Temporal Feature Alignment and Mutual Information Maximization for Video-Based Human Pose Estimation

Zhenguang Liu[1], Runyang Feng[2]*, Haoming Chen[2]*, Shuang Wu[3]*, Yixing Gao[4], Yunjun Gao[1], Xiang Wang[5]
[1] Zhejiang University, [2] Zhejiang Gongshang University, [3] Black Sesame Technologies,
[4] Jilin University, [5] National University of Singapore
{liuzhenguang2008, runyang2019.feng, chenhaomingbob}@gmail.com, wushuang@outlook.sg,
gaoyixing@jlu.edu.cn, gaoyj@zju.edu.cn, xiangwang1223@gmail.com

Abstract

Multi-frame human pose estimation has long been a compelling and fundamental problem in computer vision. This task is challenging due to fast motion and pose occlusion that frequently occur in videos. State-of-the-art methods strive to incorporate additional visual evidences from neighboring frames (supporting frames) to facilitate the pose estimation of the current frame (key frame). One aspect that has been obviated so far, is the fact that current methods directly aggregate unaligned contexts across frames. The spatial-misalignment between pose features of the current frame and neighboring frames might lead to unsatisfactory results. More importantly, existing approaches build upon the straightforward pose estimation loss, which unfortunately cannot constrain the network to fully leverage useful information from neighboring frames.

To tackle these problems, we present a novel hierarchical alignment framework, which leverages coarse-to-fine deformations to progressively update a neighboring frame to align with the current frame at the feature level. We further propose to explicitly supervise the knowledge extraction from neighboring frames, guaranteeing that useful complementary cues are extracted. To achieve this goal, we theoretically analyzed the mutual information between the frames and arrived at a loss that maximizes the task-relevant mutual information. These allow us to rank No.1 in the Multi-frame Person Pose Estimation Challenge on benchmark dataset PoseTrack2017, and obtain state-of-the-art performance on benchmarks Sub-JHMDB and PoseTrack2018. Our code is released at https://github.com/Pose-Group/FAMI-Pose, hoping that it will be useful to the community.

*Corresponding Authors

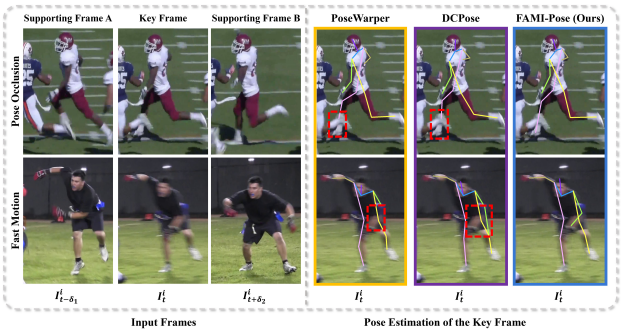

Figure 1. State-of-the-art methods like PoseWarper and DCPose directly aggregate unaligned contexts from neighboring frames, which may fail for scenes with fast motion or pose occlusion. We perform temporal feature alignment between each supporting frame and the key frame, delivering robust pose estimations.

1. Introduction

A key component of our capacity to interact with others lies in our ability to recognize the poses of humans [36, 37, 48]. Likewise, detecting human poses is crucial for an intelligent machine to adjust its action and properly allocate its attention when interacting with people. Nowadays, pose estimation finds abundant applications in a wide spectrum of scenarios including action recognition, augmented reality, surveillance, and tracking [39, 67].

An extensive body of literature focuses on pose estimation in *static images*, ranging from earlier approaches [47, 57, 59, 70] utilising tree models or random forest models to recent attempts employing deep convolutional neural networks [6, 42, 54, 60]. For pose estimation in videos, such methods are severely challenged in handling deteriorated video frames arising from scenes with fast motion and pose occlusion. Incorporating and leveraging additional contexts from neighboring frames is desirable to fill in the absent motion dynamics within a single frame and facilitate pose

estimation.

One line of work [2, 39, 58] proposes to aggregate *vanilla* sequential features of neighboring frames (supporting frames). [39] trains a convolutional LSTM to model both spatial and temporal features, and directly predicts pose sequences for videos. [58] presents a 3D-HRNet to assemble features over a tracklet. Another line of work [35, 45, 50] employs optical flow or implicit motion estimation to polish the pose estimation of the current frame (key frame). [45, 50] propose to compute dense optical flow between frames, and leverage the flow based motion field for refining pose heatmaps temporally across multiple frames. [35] aggregates the pose heatmaps of consecutive frames and models motion residuals to improve pose estimation of the key frame.

Upon scrutinizing and experimenting on the released implementations of existing methods [5, 11, 35], we observe that they suffer from performance deterioration in challenging cases such as rapid motion and pose occlusion. As illustrated in Fig. 1, in the pose occlusion scenario, existing methods like DCPose fail to recognize the right ankle of the occluded person, leading to unexpected results. In the fast motion scenario, existing methods encounter difficulties in identifying the left wrist due to motion blur. We conjecture that the reasons are **twofolds**. **(1)** It is common that the same person in the current frame and a neighboring frame is not well aligned, especially for situations involving rapid motion of human subjects or cameras. However, existing methods tend to directly aggregate unaligned contexts from neighboring frames, these spatially misaligned features potentially diminish the performances of models. **(2)** State-of-the-art approaches simply employ the conventional MSE (Mean Square Error of joints) loss to supervise the learning of pose heatmaps, while lacking an effective constraint on guaranteeing information gain from neighboring frames as well as a supervision at the intermediate feature level.

In this paper, we present a novel framework, along with theoretical analysis, to tackle the above challenges. The proposed method, termed FAMI-Pose (**F**eature **A**lignment and **M**utual **I**nformation maximization for **P**ose estimation), consists of two key components. **(i)** FAMI-Pose conducts coarse-to-fine deformations that systematically update a neighboring frame to align with the current frame at the feature level. Specifically, FAMI-Pose first performs a *global transformation*, which holistically rearranges neighboring frame feature to preliminarily rectify spatial shifts or jitter. Subsequently, a *local calibration* is exploited to adaptively move and modulate each pixel of neighboring frame feature for enhanced feature alignment. **(ii)** FAMI-Pose further engages an information-theoretic objective as an additional intermediate supervision at the feature level. Maximizing this mutual information objective allows our model to fully mine task-relevant cues within the neighboring frames, extracting purposeful complementary knowledge to enhance pose estimation on the key frame. To the best of our knowledge, we are the first to methodically investigate the problem of feature alignment in human pose estimation and provide insights from an information-theoretic perspective.

