explored.

In this paper, we present a novel framework, named **T**emporal **D**ifference Learning based on **M**utual **I**nformation (TDMI) for human pose estimation. Our TDMI consists of two key components: **(i)** A multi-stage Temporal Difference Encoder (TDE) is designed to model motion contexts conditioned on multi-stage feature differences among video frames. Specifically, we first compute the feature difference sequences across multiple stages by leveraging a temporal difference operator. Then, we perform incremental cascaded learning via intra- and inter-stage feature integration to derive the motion representation. **(ii)** We further introduce a Representation Disentanglement module (RDM) from the mutual information perspective, which distills the task-relevant motion features to enhance the frame representation for pose estimation. In particular, we first disentangle the useful

and noisy constituents of the vanilla motion representation by activating corresponding feature channels. Then, we theoretically analyze the statistical dependencies between the useful and the noisy motion features and arrive at an information-theoretic loss. Minimizing this mutual information objective encourages the useful motion components to be more discriminative and task-relevant. Our approach achieves significant and consistent performance improvements over current state-of-the-art methods on four benchmark datasets. Extensive ablation studies are conducted to validate the efficacy of each component in the proposed method.

The main contributions of this work can be summarized as follows: (1) We propose a novel framework that leverages temporal differences to model dynamic contexts for video-based human pose estimation. (2) We present a disentangled representation learning strategy to grasp discriminative task-relevant motion signals via an information-theoretic objective. (3) We demonstrate that our approach achieves new state-of-the-art results on four benchmark datasets, PoseTrack2017, PoseTrack2018, PoseTrack21, and HiEve.

2. Related Work

Image-based human pose estimation. With the recent advances in deep learning architectures [18, 50] as well as the availability of large-scale datasets [1, 11, 23, 33], various deep learning methods [2, 10, 32, 46, 58, 61, 65] are proposed for image-based human pose estimation. These approaches broadly fall into two paradigms: bottom-up and top-down. *Bottom-up* approaches [7, 29, 30] detect individual body parts and associate them with an entire person. [30] proposes a composite framework that employs a Part Intensity Field to localize human body parts and uses a Part Association Field to associate the detected body parts with each other. Conversely, *top-down* approaches [13, 32, 46, 56, 58] detect bounding boxes of persons first and predict human poses within each bounding box region. [46] presents a high-resolution convolutional architecture that preserves high-resolution features in all stages, demonstrating superior performance for human pose estimation.

Video-based human pose estimation. Existing image-based methods could not generalize well to video streams since they inherently have difficulties in capturing temporal dynamics across frames. A direct approach would be to leverage optical flow to impose motion priors [40, 45]. These approaches typically compute dense optical flow among frames and leverage such motion cues to refine the predicted pose heatmaps. However, the optical flow estimation is computationally intensive and tends to be vulnerable when encountering severe image quality degradation. Another approach [5, 34, 35, 54] considers implicit motion compensation using deformable convolutions or

3DCNNs. [5, 34] propose to model multi-granularity joint movements based on heatmap residuals and perform pose resampling or pose warping through deformable convolutions. As the above cases generally consider motion details from all pixel locations, their resulting representations are suboptimal for accurate pose estimation.

Temporal difference modeling. Temporal difference operations, *i.e.*, RGB Difference (image-level) [38, 52, 53, 71] and Feature Difference (feature-level) [25, 31, 36], are typically exploited for motion extraction, showing outstanding performance with high efficiency for many video-related tasks such as action recognition [31, 52] and video super-resolution [24]. [38, 52, 71] leverage RGB difference as an efficient alternative modality to optical flow to represent motions. [24] proposes to explicitly model temporal differences in both LR and HR space. However, the additional RGB difference branch usually replicates the feature-extraction backbone, which increases the model complexity. On the other hand, [25, 31, 36] employ a feature difference operation for network design which our work falls within this scope more closely. In contrast to previous methods that simply compute feature differences, we seek to disentangle discriminative task-relevant temporal difference representations for pose estimation.

3. Our Approach

Preliminaries. Our work follows the top-down paradigm, starting with an object detector to obtain the bounding boxes for individual persons in a video frame I_t. Then, each bounding box is enlarged by 25% to crop the same individual in a consecutive frame sequence $\boldsymbol{\mathcal{X}}_t = \langle I_{t-\delta}, ..., I_t, ..., I_{t+\delta} \rangle$ with δ being a predefined temporal span. In this way, we attain the cropped video clip $\boldsymbol{\mathcal{X}}_t^i = \langle I_{t-\delta}^i, ..., I_t^i, ..., I_{t+\delta}^i \rangle$ for person i.

Problem formulation. Given a cropped video segment $\boldsymbol{\mathcal{X}}_t^i$ centered on the key frame I_t^i, we are interested in estimating the pose in I_t^i. Our goal is to better leverage frame sequences through principled temporal difference learning and useful information disentanglement, thereby addressing the common shortcoming of existing methods in failing to adequately mine motion dynamics.

Method overview. The overall pipeline of the proposed TDMI is outlined in Fig. 2. Our framework consists of two key components: a multi-stage Temporal Difference Encoder (TDE) (Sec. 3.1) and a Representation Disentanglement module (RDM) (Sec. 3.2). Specifically, we first extract visual features of the input sequence and feed them to TDE, which computes feature differences and performs information integration to obtain the motion feature $\mathbf{M}_t^i$. Then, RDM takes motion feature $\mathbf{M}_t^i$ as input and excavates its useful constituents to yield $\mathbf{M}_{t,u}^i$. Finally, both the motion feature $\mathbf{M}_{t,u}^i$ and the visual feature of the key frame are aggregated to produce the enhanced representation $\tilde{\mathbf{F}}_t^i$.

$\tilde{\mathbf{F}}_t^i$ is handed to a detection head which outputs the pose estimation $\mathbf{H}_t^i$. In the following, we explain the two key components in depth.

3.1. Multi-Stage Temporal Difference Encoder

As multi-stage feature integration enables the network to retain diverse semantic information from fine to coarse scale [26, 39, 66], we propose to simultaneously aggregate shallow feature differences (early stages) compressing detailed motion cues and deep feature differences (late stages) encoding global semantic movements to derive informative and fine-grained motion representations. A naive approach to fuse features in multiple stages is to feed them into a convolutional network [9, 28]. However, this simple fusion solution suffers from two drawbacks: (i) redundant features might be over-emphasized, and (ii) fine-grained cues of each stage cannot be fully reserved. Motivated by these observations and insights, we present a multi-stage temporal difference encoder (TDE) with an incremental cascaded learning architecture, addressing the above issues through two designs: a *spatial modulation* mechanism to adaptively focus on important information at each stage, and a *progressive accumulation* mechanism to preserve fine-grained contexts across all stages.

Specifically, given an image sequence $\boldsymbol{\mathcal{X}}_t^i = \langle I_{t-\delta}^i, ..., I_t^i, ..., I_{t+\delta}^i \rangle$, our proposed TDE first constructs multi-stage feature difference sequences and performs both intra- and inter-stage feature fusion to yield the encoded motion representation $\mathbf{M}_t^i$. For simplicity, we take $\delta = 1$ in the following.

Feature difference sequences generation. We build TDE upon the HRNet-W48 [46] network, which includes four convolutional stages to extract feature maps of the input sequence $\{\mathbf{F}_{t-1}^{i,J}, \mathbf{F}_t^{i,J}, \mathbf{F}_{t+1}^{i,J}\}$. The superscript $J = \{1, 2, 3, 4\}$ refers to network stages. Subsequently, we compute the consecutive feature difference sequences $\boldsymbol{\mathcal{S}}_t^{i,J} = \{\mathbf{S}^1, \mathbf{S}^2, \mathbf{S}^3, \mathbf{S}^4\}$ over four stages as follows:

$$\mathbf{S}^j = \left\{ \mathbf{F}_t^{i,j} - \mathbf{F}_{t-1}^{i,j}, \mathbf{F}_{t+1}^{i,j} - \mathbf{F}_t^{i,j} \right\}, j = 1, ..., 4. \quad (1)$$

Intra-stage feature fusion. Given the feature difference sequences $\boldsymbol{\mathcal{S}}_t^{i,J}$, several residual blocks [18] are leveraged to separately aggregate the feature elements within each stage to generate stage-specific motion representations $\boldsymbol{\mathcal{D}}_t^{i,J} = \{\mathbf{D}^1, \mathbf{D}^2, \mathbf{D}^3, \mathbf{D}^4\}$. This computation can be expressed as:

$$\mathbf{D}^j = \mathbf{Conv}\left[\left(\mathbf{F}_t^{i,j} - \mathbf{F}_{t-1}^{i,j} \right) \oplus \left(\mathbf{F}_{t+1}^{i,j} - \mathbf{F}_t^{i,j} \right) \right], \quad (2)$$

where $\oplus$ is the concatenation operation and $\mathbf{Conv}(\cdot)$ is the function of convolutional blocks. In practice, we employ $\{3, 3, 2, 2\}$ residual blocks with kernel size 3×3 to aggregate the features at corresponding stages, respectively.
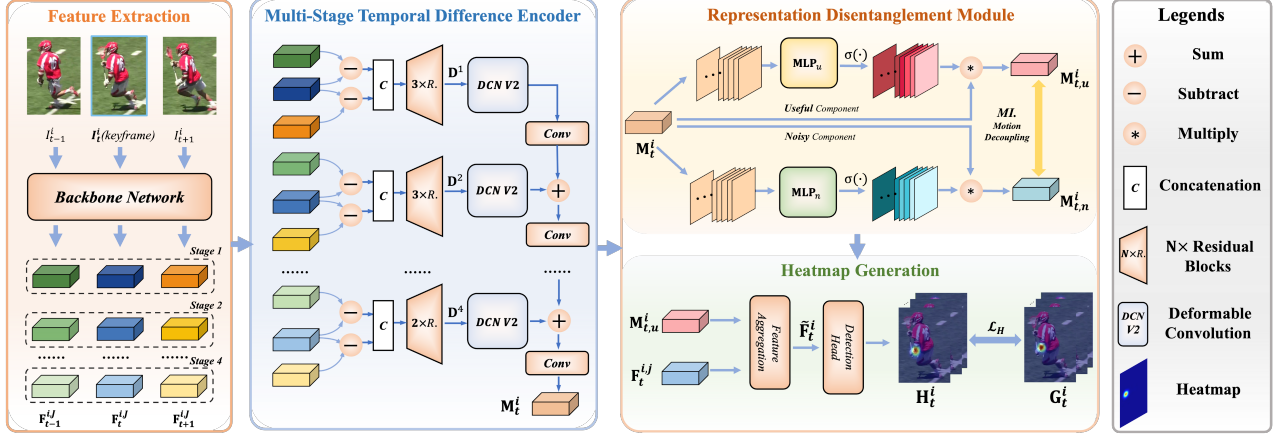
Figure 2. Overall pipeline of the proposed framework. The goal is to detect the human pose of the key frame I_t^i. Given an input sequence $\langle I_{t-1}^i, I_t^i, I_{t+1}^i \rangle$, we first extract their visual features $\{\mathbf{F}_{t-1}^{i,J}, \mathbf{F}_t^{i,J}, \mathbf{F}_{t+1}^{i,J}\}$. Our multi-stage Temporal Difference Encoder takes these features as input and outputs the motion feature $\mathbf{M}_t^i$. Then, $\mathbf{M}_t^i$ is handed to the Representation Disentanglement module which performs useful information disentanglement and outputs $\mathbf{M}_{t,u}^i$. Finally, the motion feature $\mathbf{M}_{t,u}^i$ and the visual feature $\mathbf{F}_t^{i,j}$ and are used to obtain the final pose estimation $\mathbf{H}_t^i$.