We extensively evaluate the proposed method on three widely used benchmark datasets, PoseTack2017, PoseTrack2018, and Sub-JHMDB. Empirical evaluations show that our approach significantly outperforms current state-of-the-art methods. Our method achieves **84.8** mAP, **82.2** mAP, and **96.0** mAP on PoseTrack2017, PoseTrack2018, and Sub-JHMDB, respectively. Our results are submitted to the official evaluation server of PoseTack2017, and rank *No.1* for this large benchmark dataset. We also present extensive ablation analyses on the contribution of each component, and validate the efficacy of feature alignment and the proposed mutual information loss.

The contributions of this work are summarized as:

- We propose to examine the multi-frame human pose estimation task from the perspective of effectively leveraging temporal contexts through feature alignment.
- To explicitly supervise the knowledge extraction from neighboring frames, we propose an information-theoretic loss function, which allows maximizing the task-relevant cues mined from supporting frames.
- Our approach sets new state-of-the-art results on three benchmark datasets, PoseTrack2017, PoseTrack2018, and Sub-JHMDB. Our source code has been released.

2. Related Work

In this section, we briefly review the following three topics that are closely related to our work, namely image-based human pose estimation, video-based human pose estimation, and feature alignment.

2.1. Image-Based Human Pose Estimation

Conventional solutions to image-based human pose estimation utilize pictorial structures [47, 70] to model the spatial relationships among body joints. These methods tend to rely on handcrafted features and have limited representational abilities. Fueled by the explosion of deep learning [19, 58] and the availability of large-scale pose estimation datasets such as PoseTrack [1, 27] and COCO [34], various deep learning methods [2, 8, 17, 18, 22, 51, 56, 65, 66, 68] have been proposed. These methods can be broadly categorized into two paradigms: bottom-up and top-down. *Bottom-up approaches* [6, 30–32] first detect individual body parts, and then assemble these detected constituent parts into the entire person. [6] proposes a dual convolution structure to si-

multaneously predict part confidence maps and part affinity fields (that represent the relationships between body parts). On the other hand, *top-down approaches* [41, 42, 52, 60, 62] first detect human bounding boxes and then estimate human poses within each bounding box. [62] leverages deconvolution layers to replace the commonly used bi-linear interpolation for spatial-upsampling of feature maps. A recent work in [52] presents a high resolution network (HRNet) that retains high resolution feature maps throughout the entire inference, achieving state-of-the-art results on multiple image-based benchmarks.

2.2. Video-Based Human Pose Estimation

Pose estimation models trained for image-based data could not generalize well to video sequences due to their inability to incorporate abundant cues from neighboring frames. To model and leverage temporal contexts across frames, one direct approach would be employing convolutional LSTMs as proposed in [2, 39]. A key shortcoming of such models might be their tendency to misalign features across different frames, which unfavourably reduces the potency of the supporting frames. [45, 50] explicitly estimate motion fields by computing optical flow between consecutive frames, and these motion cues are subsequently used for aligning pose heatmaps. [35] estimates motion offsets between the key frame and supporting frames, and these offsets provide the basis to perform resampling of pose heatmaps on consecutive frames. In both cases, the pose estimation accuracy would be heavily dependent on the performance of the optical flow or motion offset estimation. Furthermore, the lack of an effective supervision at the intermediate features level for these approaches could lead to inaccurate pose estimations.

2.3. Feature Alignment

Feature alignment is an important topic for many computer vision tasks (*e.g.*, semantic segmentation [33, 40], object detection [7, 20]), and numerous efforts have recently been made to address this problem. [38] presents an index-guided framework that employs indices to guide the pooling and upsampling. [23] proposes to learn the transformation offsets of pixels to align upsampled feature maps. [24] presents an aligned feature aggregation module to align the features of multiple different resolutions for better aggregation. Whereas previous methods mostly tackle spatial misalignment between network inputs and outputs, we focus on temporal (*i.e.*, across frames) feature alignment in this work.

3. Our Approach

Preliminaries To detect human poses from the video frames, we first extract the bounding box of each individual person. Technically, for a video frame I_t, we first employ an object detector to extract the bounding box for each individual person. This bounding box is then enlarged by 25% to crop the same individual on a predefined window $\mathcal{N}$ of neighboring frames. Overall, for person i, we obtain the cropped image I_t^i for the key frame and $\{I_{t+\delta}^i \mid \delta \in \mathcal{N}\}$ for the supporting (neighboring) frames.

Problem Formulation Presented with a key frame I_t^i along with its supporting frames $\{I_{t+\delta}^i \mid \delta \in \mathcal{N}\}$, our goal is to estimate the pose in I_t^i. We seek to better leverage the supporting frames through a principled feature alignment and mining task relevant information, thereby addressing the common drawback of existing approaches in failing to adequately tap into the temporal information.

Method Overview An overview of our pipeline is illustrated in Fig. 2. For each supporting frame $I_{t+\delta}^i$, FAMI-pose performs a two-stage hierarchical transformation to align $I_{t+\delta}^i$ with the key frame I_t^i at the feature level. Specifically, FAMI-Pose consists of two main modules, a global transformation module and a local calibration module. We first perform feature extraction on I_t^i and $I_{t+\delta}^i$ to obtain z_t^i and $z_{t+\delta}^i$, respectively. These features are then fed into our global transformation module, which learns the parameters of an affine transformation to obtain a coarsely aligned supporting frame feature $\bar{z}_{t+\delta}^i$. z_t^i and $\bar{z}_{t+\delta}^i$ are then handed to the local calibration module, which performs pixel-wise deformation to produce finely aligned features $\bar{\bar{z}}_{t+\delta}^i$. Finally, we aggregate all aligned supporting frame features $\{\bar{\bar{z}}_{t+\delta}^i \mid \delta \in \mathcal{N}\}$ and the key frame feature z_t^i to obtain our enhanced feature $\tilde{z}_t^i$. $\tilde{z}_t^i$ is passed to a detection head that outputs pose estimations $\widehat{H}_t^i$. The task objective is to minimize the heatmap estimation loss $\mathcal{L}_H$ which measures the discrepancy between $\widehat{H}_t^i$ and the ground truth H_t^i. On top of this, we also design a mutual information objective $\mathcal{L}_{MI}$ which effectuates a feature level supervision for maximizing the amount of complementary task-relevant information encoded in $\tilde{z}_t^i$. In what follows, we introduce the complete FAMI-Pose architecture and the mutual information objective in detail.