Inter-stage feature fusion. After obtaining the motion features of each stage $\mathcal{D}_t^{i,J}$, we perform feature integration across stages to obtain the fused motion representation $\mathbf{M}_t^i$, via the proposed *spatial modulation* and *progressive accumulation*. **(1)** We first employ deformable convolutions (DCN V2 [72]) to adaptively modulate the spatial-wise responses of each stage feature. Specifically, given $\mathcal{D}_t^{i,J}$, we independently estimate the kernel sampling offsets $\mathcal{O}$ and modulated scalars $\mathcal{W}$:

$$\{\mathbf{D}^1, \cdots, \mathbf{D}^4\} \xrightarrow[\text{blocks}]{\text{residual}} \xrightarrow[\text{convolution}]{\text{regular}} \{O^1, \cdots, O^4\},$$
$$\{\mathbf{D}^1, \cdots, \mathbf{D}^4\} \xrightarrow[\text{blocks}]{\text{residual}} \xrightarrow[\text{convolution}]{\text{regular}} \{W^1, \cdots, W^4\}. \quad (3)$$

The adaptively learned offsets $\mathcal{O}$ reveal the pixel movement association fields while the modulated scalars $\mathcal{W}$ reflect the magnitude of motion information in each pixel location. Then, we apply deformable convolutions which take the motion features $\mathcal{D}_t^{i,J}$, the kernel offsets $\mathcal{O}$, and the modulated scalars $\mathcal{W}$ as input, and output the spatially calibrated motion features of each stage $\bar{\mathcal{D}}_t^{i,J} = \{\bar{\mathbf{D}}^1, \bar{\mathbf{D}}^2, \bar{\mathbf{D}}^3, \bar{\mathbf{D}}^4\}$:

$$\left(\mathbf{D}^j, O^j, W^j\right) \xrightarrow[\text{convolution}]{\text{deformable}} \bar{\mathbf{D}}^j, j = 1, ..., 4. \quad (4)$$

(2) As shown in Fig. 2, after the spatial modulation, TDE adds the motion feature in the previous stage to the modulated feature, followed by several 3×3 convolutions. Such processing is executed progressively until all stages of features are converged to $\mathbf{M}_t^i$. The above computation is formulated as:

$$\mathbf{M}_t^i = \mathbf{Conv}\left(\cdots\left(\mathbf{Conv}\left(\bar{\mathbf{D}}^1\right) + \cdots + \bar{\mathbf{D}}^4\right)\right). \quad (5)$$

By selectively and comprehensively aggregating multi-stage features, our TDE is able to encode informative and fine-grained motion representation $\mathbf{M}_t^i$.

3.2. Representation Disentanglement Module

Directly leveraging the encoded motion feature $\mathbf{M}_t^i$ for subsequent pose estimation is prone to inevitable task-irrelevant pixel movements (*e.g.*, background, occlusion). To alleviate this limitation, one can train a Vision Attention module end-to-end via a heatmap loss to further distill meaningful motion cues. While being straightforward, the learned features in this approach tend to be plain and undistinguished, which results in limited performance improvement (see Table 7). After manual examination of the extracted temporal dynamic features for pose estimation, it would be fruitful to investigate whether introducing *supervision to the meaningful information distillation* would facilitate the task.

Ideally, incorporating motion information annotations as feature constraints would simplify this problem. Unfortunately, such movement labels in videos are typically absent in most cases. In light of the above observations, we propose to approach this issue through a mutual information viewpoint. In specific, we explicitly define both useful composition $\mathbf{M}_{t,u}^i$ and noisy composition $\mathbf{M}_{t,n}^i$ of the vanilla motion feature $\mathbf{M}_t^i$, in which $\mathbf{M}_{t,u}^i$ is used for subsequent task while $\mathbf{M}_{t,n}^i$ serves as a contrastive landmark. By introducing mutual information supervision to reduce the statistical dependence between $\mathbf{M}_{t,u}^i$ and $\mathbf{M}_{t,n}^i$, we can grasp discriminative task-relevant motion signals. Coupling these network objectives, the representation disentanglement module (RDM) is designed.

Representation factorization. Given the vanilla motion representation $\mathbf{M}_t^i$, we factorize it into useful motion component $\mathbf{M}_{t,u}^i$ and noisy component $\mathbf{M}_{t,n}^i$ by activating corresponding feature channels. Specifically, we first squeeze the global spatial information into a channel de-

scriptor via a global average pooling (GAP) layer. Then, a Multilayer Perceptron (MLP) is used to capture channel-wise interactions followed by a sigmoid function to output an attention mask. This channel-wise attention matrix is finally exploited to rescale the input feature $\mathbf{M}_t^i$ to output $\mathbf{M}_{t,u}^i$ and $\mathbf{M}_{t,n}^i$, respectively. The above factorization process is formulated as:

$$
\begin{aligned}
\mathbf{M}_{t,u}^i &= \sigma\left(\mathbf{MLP}_u\left(\mathbf{GAP}\left(\mathbf{M}_t^i\right)\right)\right) \odot \mathbf{M}_t^i, \\
\mathbf{M}_{t,n}^i &= \sigma\left(\mathbf{MLP}_n\left(\mathbf{GAP}\left(\mathbf{M}_t^i\right)\right)\right) \odot \mathbf{M}_t^i.
\end{aligned}
\tag{6}
$$

The symbol σ denotes the sigmoid function and $\odot$ refers to the channel-wise multiplication. The network parameters of $\mathbf{MLP}_u$ and $\mathbf{MLP}_n$ are learned independently.

Heatmap generation. We integrate the useful motion feature $\mathbf{M}_{t,u}^i$ and the visual feature of the key frame $\mathbf{F}_t^{i,j}$ through several residual blocks to obtain the enhanced representation $\tilde{\mathbf{F}}_t^i$. $\tilde{\mathbf{F}}_t^i$ is fed to a detection head to yield the estimated pose heatmaps $\mathbf{H}_t^i$. We implement the detection head using a 3×3 convolutional layer.

Mutual information objective. Mutual information (MI) measures the amount of information one variable reveals about the other [20, 35]. Formally, the MI between two random variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is defined as:

$$
\mathcal{I}(\boldsymbol{x}_1; \boldsymbol{x}_2) = \mathbb{E}_{p(\boldsymbol{x}_1, \boldsymbol{x}_2)}\left[\log \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2)}{p(\boldsymbol{x}_1)p(\boldsymbol{x}_2)}\right],
\tag{7}
$$

where $p(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is the joint probability distribution between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, while $p(\boldsymbol{x}_1)$ and $p(\boldsymbol{x}_2)$ are their marginals. Within this framework, our main objective for learning effective temporal differences is formulated as:

$$
\min \mathcal{I}\left(\mathbf{M}_{t,u}^i; \mathbf{M}_{t,n}^i\right).
\tag{8}
$$

The term $\mathcal{I}\left(\mathbf{M}_{t,u}^i; \mathbf{M}_{t,n}^i\right)$ rigorously quantifies the amount of information shared between the useful motion feature $\mathbf{M}_{t,u}^i$ and the noisy feature $\mathbf{M}_{t,n}^i$. Recall that $\mathbf{M}_{t,u}^i$ is used for the subsequent pose estimation task. Intuitively, at the beginning of model training, both meaningful and noisy motion cues will be encoded into $\mathbf{M}_{t,u}^i$ and $\mathbf{M}_{t,n}^i$ simultaneously under the constraint of mutual information minimization. As the training progresses, the features encoded by $\mathbf{M}_{t,u}^i$ and $\mathbf{M}_{t,n}^i$ are gradually systematized into task-relevant motion information and irrelevant motion clues. Under this contrastive setting of $\mathbf{M}_{t,n}^i$, the useful motion feature $\mathbf{M}_{t,u}^i$ would be more discriminative and beneficial for the pose estimation task, as shown in Fig. 1 (c).

Furthermore, two regularization functions are introduced to facilitate model optimization. **(i)** We propose to improve the capacity of the multi-stage temporal difference encoder (TDE) to perceive meaningful information:

$$
\max \mathcal{I}\left(\mathbf{M}_t^i; \mathbf{M}_{t,u}^i\right),
\tag{9}
$$

where $\mathcal{I}\left(\mathbf{M}_t^i; \mathbf{M}_{t,u}^i\right)$ measures the useful motion information compressed in the vanilla motion feature $\mathbf{M}_t^i$. Maximizing this term encourages more abundant and effective motion cues to be encoded in our TDE. **(ii)** We propose to mitigate the information dropping during the feature enhancement of $\tilde{\mathbf{F}}_t^i$:

$$
\min\left[\mathcal{I}\left(\mathbf{M}_{t,u}^i; y_t^i \mid \tilde{\mathbf{F}}_t^i\right) + \mathcal{I}\left(\mathbf{F}_t^{i,j}; y_t^i \mid \tilde{\mathbf{F}}_t^i\right)\right],
\tag{10}
$$

where y_t^i refers to the label, and terms $\mathcal{I}\left(\mathbf{M}_{t,u}^i; y_t^i \mid \tilde{\mathbf{F}}_t^i\right)$ and $\mathcal{I}\left(\mathbf{F}_t^{i,j}; y_t^i \mid \tilde{\mathbf{F}}_t^i\right)$ quantify the vanishing task-relevant information in $\mathbf{M}_{t,u}^i$ and $\mathbf{F}_t^{i,j}$ respectively when aggregating them into the enhanced feature $\tilde{\mathbf{F}}_t^i$. Minimizing these two regularization terms fosters the nondestructive propagation of information. Given the notorious difficulty of conditional MI calculations [19, 48], we adopt a simplified version as done in [35, 70]. We approximate the first term $\mathcal{I}\left(\mathbf{M}_{t,u}^i; y_t^i \mid \tilde{\mathbf{F}}_t^i\right)$ as:

$$
\mathcal{I}\left(\mathbf{M}_{t,u}^i; y_t^i \mid \tilde{\mathbf{F}}_t^i\right) \to \mathcal{I}\left(\mathbf{M}_{t,u}^i; y_t^i\right) - \mathcal{I}\left(\mathbf{M}_{t,u}^i; \tilde{\mathbf{F}}_t^i\right).
\tag{11}
$$

Similarly, the second term $\mathcal{I}\left(\mathbf{F}_t^{i,j}; y_t^i \mid \tilde{\mathbf{F}}_t^i\right)$ in Eq. 10 can be simplified as follows:

$$
\mathcal{I}\left(\mathbf{F}_t^{i,j}; y_t^i \mid \tilde{\mathbf{F}}_t^i\right) \to \mathcal{I}\left(\mathbf{F}_t^{i,j}; y_t^i\right) - \mathcal{I}\left(\mathbf{F}_t^{i,j}; \tilde{\mathbf{F}}_t^i\right).
\tag{12}
$$

Ultimately, the proposed primary objective (Eq. 8) and the regularization terms (Eq. 9 and Eq. 10) are simultaneously optimized to provide feature supervision:

$$
\begin{aligned}
\mathcal{L}_{\mathrm{MI}} = {} & \mathcal{I}\left(\mathbf{M}_{t,u}^i; \mathbf{M}_{t,n}^i\right) - \mathcal{I}\left(\mathbf{M}_t^i; \mathbf{M}_{t,u}^i\right) \\
& + \mathcal{I}\left(\mathbf{M}_{t,u}^i; y_t^i \mid \tilde{\mathbf{F}}_t^i\right) + \mathcal{I}\left(\mathbf{F}_t^{i,j}; y_t^i \mid \tilde{\mathbf{F}}_t^i\right).
\end{aligned}
\tag{13}
$$

We adopt Variational Self-Distillation (VSD) [48] to compute the MI for each term.