3.1. Feature Alignment

Feature alignment starts with feature extraction, which is done with the HRNet-W48 network [52] (the state-of-the-art method for image-based human pose estimation) as the backbone. The extracted features z_t^i and $z_{t+\delta}^i$ are then passed through a global transformation module and a local calibration module, to progressively align $z_{t+\delta}^i$ with z_t^i. We would like to highlight that we do not pursue an image-level alignment, instead we drive the network to learn a feature-level alignment between a supporting frame and the key frame.

Global Transformation We observe that most failure cases for pose estimation in videos occur due to rapid movements of persons or cameras, which inevitably lead to large

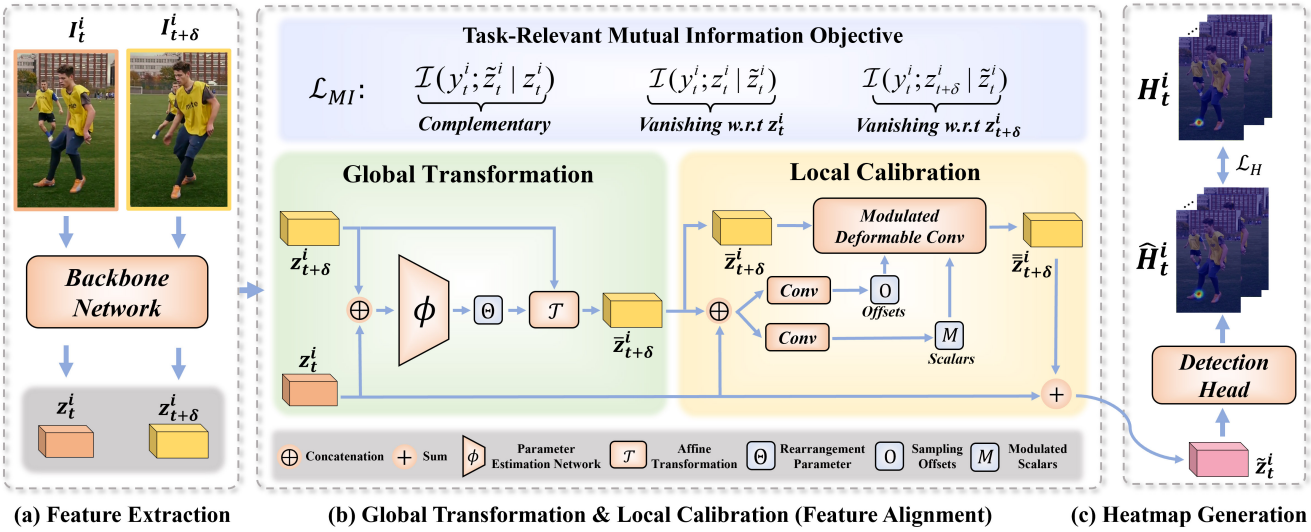

Figure 2. Overall pipeline of our FAMI-Pose framework. The goal is to detect the pose of person i in the key frame I_t^i, with the assistance of its supporting frames. For clarity of illustration, we only show a single supporting frame $I_{t+\delta}^i$ in this figure. We first extract their respective features z_t^i and $z_{t+\delta}^i$. These features are then handed to our global transformation module and the local calibration module for temporal alignment. The key frame feature z_t^i and aligned features $\bar{\bar{z}}_{t+\delta}^i$ for all supporting frames are aggregated to $\tilde{z}_t^i$, which is passed to a detection head that outputs pose estimates $\widehat{H}_t^i$. Besides the heatmap estimation loss $\mathcal{L}_H$ that measures the discrepancy between $\widehat{H}_t^i$ and the ground truth H_t^i, we introduce an additional feature level supervision through our Mutual Information objective $\mathcal{L}_{MI}$ to extract maximal task-relevant complementary information from supporting frames.

spatial shifts or jitters between neighboring frames. In order to align a supporting frame to the key frame, we design a global transformation module (GTM). The GTM computes spatial rearrangement parameters of a global affine transformation to obtain a coarse preliminary alignment of supporting frame feature $z_{t+\delta}^i$ with the key frame feature z_t^i.

More specifically, the GTM includes two submodules:

1. A spatial rearrangement parameter estimation network ϕ that estimates affine transformation parameters Θ from the input feature pair as $\phi : (z_t^i, z_{t+\delta}^i) \rightarrow \Theta \in \mathbb{R}^{2\times 3}$. The elements of Θ correspond to translation, rotation, shear, and scaling operations.

2. Subsequently, a global affine transformation $\mathcal{T}$ is performed to obtain the preliminarily aligned supporting frame feature $\mathcal{T} : (z_{t+\delta}^i, \Theta) \rightarrow \bar{z}_{t+\delta}^i$.

The operations of the GTM can be expressed as follows:

$$\Theta = \phi\left(z_t^i \oplus z_{t+\delta}^i\right),$$
$$\begin{pmatrix} x_p \\ y_p \end{pmatrix} = \underbrace{\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix}}_{\Theta} \begin{pmatrix} \bar{x}_p \\ \bar{y}_p \\ 1 \end{pmatrix}, \tag{1}$$

where (x_p, y_p) and $(\bar{x}_p, \bar{y}_p)$ denote the coordinates of pixel p for $z_{t+\delta}^i$ and $\bar{z}_{t+\delta}^i$, respectively.

Local Calibration The global transformation module produces a coarse alignment. We then design our local calibration module (LCM) to perform meticulous fine-tuning at a pixel-level, yielding finely aligned features $\bar{\bar{z}}_{t+\delta}^i$.