3.3. Loss Functions

Overall, our loss function consists of two portions. (1) We employ the standard pose heatmap loss $\mathcal{L}_{\mathrm{H}}$ to supervise the final pose estimation:

$$
\mathcal{L}_{\mathrm{H}} = \left\|\mathbf{H}_t^i - \mathbf{G}_t^i\right\|_2^2,
\tag{14}
$$

where $\mathbf{H}_t^i$ and $\mathbf{G}_t^i$ denote the predicted and ground truth pose heatmaps, respectively. (2) We also leverage the proposed mutual information objective $\mathcal{L}_{\mathrm{MI}}$ to supervise the learning of motion features. Our total loss is given by:

$$
\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{H}} + \alpha \mathcal{L}_{\mathrm{MI}},
\tag{15}
$$

where α is a hyper-parameter to balance the ratio of different loss terms.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
PoseTracker [16]	67.5	70.2	62.0	51.7	60.7	58.7	49.8	60.6
PoseFlow [60]	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
JointFlow [12]	-	-	-	-	-	-	-	69.3
FastPose [68]	80.0	80.3	69.5	59.1	71.4	67.5	59.4	70.3
TML++ [22]	-	-	-	-	-	-	-	71.5
Simple (R-50) [58]	79.1	80.5	75.5	66.0	70.8	70.0	61.7	72.4
Simple (R-152) [58]	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
STEmbedding [27]	83.8	81.6	77.1	70.0	77.4	74.5	70.8	77.0
HRNet [46]	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
MDPN [17]	85.2	88.5	83.9	77.5	79.0	77.0	71.4	80.7
CorrTrack [42]	86.1	87.0	83.4	76.4	77.3	79.2	73.3	80.8
Dynamic-GNN [62]	88.4	88.4	82.0	74.5	79.1	78.3	73.1	81.1
PoseWarper [5]	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
DCPose [34]	88.0	88.7	84.1	78.4	83.0	81.4	74.2	82.8
DetTrack [54]	89.4	89.7	85.5	79.5	82.4	80.8	76.4	83.8
FAMI-Pose [35]	89.6	90.1	86.3	80.0	84.6	83.4	77.0	84.8
TDMI (Ours)	90.0	**91.1**	87.1	81.4	85.2	84.5	78.5	85.7
TDMI-ST (Ours)	**90.6**	91.0	**87.2**	**81.5**	**85.2**	**84.5**	**78.7**	**85.9**

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

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
STAF [41]	-	-	-	64.7	-	-	62.0	70.4
AlphaPose [13]	63.9	78.7	77.4	71.0	73.7	73.0	69.7	71.9
TML++ [22]	-	-	-	-	-	-	-	74.6
MDPN [17]	75.4	81.2	79.0	74.1	72.4	73.0	69.9	75.0
PGPT [3]	-	-	-	72.3	-	-	72.2	76.8
Dynamic-GNN [62]	80.6	84.5	80.6	74.4	75.0	76.7	71.8	77.9
PoseWarper [5]	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
PT-CPN++ [63]	82.4	88.8	86.2	79.4	72.0	80.6	76.2	80.9
DCPose [34]	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
DetTrack [54]	84.9	87.4	84.8	79.2	77.6	79.7	75.3	81.5
FAMI-Pose [35]	85.5	87.7	84.2	79.2	81.4	81.1	74.9	82.2
TDMI (Ours)	86.2	88.7	85.4	80.6	82.4	82.1	77.5	83.5
TDMI-ST (Ours)	**86.7**	**88.9**	**85.4**	**80.6**	**82.4**	**82.1**	**77.6**	**83.6**

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

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Tracktor++ w. poses [4, 11]	-	-	-	-	-	-	-	71.4
CorrTrack [11,42]	-	-	-	-	-	-	-	72.3
CorrTrack w. ReID [11,42]	-	-	-	-	-	-	-	72.7
Tracktor++ w. corr. [4, 11]	-	-	-	-	-	-	-	73.6
DCPose [34]	83.2	84.7	82.3	78.1	80.3	79.2	73.5	80.5
FAMI-Pose [35]	83.3	85.4	82.9	78.6	81.3	80.5	75.3	81.2
TDMI (Ours)	85.8	**87.5**	85.1	81.2	83.5	82.4	77.9	83.5
TDMI-ST (Ours)	**86.8**	87.4	85.1	81.4	83.8	82.7	78.0	83.8

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

4. Experiments

4.1. Experimental Settings

Datasets. PoseTrack is a large-scale benchmark dataset for video-based human pose estimation. Specifically, **PoseTrack2017** contains 250 video sequences for training and 50 video sequences for validation (following the official protocol), with a total of $80,144$ pose annotations. **PoseTrack2018** greatly increases the number of videos and includes 593 for training and 170 for validation (with $153,615$ pose annotations). Both datasets are annotated with 15 keypoints, with an additional flag for joint visibility. **PoseTrack21** further extends pose annotations of the PoseTrack2018 dataset, especially for particular small persons and persons in crowds, including $177,164$ pose annotations. The flag of joint visibility is re-defined in

Method	w_AP@avg	w_AP@50	w_AP@75	w_AP@90	AP@avg	AP@50	AP@75	AP@90
RSN18 [6]	48.3	52.5	48.9	43.6	52.4	56.2	53.2	47.8
DHRN FT [46]	51.6	59.6	50.1	45.3	55.3	63.0	53.9	49.1
ADAM+PRM [33]	54.0	63.0	52.2	46.7	57.7	66.5	55.9	50.6
DH_IBA [64]	55.2	66.7	52.5	46.3	59.3	70.5	56.8	50.5
TryNet [34]	55.2	68.7	51.7	45.3	59.3	72.6	56.1	49.3
ccc† [8]	56.3	**69.6**	53.0	46.4	60.0	**73.1**	56.9	49.9
TDMI-ST (Ours)	**58.5**	69.0	**56.2**	**50.4**	**62.1**	72.4	**60.0**	**54.0**

Table 4. Quantitative results from the **HiEve** *test leaderboard*. † means using extra data. AP@β denotes the AP value calculated under given distance threshold β, and the prefix "w" indicates that the value is calculated with frame weight. The official leaderboard ranks different approaches leveraging the metric of **w_AP@avg**.