Specifically, given $\bar{z}_{t+\delta}^i$ and z_t^i, we independently estimate convolution kernel sampling offsets O and modulated scalars M for the feature $\bar{z}_{t+\delta}^i$:

$$\begin{aligned} \bar{z}_{t+\delta}^i \oplus z_t^i &\xrightarrow[\text{blocks}]{\text{residual}} \xrightarrow[\text{convolution}]{\text{regular}} O, \\ \bar{z}_{t+\delta}^i \oplus z_t^i &\xrightarrow[\text{blocks}]{\text{residual}} \xrightarrow[\text{convolution}]{\text{regular}} M. \end{aligned} \tag{2}$$

The adaptively learned kernel offsets O and modulated scalars M respectively correspond to *location shifts* and *intensity fluctuations* of each pixel in $\bar{z}_{t+\delta}^i$ with respect to the key frame feature z_t^i.

Subsequently, we implement the local calibration operation through the modulated deformable convolution [73]. Given the preliminarily aligned features $\bar{z}_{t+\delta}^i$, the kernel sampling offsets O, and the modulated scalars M as inputs, the modulated deformable convolution outputs the fine-tuned feature $\bar{\bar{z}}_{t+\delta}^i$:

$$\left(\bar{z}_{t+\delta}^i, O, M\right) \xrightarrow[\text{convolution}]{\text{modulated deformable}} \bar{\bar{z}}_{t+\delta}^i. \tag{3}$$

To anticipate the discussion of the mutual information loss, we would like to point out that the key frame fea-

ture z_t^i is only used for computing the global transformation parameters in GTM and convolutional parameters in LCM. Its information will not be propagated into the final aligned supporting frame feature $\bar{\bar{z}}_{t+\delta}^i$.

Heatmap Generation Ultimately, we aggregate over all final aligned supporting frame features $\{\bar{\bar{z}}_{t+\delta}^i \mid \delta \in \mathcal{N}\}$ and the key frame feature z_t^i via element-wise addition to obtain the enhanced feature $\tilde{z}_t^i$. $\tilde{z}_t^i$ is fed to a detection head to produce pose heatmap estimations $\widehat{H}_t^i$. We implemented the detection head using a stack of 3×3 convolutions. By effectively leveraging temporal information from supporting frames through our coarse-to-fine alignment modules, our FAMI-Pose is more adept at tackling visual degeneration issues and therefore gives more accurate pose heatmaps.

3.2. Mutual Information Objective

We can certainly train the FAMI-Pose in a direct end-to-end manner with a pose heatmap loss, as is done in most previous methods [5, 35, 52, 58, 62]. Given our systematic examination of extracting temporal features for pose estimation, it would be fruitful to investigate whether introducing *supervision at the feature level* would facilitate the task.

Naively, we could formulate the feature level objective as the $L1$ or $L2$ difference between supporting frames feature $z_{t+\delta}^i$ and the key frame feature z_t^i. However, such rigid-alignment is likely to lead to erosion of complementary task-specific information from supporting frames. Consequently, the temporal features thus optimized would be inadequate for providing relevant supporting information to facilitate pose estimation.

It is therefore crucial that we highlight the purposeful complementary information from the supporting frames. Towards this end, inspired by [21, 72], we propose a mutual information objective, which seeks to maximize the amount of complementary task-relevant information in the enhanced feature $\tilde{z}_t^i$.

Mutual Information Mutual information (MI) is a measure of the amount of information shared between random variables. Formally, MI quantifies the statistical dependency of two random variables $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$:

$$\mathcal{I}(\boldsymbol{v}_1;\boldsymbol{v}_2) = \mathbb{E}_{p(\boldsymbol{v}_1,\boldsymbol{v}_2)}\left[\log \frac{p(\boldsymbol{v}_1,\boldsymbol{v}_2)}{p(\boldsymbol{v}_1)p(\boldsymbol{v}_2)}\right], \tag{4}$$

where $p(\boldsymbol{v}_1,\boldsymbol{v}_2)$ is the joint probability distribution between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, while $p(\boldsymbol{v}_1)$ and $p(\boldsymbol{v}_2)$ are their marginals.

Mutual Information Loss Within this framework, our primary objective for learning effective temporal feature alignment can be formulated as:

$$\max \mathcal{I}\left(y_t^i;\tilde{z}_t^i \mid z_t^i\right), \tag{5}$$

where y_t^i represents the label, and $\mathcal{I}\left(y_t^i;\tilde{z}_t^i \mid z_t^i\right)$ denotes the amount of task-relevant information in the enhanced feature $\tilde{z}_t^i$, complementary to (*i.e.*, excluding) the information from the key frame feature z_t^i. Intuitively, optimizing this objective will maximize the additional relevant and complementary information we seek to extract from neighboring frames to support the pose estimation task.

Due to the notorious difficulty of the conditional MI computations especially in neural networks [21,53], we perform a simplification. We first factorize Eq. 5 as follows:

$$\begin{aligned}\mathcal{I}\left(y_t^i;\tilde{z}_t^i \mid z_t^i\right) = \mathcal{I}\left(y_t^i;\tilde{z}_t^i\right) - \mathcal{I}\left(\tilde{z}_t^i;z_t^i\right) \\ + \mathcal{I}\left(\tilde{z}_t^i;z_t^i \mid y_t^i\right),\end{aligned} \tag{6}$$

where $\mathcal{I}\left(y_t^i;\tilde{z}_t^i\right)$ measures the relevance of the label y_t^i and feature $\tilde{z}_t^i$, $\mathcal{I}\left(\tilde{z}_t^i;z_t^i\right)$ indicates the dependence between the two features $\tilde{z}_t^i$ and z_t^i, and $\mathcal{I}\left(\tilde{z}_t^i;z_t^i \mid y_t^i\right)$ represents the *task-irrelevant information* in both $\tilde{z}_t^i$ and z_t^i. Heuristically, when optimizing over the task objective, the task-specific information will have an overwhelming presence over the task-irrelevant information. Therefore, we may assume that the task-irrelevant information will be negligible upon sufficient training [14, 72]. This simplifies Eq. 6 to:

$$\mathcal{I}\left(y_t^i;\tilde{z}_t^i \mid z_t^i\right) \to \mathcal{I}\left(y_t^i;\tilde{z}_t^i\right) - \mathcal{I}\left(z_t^i;\tilde{z}_t^i\right). \tag{7}$$

Moreover, we introduce two regularization terms to alleviate information dropping:

$$\min\left[\mathcal{I}\left(y_t^i;z_{t+\delta}^i \mid \tilde{z}_t^i\right) + \mathcal{I}\left(y_t^i;z_t^i \mid \tilde{z}_t^i\right)\right]. \tag{8}$$

The terms $\mathcal{I}\left(y_t^i;z_{t+\delta}^i \mid \tilde{z}_t^i\right)$ and $\mathcal{I}\left(y_t^i;z_t^i \mid \tilde{z}_t^i\right)$ respectively measure the vanishing task-relevant information in $z_{t+\delta}^i$ and z_t^i during feature alignment. They serve to facilitate the nondestructive propagation of information. Simultaneously minimizing these two terms would prevent excessive information loss in $z_{t+\delta}^i$ and z_t^i while maximizing the primary complementary task-relevant mutual information objective.