PoseTrack21 so that occlusion information can be utilized. **HiEve** [33] is a very challenging benchmark dataset for human-centric video analysis in various realistic crowded and complex events (*e.g.*, earthquake, getting-off train, and bike collision), containing 32 video clips where 19 for training and 13 for testing. This dataset possesses a substantially larger data scale and includes the currently *largest* number of pose annotations ($1,099,357$).

Evaluation metric. We employ the standard pose estimation metric namely average precision (**AP**) to evaluate our model. We first compute the AP for each joint and then obtain the final performance (**mAP**) by averaging all joints.

Implementation details. Our TDMI framework is implemented with PyTorch. For visual feature extraction, we leverage the HRNet-W48 [46] model that is pre-trained on the COCO dataset. We incorporate several data augmentation strategies as adopted in [5, 46, 58], including random rotation $[-45°, 45°]$, random scale $[0.65, 1.35]$, truncation (half body), and flipping. The temporal span δ is set to 2, *i.e.*, 2 previous and 2 future frames. We use the Adam optimizer with an initial learning rate of $1e-3$ (decays to $1e-4$, $1e-5$, and $1e-6$ at the 6th, 12th, and 18th epochs, respectively). We train the model using 2 TITAN RTX GPUs. The training process is terminated within 20 epochs. To weight difference losses in Eq. 15, we empirically found that $\alpha = 1$ is the most effective. During the feature enhancement of $\tilde{\mathbf{F}}_t^i$, the final stage feature $\mathbf{F}_t^{i,4}$ is adopted as the visual feature.

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

Results on the PoseTrack2017 dataset. We first evaluate our method on the PoseTrack2017 dataset. A total of 18 methods are compared and their performances on the validation set are reported in Table 1. Our TDMI model consistently outperforms existing state-of-the-art methods, reaching an mAP of 85.7. To fully leverage temporal clues, we extend the TDMI to **TDMI-ST** where we replace the keyframe visual feature with the spatiotemporal feature of the input sequence to yield the final representation $\tilde{\mathbf{F}}_t^i$. The proposed TDMI-ST further pushes forward the performance boundary and achieves an mAP of 85.9. Remarkably, our TDMI-ST improves mAP by 8.6 points over the adopted backbone network HRNet-W48 [46]. Compared to

Figure 3. Visual results of our TDMI on benchmark datasets. Challenging scenes such as fast motion or pose occlusion are involved.

the previous best-performed method FAMI-Pose [35], our TDMI-ST also delivers a 1.1 mAP gain. The performance boost for challenging joints (*i.e.*, wrist, ankle) is also encouraging: we obtain an mAP of 81.5 (↑ 1.5) for wrists and an mAP of 78.7 (↑ 1.7) for ankles. Such consistent and significant performance improvements suggest the importance of explicitly embracing meaningful motion information. Moreover, we display the visualized results for scenarios with complex spatio-temporal interactions (*e.g.*, occlusion, blur) in Fig. 3, which attest to the robustness of the proposed method.

Results on the PoseTrack2018 dataset. We further benchmark our model on the PoseTrack2018 dataset. Empirical comparisons on the validation set are tabulated in Table 2. As presented in this table, our basic model, TDMI, surpasses all other approaches and achieves an mAP of 83.5. Our TDMI-ST model, on the other hand, attains the new state-of-the-art results over all joints and obtains the final performance of 83.6 mAP, with an mAP of 85.4, 80.6, and 77.6 for the elbow, wrist, and ankle, respectively.

Results on the PoseTrack21 dataset. Table 3 reports the results of our method as well as other state-of-the-art methods on the PoseTrack21 dataset. Quantitive results of the first four baselines [4, 11, 42] are officially provided by the dataset [11]. We further reproduce several previous impressive approaches (*i.e.*, DCPose [34] and FAMI-Pose [35]) according to their released implementations on GitHub, and evaluate their performances in this dataset. From Table 3, we observe that FAMI-Pose [35] achieves favorable pose estimation results with an mAP of 81.2. In contrast, our TDMI and TDMI-ST are able to obtain 83.5 mAP and 83.8 mAP, respectively. Another observation is that PoseTrack21 mainly increases the pose annotations of small persons and persons in crowds for PoseTrack2018. Interestingly, the proposed TDMI-ST attains a better performance on PoseTrack21 with respect to PoseTrack2018 (↑ 0.2 mAP), which might be an evidence to show the effectiveness and robustness of our method especially for challenging scenes.

Results on the HiEve dataset. Furthermore, we evaluate our model on the *largest* HiEve benchmark dataset. The detailed results of the test set are tabulated in Table 4. We upload the prediction results of our model to the HiEve test server[1] to obtain results. Our TDMI-ST achieves top scoring weight-average AP (w_AP@avg) of 58.5 on the HiEve leaderboard. Significantly, compared to ccc† [8] that uses extra data to train the model, we still achieve performance improvements by 2.2 w_AP and 2.1 AP, respectively.

Comparison of visual results. In addition to the quantitative analysis, we also qualitatively examine the capability of our approach in handling complex situations such as occlusion and blur. We depict in Fig. 4 the side-by-side comparisons of a) our TDMI against state-of-the-art methods b) HRNet [46] and c) FAMI-Pose [35]. It is observed that our approach consistently provides more accurate and robust pose detection for various challenging scenarios. HRNet is designed for static images and does not incorporate temporal dynamics, resulting in suboptimal results in degraded frames. In contrast, FAMI-Pose performs implicit motion compensation yet lacks effective information distillation. Through the principled design of TDE and RDM for temporal difference modeling and useful information disentanglement, our TDMI is more adept at handling challenging scenes such as occlusion and blur.