Similar to Eq. 7, we simplify the two regularization terms in Eq. 8 as follows:

$$\begin{aligned}\mathcal{I}\left(y_t^i;z_{t+\delta}^i \mid \tilde{z}_t^i\right) \to \mathcal{I}\left(y_t^i;z_{t+\delta}^i\right) - \mathcal{I}\left(z_{t+\delta}^i;\tilde{z}_t^i\right), \\ \mathcal{I}\left(y_t^i;z_t^i \mid \tilde{z}_t\right) \to \mathcal{I}\left(y_t^i;z_t^i\right) - \mathcal{I}\left(z_t^i;\tilde{z}_t^i\right).\end{aligned} \tag{9}$$

Finally, we simultaneously optimize the complementary information term in Eq. 5 and the two regularization terms in Eq. 8 to provide feature level supervision:

$$\begin{aligned}\mathcal{L}_{\text{MI}} = \overbrace{\mathcal{I}\left(y_t^i;z_t^i \mid \tilde{z}_t^i\right)}^{\text{Vanishing w.r.t. } z_t^i} + \overbrace{\mathcal{I}\left(y_t^i;z_{t+\delta}^i \mid \tilde{z}_t^i\right)}^{\text{Vanishing w.r.t. } z_{t+\delta}^i} \\ - \alpha \cdot \underbrace{\mathcal{I}\left(y_t^i;\tilde{z}_t^i \mid z_t^i\right)}_{\text{Complementary}},\end{aligned} \tag{10}$$

where α serves as a hyper-parameter in our network to balance the ratios of different terms. These MI terms can be estimated by existing MI estimators [4, 9, 53, 55]. In our experiments, we employ the Variational Self-Distillation (VSD) [53] to estimate the MI for each term.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
PoseTracker [15]	67.5	70.2	62.0	51.7	60.7	58.7	49.8	60.6
PoseFlow [64]	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
JointFlow [10]	-	-	-	-	-	-	-	69.3
FastPose [69]	80.0	80.3	69.5	59.1	71.4	67.5	59.4	70.3
TML++ [25]	-	-	-	-	-	-	-	71.5
Simple (ResNet-50) [62]	79.1	80.5	75.5	66.0	70.8	70.0	61.7	72.4
Simple (ResNet-152) [62]	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
STEmbedding [29]	83.8	81.6	77.1	70.0	77.4	74.5	70.8	77.0
HRNet [52]	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
MDPN [16]	85.2	88.5	83.9	77.5	79.0	77.0	71.4	80.7
Dynamic-GNN [67]	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 [35]	88.0	88.7	84.1	78.4	83.0	81.4	74.2	82.8
FAMI-Pose (Ours)	**89.6**	**90.1**	**86.3**	**80.0**	**84.6**	**83.4**	**77.0**	**84.8**

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

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Total**
PoseTracker [15]	-	-	-	51.5	-	-	50.2	59.6
PoseFlow [64]	64.9	67.5	65.0	59.0	62.5	62.8	57.9	63.0
JointFlow [10]	-	-	-	53.1	-	-	50.4	63.4
TML++ [25]	-	-	-	60.9	-	-	56.0	67.8
KeyTrack [49]	-	-	-	71.9	-	-	65.0	74.0
DetTrack [58]	-	-	-	69.8	-	-	65.9	74.1
Simple (ResNet-152) [62]	80.1	80.2	76.9	71.5	72.5	72.4	65.7	74.6
HRNet [52]	80.1	80.2	76.9	72.0	73.4	72.5	67.0	74.9
PoseWarper [5]	79.5	84.3	80.1	75.8	77.6	76.8	70.8	77.9
DCPose [35]	84.3	84.9	80.5	76.1	77.9	77.1	71.2	79.2
FAMI-Pose (Ours)	**86.1**	**86.1**	**81.8**	**77.4**	**79.5**	**79.1**	**73.6**	**80.9**

Table 2. Performance comparisons on the PoseTrack2017 **test** set. These results are published in the *PoseTrack2017 leaderboard*.

3.3. Training Objective

Our training objective consists of two parts. (1) We adopt the heatmap estimation loss function $\mathcal{L}_{\text{H}}$ to supervise the learning of final pose estimates:

$$\mathcal{L}_{\text{H}} = \left\| \widehat{H}_t^i - H_t^i \right\|_2^2, \tag{11}$$

where $\widehat{H}_t^i$ and H_t^i denotes the prediction heatmap and ground truth heatmap, respectively. (2) We also leverage the proposed MI loss to supervise the temporal features as described in Sec. 3.2. The overall loss function is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{H}} + \beta \cdot \mathcal{L}_{\text{MI}}. \tag{12}$$

4. Experiments

In this section, we present our experimental results on three widely used benchmark datasets, namely PoseTrack2017 [27], PoseTrack2018 [1], and Sub-JHMDB [28].

4.1. Experimental Settings

Datasets PoseTrack is a large-scale benchmark for human pose estimation and articulated tracking in videos, containing challenging sequences of people in crowded scenarios and performing rapid movement. The **PoseTrack2017** dataset includes 514 video sequences with a total of 16,219 pose annotations. These are split (following the official protocol) into 250, 50, and 214 video sequences for training, validation, and testing. The **PoseTrack2018** dataset contains 1,138 video sequences (and 153,615 pose annotations), with 593 for training, 170 for validation, and 375 for testing. Both datasets are annotated with 15 joints, with an additional label for joint visibility. Training videos provide dense pose annotations in the center 30 frames, and validation videos further provide pose annotations every four frames. The **Sub-JHMDB** dataset contains 316 videos for a total of 11,200 frames. Annotations are done for 15 joints but only visible joints are annotated. Three different data splits are performed for this dataset, each with a training to testing ratio of 3 : 1. Following previous works [39, 43, 71], we report the mean accuracy over the three splits.

Implementation Details Our FAMI-Pose is implemented with PyTorch. The input image size is fixed to 384×288. We perform data augmentation including random rotation $[-45°, 45°]$, random scaling $[0.65, 1.35]$, random truncation, and horizontal flipping. The predefined window $\mathcal{N}$ of neighboring frames is set to $\{-2, -1, 1, 2\}$, *i.e.*, 2 previous and 2 future frames. We employ the HRNet-W48 model pre-trained on the COCO dataset for feature ex-

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
STAF [46]	-	-	-	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++ [25]	-	-	-	-	-	-	-	74.6
MDPN [16]	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 [67]	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
DCPose [35]	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
FAMI-Pose (Ours)	**85.5**	**87.7**	**84.2**	**79.2**	**81.4**	**81.1**	**74.9**	**82.2**

Table 3. Quantitative results on the PoseTrack2018 **validation** set.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Total**
TML++ [25]	-	-	-	60.2	-	-	56.9	67.8
AlphaPose++ [13, 16]	-	-	-	66.2	-	-	65.0	67.6
DetTrack [58]	-	-	-	69.8	-	-	67.1	73.5
MDPN [16]	-	-	-	74.5	-	-	69.0	76.4
PoseWarper [5]	78.9	84.4	80.9	76.8	75.6	77.5	71.8	78.0
DCPose [35]	82.8	84.0	80.8	77.2	76.1	77.6	72.3	79.0
FAMI-Pose (Ours)	**83.6**	**84.5**	**81.4**	**77.9**	**76.8**	**78.3**	**72.9**	**79.6**

Table 4. Performance comparisons on the PoseTrack2018 **test** set.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Avg**
Part Models [44]	79.0	60.3	28.7	16.0	74.8	59.2	49.3	52.5
Joint Action [63]	83.3	63.5	33.8	21.6	76.3	62.7	53.1	55.7
Pose-Action [26]	90.3	76.9	59.3	55.0	85.9	76.4	73.0	73.8
CPM [61]	98.4	94.7	85.5	81.7	97.9	94.9	90.3	91.9
Thin-slicing Net [50]	97.1	95.7	87.5	81.6	98.0	92.7	89.8	92.1
LSTM PM [39]	98.2	96.5	89.6	86.0	98.7	95.6	90.0	93.6
DKD(ResNet-50) [43]	98.3	96.6	90.4	87.1	99.1	96.0	92.9	94.0
K-FPN(ResNet-18) [71]	94.7	96.3	95.2	90.2	96.4	95.5	93.2	94.5
K-FPN(ResNet-50) [71]	95.1	96.4	**95.3**	91.3	96.3	95.6	92.6	94.7
MotionAdaptive [12]	98.2	97.4	91.7	85.2	**99.2**	**96.7**	92.2	94.7
FAMI-Pose (Ours)	**99.3**	**98.6**	94.5	**91.7**	**99.2**	91.8	**95.4**	**96.0**

Table 5. Performance comparisons on the **Sub-JHMDB** dataset.

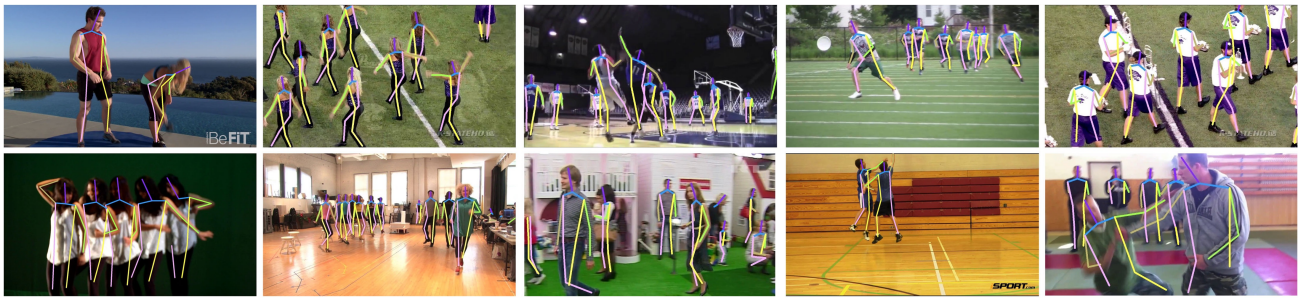

Figure 3. Visual results of our FAMI-Pose on benchmark datasets. Challenging scenes such as high-speed motion or pose occlusion are involved.

traction. Subsequent weight parameters are initialized from a standard Gaussian distribution, while biases are initialized to 0. We employ the Adam optimizer with a base learning rate of $1e-4$ (decays to $1e-5$, $1e-6$, and $1e-7$ at the 8^{th}, 12^{th}, and 16^{th} epochs, respectively). Training is done with 4 Nvidia Geforce RTX 2080 Ti GPUs and a batch size of 48. All training process is terminated within 20 epochs. To weigh different losses in Eq. 10 and Eq. 12, we set $\alpha = 1.0$ and $\beta = 0.1$, and have not densely tuned them.

Evaluation Metric We benchmark our model using the standard human pose estimation protocol [52, 62], namely the average precision (**AP**) metric. We compute the AP for each body joint, and then average over all joints to get the final results (**mAP**). Note that only visible joints are calculated in performance evaluation.

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

Results on the PoseTrack2017 Dataset We first evaluate our model on the PoseTrack2017 validation set and test set. A total of 14 methods are compared, including PoseTracker [15], PoseFlow [64], JointFlow [10], FastPose [69], TML++ [25], SimpleBaseline (ResNet-50 *and* ResNet-152), STEmbedding [29], HRNet [52], MDPN [16], Dynamic-GNN [67], PoseWarper [5], DCPose [35], and our FAMI-Pose. Their performance on the PoseTrack2017 validation set is reported in Table 1. The proposed FAMI-Pose consistently outperforms existing methods, achieving an mAP of 84.8. Significantly, our FAMI-Pose is able to improve the mAP by 7.5 points over the widely adopted backbone network HRNet-W48 [52]. Our model also achieves a 2.0 mAP gain over the previous state-of-the-art approach DCPose [35]. In particular, we obtain encouraging improvements for the more challenging joints (*i.e.*, wrist, ankle): with an mAP of 80.0 ($\uparrow$ 1.6) for wrists and an mAP of 77.0 ($\uparrow$ 2.8) for ankles. Another interesting observation is that pose estimation approaches that incorporate neighboring frames (such as PoseWarper and DCPose) outperforms methods that use only the single key frame. This suggests the importance of embracing complementary cues from neighboring frames.