4.3. Ablation Study

We perform ablation studies focused on examining the contribution of each component in our TDMI framework, including the multi-stage Temporal Difference Encoder (TDE) and the Representation Disentanglement module (RDM). We also investigate the effectiveness of various micro designs within each component. All experiments are conducted on the PoseTrack2017 validation set.

Study on components of TDMI. We empirically evaluate the efficacy of each component in the proposed TDMI and report quantitative results in Table 5. Op-Flow is a baseline where we employ optical flows predicted by the RAFT [47] as motion representations. **(a)** For the first set-

---

[1] http://humaninevents.org/oltp.html?title=3

Method	TDE	RDM	Mean
HRNet [46]			77.3
Op-Flow			84.0
(a)	✓		84.5
(b)	✓	✓	85.7

Table 5. Ablation of different components in **TDMI**.

Method	Multi-stage	Spatial modulation	Progressive fusion	Mean
(a)				84.4
(b)	✓			84.6
(c)	✓		✓	84.9
(d)	✓	✓	✓	85.7

Table 6. Ablation of various designs in **TDE**.

Method	Factorization	MI objective	Mean
TDMI, w/o RDM			84.5
(a)	✓		84.7
(b)	✓	✓	85.7
(c)‡	✓	✓	83.8

Table 7. Ablation of various designs in **RDM**. "w/o X" refers to removing X module.
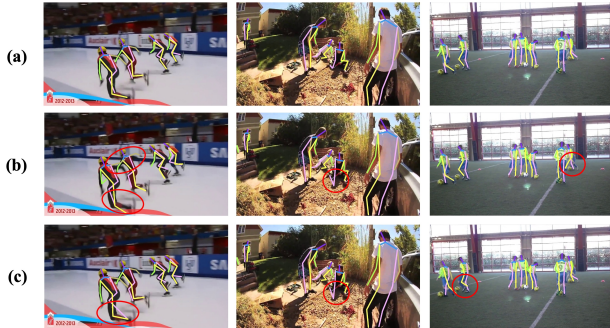
Figure 4. Qualitative comparisons of the prediction results of our TDMI (a), HRNet-W48 (b), and FAMI-Pose (c) on the challenging cases from PoseTrack dataset. Inaccurate detections are highlighted by the red circles.

ting, we introduce TDE to the HRNet-W48 [46] baseline for capturing motion contexts. Remarkably, the motion clues encoded by the TDE already improve over the baseline by a large margin of 7.2 mAP. This demonstrates the effectiveness of our TDE in introducing motion information to facilitate video-based human pose estimation. On the other hand, our TDE also delivers a performance gain of 0.5 mAP over the baseline Op-Flow. This highlights the great potential of temporal differences in representing motions as compared to complex optical flow. **(b)** For the next setting, we further incorporate the RDM to discover meaningful motions. The results in mAP increase to 85.7 by 1.2. This significant performance improvement on top of the well-established TDE corroborates the importance of excavating meaningful temporal dynamics in guiding accurate pose estimation.

Study on multi-stage Temporal Difference Encoder. We modify the TDE with various designs to investigate their influences on the final performance. As reported in Table 6, four experiments are conducted: **(a)** fusing only the features of the final stage $\mathbf{S}^4$, **(b)** directly aggregating multi-stage features via simple concatenation and convolution, **(c)** progressively integrating multi-stage features, and **(d)** our complete TDE with spatial modulation and progressive fusion. From this table, we observe that multi-stage feature fusion indeed outperforms using only single-stage features, yet the simple fusion scheme **(b)** yields a slight performance improvement (↑ 0.2 mAP). By progressively aggregating multi-stage features **(c)**, detailed information at each stage is preserved which provides an mAP gain of 0.5 points. Our complete TDE **(d)** further incorporates a spatial modulation mechanism to adaptively select important

information of each stage, achieving the best performance.

Study on Representation Disentanglement module. In addition, we examine the effects of RDM under different settings and present the results in Table 7. We first conduct a simple baseline **(a)** in which the MI objective is removed. This method partially distills the useful motion information and marginally improves the mAP by 0.2 ($84.5 \rightarrow 84.7$). We then incorporate the MI objective, which corresponds to our full RDM **(b)**. The noticeable performance enhancement of 1.0 mAP provides empirical evidence that our proposed MI objective is effective as additional supervision to facilitate the learning of discriminative task-relevant motion clues. We also attempt to perform the feature factorization by splitting channels **(c)**‡ and the mAP drops 1.9, which suggests that the channel splitting scheme might be inapplicable to our TDMI framework.

5. Conclusion and Future Works

In this paper, we investigate the video-based human pose estimation task from the perspective of effectively exploiting dynamic contexts through temporal difference learning and useful information disentanglement. We present a multi-stage Temporal Difference Encoder (TDE) to capture motion clues conditioned on explicit feature difference representation. Theoretically, we further build a Representation Disentanglement module (RDM) on top of mutual information to grasp task-relevant information. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art approaches on four benchmark datasets, including PoseTrack2017, PoseTrack2018, PoseTrack21, and HiEve. Future works include applications to other video-related tasks such as 3D human pose estimation and action recognition. The temporal difference features can also be integrated into existing pose-tracking pipelines to assess the similarity of human motions for data association.