Method	Global Transformation	Local Calibration	MI Loss	Wrist	Ankle	Mean
HRNet [52]				73.3	68.5	77.3
(a)	✓			78.1	74.3	82.9
(b)	✓	✓		79.7	76.0	84.0
(c)	✓	✓	✓	**80.0**	**77.0**	**84.8**

Table 6. Ablation of different components in FAMI-Pose.

Supp. Frame Window $\mathcal{N}$	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
$\mathcal{N} = \{-1\}$	88.1	89.2	83.9	78.0	83.5	80.7	73.4	82.8
$\mathcal{N} = \{-1, 1\}$	89.1	89.5	84.8	79.0	84.2	82.3	74.9	83.9
$\mathcal{N} = \{-2, -1, 1\}$	89.3	89.8	85.3	79.8	84.2	82.6	76.2	84.5
$\mathcal{N} = \{-2, -1, 1, 2\}$	**89.6**	**90.1**	**86.3**	**80.0**	**84.6**	**83.4**	**77.0**	**84.8**

Table 7. Impact of modifying the supporting frame window.

The quantitative comparisons on the PoseTrack2017 test set are reported in Table 2. Since the pose annotations are not publicly available, we upload our model predictions to the PoseTrack official evaluation server: https://posetrack.net/leaderboard.php to obtain results. FAMI-Pose again surpasses previous state-of-the-art, attaining an mAP of 80.9 ($\uparrow$ 1.7), with an mAP of 81.8, 77.4, 79.1, and 73.6 for the elbow, wrist, knee, and ankle, respectively. As illustrated in in Fig. 3, the visualized results for scenes with rapid motion or pose occlusions attest to the robustness of our method. More visualized results can be found on our project page[1].

Results on the PoseTrack2018 Dataset We further benchmark our model on the PoseTrack2018 dataset. The detailed results on the validation and test sets are tabulated in Table 3 and Table 4, respectively. From these tables, we observe that our FAMI-Pose consistently attains the new state-of-the-art results for all joints. We obtain a 82.2 mAP on the validation set and a 79.6 mAP for the test set.

Results on the Sub-JHMDB Dataset Results for the Sub-JHMDB dataset are reported in Table 5. We observe that existing methods have already achieved an impressive accuracy. Specifically, the current state-of-the-art method MotionAdaptive obtains a 94.7 mAP on this dataset. In

[1] https://github.com/Pose-Group/FAMI-Pose

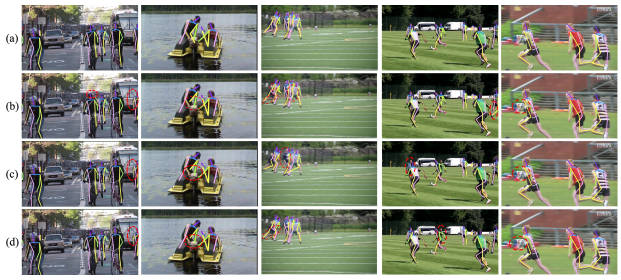

Figure 4. Visual comparisons of the predictions of our FAMI-Pose (a), HRNet-W48 (b), PoseWarper (c), and DCPose (d) on the challenging cases from PoseTrack2017 and PoseTrack2018 datasets. Inaccurate pose estimations are highlighted by the red dotted circles.

contrast, our method is able to achieve a 96.0 mAP. We also obtain a 99.3 mAP for the head joint and a 99.2 mAP for the hip joint. The 1.3 mAP improvement over the already impressive state-of-the-art methods might be an evidence to show the effectiveness of the proposed method.

4.3. Ablation Study

We perform ablation experiments to examine the contribution of feature alignment as well as the influence of each component in our method (*i.e.*, Global Transformation Module, Local Calibration Module, and MI Loss). We also investigate the impact of modifying the predefined window $\mathcal{N}$ of supporting frames. These experiments are conducted on the PoseTrack2017 validation dataset.

Feature Alignment We empirically evaluate the efficacy of proposed components for facilitating and guiding feature alignment in our FAMI-Pose framework. We report the AP for the wrist and ankle joints as well as the mAP for all joints in Table 6. **(a)** For the first setting, we remove the local calibration module and MI loss in FAMI-Pose, employing only the global transformation module (GTM) for feature alignment. Remarkably, the coarse feature alignment with the GTM already improves upon the baseline (HRNet-W48 backbone) by a significant margin of 5.6 mAP and the 82.9 mAP is in fact on par with the previous state-of-the-art 82.8 mAP of DCPose [35]. This corroborates the effectiveness of our approach in introducing feature alignment to facilitate video-based pose estimation. Feature alignment is noticeably more effective in leveraging temporal information from supporting frames as compared to previous methods which adopt optical flow or motion offset estimations. **(b)** For the next setting, we incorporate the local calibration module (LCM) on top of the global alignment to obtain fine-tuned feature alignment. This fine-tuning improves the mAP by 1.1 to 84.0. **(c)** The final setting includes the MI objective and corresponds to our complete FAMI-Pose framework. The improvement of 0.8 mAP provides empirical evidence that our proposed MI loss is effective as an additional supervision to facilitate the learning of complementary task-specific information in temporal features.

Supporting Frames In addition, we investigate the effects of adopting different supporting frame windows $\mathcal{N}$ for pose estimation. The results in Table 7 suggest a performance improvement with higher number of supporting frames, whereby the mAP increases from 82.8 for $\mathcal{N} = \{-1\}$ to 83.9, 84.5, 84.8 at $\mathcal{N} = \{-1, 1\}$, $\mathcal{N} = \{-2, -1, 1\}$, $\mathcal{N} = \{-2, -1, 1, 2\}$, respectively. This is in line with our intuitions, *i.e.*, incorporating more supporting frames enables accessing a larger temporal context with more complementary and useful information that are beneficial for improving the pose estimation on the key frame.