6. Acknowledgements

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[2] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7035–7044, 2020. 2

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

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 6, 7

[5] Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, and Lorenzo Torresani. Learning temporal pose estimation from sparsely-labeled videos. In *Advances in Neural Information Processing Systems*, pages 3027–3038, 2019. 2, 3, 6

[6] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xinyu Zhou, Erjin Zhou, Xiangyu Zhang, and Jian Sun. Learning delicate local representations for multi-person pose estimation. *arXiv preprint arXiv:2003.04030*, 2020. 6

[7] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[8] Shuning Chang, Li Yuan, Xuecheng Nie, Ziyuan Huang, Yichen Zhou, Yupeng Chen, Jiashi Feng, and Shuicheng Yan. Towards accurate human pose estimation in videos of crowded scenes. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4630–4634, 2020. 6, 7

[9] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7103–7112, 2018. 3

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

[11] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20963–20972, 2022. 2, 6, 7

[12] Andreas Doering, Umar Iqbal, and Juergen Gall. Joint flow: Temporal flow fields for multi person tracking. *arXiv preprint arXiv:1805.04596*, 2018. 6

[13] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 2, 6

[14] Yixing Gao, Hyung Jin Chang, and Yiannis Demiris. User modelling for personalised dressing assistance by humanoid robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1840–1845. IEEE, 2015. 1

[15] Yixing Gao, Hyung Jin Chang, and Yiannis Demiris. Iterative path optimisation for personalised dressing assistance using vision and force information. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4398–4403. IEEE, 2016. 1

[16] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 350–359, 2018. 6

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

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 3

[19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018. 5

[20] Xuege Hou, Yali Li, and Shengjin Wang. Disentangled representation for age-invariant face recognition: A mutual information minimization perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3692–3701, 2021. 5

[21] Ziyuan Huang, Shiwei Zhang, Jianwen Jiang, Mingqian Tang, Rong Jin, and Marcelo H Ang. Self-supervised motion learning from static images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1276–1285, 2021. 2

[22] Jihye Hwang, Jieun Lee, Sungheon Park, and Nojun Kwak. Pose estimator and tracker using temporal flow maps for limbs. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 6

[23] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[24] Takashi Isobe, Xu Jia, Xin Tao, Changlin Li, Ruihuang Li, Yongjie Shi, Jing Mu, Huchuan Lu, and Yu-Wing Tai. Look back and forth: Video super-resolution with explicit temporal difference modeling. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 17411–17420, 2022. 1, 2, 3

[25] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2000–2009, 2019. 2, 3

[26] Chenru Jiang, Kaizhu Huang, Shufei Zhang, Xinheng Wang, and Jimin Xiao. Pay attention selectively and comprehensively: Pyramid gating network for human pose estimation without pre-training. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2364–2371, 2020. 3

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

[28] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 713–728, 2018. 3

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

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

[31] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2020. 3

[32] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11313–11322, 2021. 1, 2

[33] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Rui Qian, Tao Wang, Ning Xu, Hongkai Xiong, Guo-Jun Qi, and Nicu Sebe. Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint arXiv:2005.04490*, 2020. 2, 6

[34] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 525–534, 2021. 1, 2, 3, 6, 7

[35] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang Wang. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11016, 2022. 1, 2, 5, 6, 7

[36] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong

Lu. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11669–11676, 2020. 3

[37] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. Lstm pose machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5207–5215, 2018. 2

[38] Joe Yue-Hei Ng and Larry S Davis. Temporal difference networks for video action recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1587–1596. IEEE, 2018. 3

[39] Yanwei Pang, Yazhao Li, Jianbing Shen, and Ling Shao. Towards bridging semantic gap to improve semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4230–4239, 2019. 3

[40] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015. 2

[41] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4620–4628, 2019. 6

[42] Umer Rafi, Andreas Doering, Bastian Leibe, and Juergen Gall. Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In *European Conference on Computer Vision*, pages 36–52. Springer, 2020. 6, 7

[43] Benjamin Sapp, Alexander Toshev, and Ben Taskar. Cascaded models for articulated pose estimation. In *European conference on computer vision*, pages 406–420. Springer, 2010. 1

[44] Luca Schmidtke, Athanasios Vlontzos, Simon Ellershaw, Anna Lukens, Tomoki Arichi, and Bernhard Kainz. Unsupervised human pose estimation through transforming shape templates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2484–2494, 2021. 1

[45] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4220–4229, 2017. 2

[46] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019. 2, 3, 6, 7, 8

[47] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 7

[48] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 1522–1531, 2021. 5

[49] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1653–1660, 2014. 1

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[51] Fang Wang and Yi Li. Beyond physical connections: Tree models in human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 596–603, 2013. 1

[52] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1895–1904, 2021. 2, 3

[53] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 3

[54] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11088–11096, 2020. 2, 6

[55] Yang Wang and Greg Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *European Conference on Computer Vision*, pages 710–724. Springer, 2008. 1

[56] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. 1, 2

[57] Jiamin Wu, Tianzhu Zhang, Zhe Zhang, Feng Wu, and Yongdong Zhang. Motion-modulated temporal fragment alignment network for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9160, 2022. 1

[58] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1, 2, 6

[59] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3252–3262, 2022. 2

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

[61] 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. 1, 2

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

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

[64] Lei Yuan, Shu Zhang, Feng Fubiao, Naike Wei, and Huadong Pan. Combined distillation pose. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4635–4639, 2020. 6

[65] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in Neural Information Processing Systems*, 34:7281–7293, 2021. 1, 2

[66] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 2, 3

[67] Dingwen Zhang, Guangyu Guo, Dong Huang, and Junwei Han. Poseflow: A deep motion representation for understanding human behaviors in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6762–6770, 2018. 2

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

[69] Xiaoqin Zhang, Changcheng Li, Xiaofeng Tong, Weiming Hu, Steve Maybank, and Yimin Zhang. Efficient human pose estimation via parsing a tree structure based human model. In *IEEE 12th International Conference on Computer Vision*, pages 1349–1356. IEEE, 2009. 1

[70] Long Zhao, Yuxiao Wang, Jiaping Zhao, Liangzhe Yuan, Jennifer J Sun, Florian Schroff, Hartwig Adam, Xi Peng, Dimitris Metaxas, and Ting Liu. Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12793–12802, 2021. 5

[71] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Recognize actions by disentangling components of dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6566–6575, 2018. 1, 3

[72] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9308–9316, 2019. 4