4.4. Comparison of Visual Results

In addition to the quantitative analysis, we further examine the ability of our model to handle challenging scenarios such as rapid motion or pose occlusions. We illustrate in Fig. 4 the side-by-side comparisons of a) our FAMI-Pose against state-of-the-art methods, namely b) HRNet-W48 [52], c) PoseWarper [5], and d) DCPose [35]. It is observed that our approach yields more robust and accurate pose estimates for such challenging scenes. HRNet-W48 is designed for image-based pose estimation and does not incorporate information from supporting frames, resulting in poor performance on degraded video frames. On the other hand, PoseWarper and DCPose implicitly estimate motion cues between frames to improve pose estimation but lack feature alignment and effective supervision on information gain. Through a principled design of the GTM and LCM for progressive feature alignment as well as the MI objective to enhance complementary information mining, FAMI-Pose shows a better ability to handle visual degradation.

5. Conclusion

In this paper, we examine the multi-frame human pose estimation task from the perspective of effectively leveraging temporal contexts through feature alignment and complementary information mining. We present a hierarchical coarse-to-fine network to progressively align supporting frame feature with the key frame feature. Theoretically, we further introduce a mutual information objective for effective supervision on intermediate features. Extensive experiments show that our method delivers state-of-the-art results on three benchmark datasets, PoseTrack2017, PoseTrack2018, and Sub-JHMDB.

6. Acknowledgements

This paper is supported by the National Natural Science Foundation of China (No. 61902348) and the Key R&D Program of Zhejiang Province (No. 2021C01104).

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 6

[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

[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] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018. 5

[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, 5, 6, 7, 8

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[7] Yuntao Chen, Chenxia Han, Naiyan Wang, and Zhaoxiang Zhang. Revisiting feature alignment for one-stage object detection. *arXiv preprint arXiv:1908.01570*, 2019. 3

[8] 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

[9] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pages 1779–1788. PMLR, 2020. 5

[10] Andreas Doering, Umar Iqbal, and Juergen Gall. Joint flow: Temporal flow fields for multi person tracking. *arXiv preprint arXiv:1805.04596*, 2018. 6, 7

[11] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

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

[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. 6

[14] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2019. 5

[15] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 350–359, 2018. 6, 7

[16] Hengkai Guo, Tang Tang, Guozhong Luo, Riwei Chen, Yongchen Lu, and Linfu Wen. Multi-domain pose network for multi-person pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 6, 7

[17] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Alberto Del Bimbo. Adavqa: Overcoming language priors with adapted margin cosine loss. In *IJCAI*, pages 708–714. ijcai.org, 2021. 2

[18] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. Loss re-scaling VQA: revisiting the language prior problem from a class-imbalance view. *TIP*, 31:227–238, 2022. 2

[19] Yanbin Hao, Zi-Niu Liu, Hao Zhang, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. Person-level action recognition in complex events via tsd-tsm networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4699–4702, 2020. 2

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[21] 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

[22] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5700–5709, 2020. 2

[23] Shihua Huang, Zhichao Lu, Ran Cheng, and Cheng He. Fapn: Feature-aligned pyramid network for dense image prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 864–873, 2021. 3

[24] Zilong Huang, Yunchao Wei, Xinggang Wang, Humphrey Shi, Wenyu Liu, and Thomas S Huang. Alignseg: Feature-aligned segmentation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[25] Jihye Hwang, Jieun Lee, Sungheon Park, and Nojun Kwak. Pose estimator and tracker using temporal flow maps for limbs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 6, 7

[26] Umar Iqbal, Martin Garbade, and Juergen Gall. Pose for action-action for pose. In *2017 12th IEEE International*

Conference on Automatic Face & Gesture Recognition (FG 2017), pages 438–445. IEEE, 2017. 6

[27] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 6

[28] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *International Conf. on Computer Vision (ICCV)*, pages 3192–3199, Dec. 2013. 6

[29] Sheng Jin, Wentao Liu, Wanli Ouyang, and Chen Qian. Multi-person articulated tracking with spatial and temporal embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5664–5673, 2019. 6, 7

[30] 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

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

[32] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019. 2

[33] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *European Conference on Computer Vision*, pages 775–793. Springer, 2020. 3

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[35] 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. 2, 3, 5, 6, 7, 8

[36] Zhenguang Liu, Kedi Lyu, Shuang Wu, Haipeng Chen, Yanbin Hao, and Shouling Ji. Aggregated multi-gans for controlled 3d human motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2225–2232, 2021. 1

[37] Zhenguang Liu, Pengxiang Su, Shuang Wu, Xuanjing Shen, Haipeng Chen, Yanbin Hao, and Meng Wang. Motion prediction using trajectory cues. In *ICCV*, pages 13279–13288, 2021. 1

[38] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3266–3275, 2019. 3

[39] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. Lstm pose machines. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5207–5215, 2018. 1, 2, 3, 6

[40] Davide Mazzini. Guided upsampling network for real-time semantic segmentation. *arXiv preprint arXiv:1807.07466*, 2018. 3

[41] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7773–7781, 2019. 3

[42] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 1, 3

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

[44] Dennis Park and Deva Ramanan. N-best maximal decoders for part models. In *2011 International Conference on Computer Vision*, pages 2627–2634. IEEE, 2011. 6

[45] 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, 3

[46] 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

[47] 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, 2

[48] 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

[49] Michael Snower, Asim Kadav, Farley Lai, and Hans Peter Graf. 15 keypoints is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2020. 6

[50] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4220–4229, 2017. 2, 3, 6

[51] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5674–5682, 2019. 2

[52] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3, 5, 6, 7, 8

[53] 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

[54] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1

[55] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018. 5

[56] Ali Varamesh and Tinne Tuytelaars. Mixture dense regression for object detection and human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13086–13095, 2020. 2

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

[58] 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, 5, 6

[59] 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

[60] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 3

[61] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 6

[62] 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. 3, 5, 6, 7

[63] Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. Joint action recognition and pose estimation from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1293–1301, 2015. 6

[64] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977*, 2018. 6, 7

[65] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. 2

[66] Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2):791–805, 2017. 2

[67] 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. 1, 6, 7

[68] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2020. 2

[69] 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, 7

[70] 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 *2009 IEEE 12th International Conference on Computer Vision*, pages 1349–1356. IEEE, 2009. 1, 2

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

[72] 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

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