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

Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, Xiang Wang
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.

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 (Feature Alignment and Mutual Information maximization for Pose 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, PoseTrack2017, 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 deconvo-
uolution layers to replace the commonly used bi-linear inter-
polation 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 en-
tire 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 convolu-
tional 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 po-
tency of the supporting frames. \[45, 50\] explicitly estimate
motion fields by computing optical flow between consecu-
tive 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 es-
timation accuracy would be heavily dependent on the per-
formance of the optical flow or motion offset estimation.
Furthermore, the lack of an effective supervision at the in-
termediate features level for these approaches could lead to
inaccurate pose estimations.

2.3. Feature Alignment

Feature alignment is an important topic for many com-
puter vision tasks (e.g., semantic segmentation \[33, 40\], ob-
ject 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 pool-
ing and upsampling. \[23\] proposes to learn the transforma-
cion 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 aggrega-
tion. Whereas previous methods mostly tackle spatial mis-
alignment 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 indi-
vidual 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 \(I^i_t\) for the key frame and \(\{I^i_{t+\delta} | \delta \in \mathcal{N}\}\) for
the supporting (neighboring) frames.

Problem Formulation Presented with a key frame \(I^i_t\)
along with its supporting frames \(\{I^i_{t+\delta} | \delta \in \mathcal{N}\}\), our goal
is to estimate the pose in \(I^i_t\). 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 il-
lustrated in Fig. 2. For each supporting frame \(I^i_{t+\delta}\), FAMI-
pose performs a two-stage hierarchical transformation to
align \(I^i_{t+\delta}\) with the key frame \(I^i_t\) at the feature level. Specif-
ically, FAMI-Pose consists of two main modules, a global
transformation module and a local calibration module. We
first perform feature extraction on \(I^i_t\) and \(I^i_{t+\delta}\) to obtain \(z^i_t\)
and \(z^i_{t+\delta}\), respectively. These features are then fed into
our global transformation module, which learns the parameters
of an affine transformation to obtain a coarsely aligned sup-
porting frame feature \(\frac{1}{\delta} z^i_t\). \(z^i_t\) and \(z^i_{t+\delta}\) are then handed
to the local calibration module, which performs pixel-wise
deformation to produce finely aligned features \(\tilde{z}^i_t\) and \(\tilde{z}^i_{t+
\delta}\). Finally, we aggregate all aligned supporting frame features
\(\{\tilde{z}^i_{t+\delta} | \delta \in \mathcal{N}\}\) and the key frame feature \(z^i_t\) to obtain our
enhanced feature \(\tilde{z}^i_t\). \(\tilde{z}^i_t\) is passed to a detection head that
outputs pose estimations \(\tilde{H}^i_t\). The task objective is to min-
imize the heatmap estimation loss \(L_{H}\) which measures the
discrepancy between \(\tilde{H}^i_t\) and the ground truth \(H^i_t\). On top
of this, we also design a mutual information objective \(L_{MI}\)
which effectuates a feature level supervision for maximiz-
ing the amount of complementary task-relevant information
encoded in \(\tilde{z}^i_t\). In what follows, we introduce the complete
FAMI-Pose architecture and the mutual information objec-
tive 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 \(z^i_t\) and \(z^i_{t+\delta}\) are then
passed through a global transformation module and a local
calibration module, to progressively align \(z^i_{t+\delta}\) with \(z^i_t\). 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 move-
ments 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 \( I_t \) with the assistance of its supporting frames. For clarity of illustration, we only show a single supporting frame \( I_{t+\delta} \) 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 \( \tilde{z}_{t+\delta}^i \) for all supporting frames are aggregated to \( \tilde{z}_t^i \), which is passed to a detection head that outputs pose estimates \( \hat{H}_t \). Besides the heatmap estimation loss \( L_H \), we introduce an additional feature level supervision through our Mutual Information objective \( L_{M1} \) 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 \( \phi \) that estimates affine transformation parameters \( \Theta \) 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 \( \Theta \) 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 \( \mathcal{T} : (z_{t+\delta}^i, \Theta) \rightarrow \tilde{z}_{t+\delta}^i \).

The operations of the GTM can be expressed as follows:

\[
\Theta = \phi(z_t^i \oplus z_{t+\delta}^i),
\]

\[
\begin{pmatrix}
{x_p} \\
{y_p}
\end{pmatrix} = \begin{bmatrix}
\theta_{11} & \theta_{12} & \theta_{13} \\
\theta_{21} & \theta_{22} & \theta_{23}
\end{bmatrix}
\begin{pmatrix}
\bar{x}_p \\
\bar{y}_p \\
1
\end{pmatrix},
\]

(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 \( \tilde{z}_{t+\delta}^i \), 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 \( \tilde{z}_{t+\delta}^i \).

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

\[
\tilde{z}_{t+\delta}^i \oplus z_t^i \xrightarrow{\text{residual}} \tilde{z}_{t+\delta}^i \xrightarrow{\text{regular}} O, \\
\tilde{z}_{t+\delta}^i \oplus z_t^i \xrightarrow{\text{residual}} \tilde{z}_{t+\delta}^i \xrightarrow{\text{regular}} M.
\]

(2)

The adaptively learned kernel offsets \( O \) and modulated scalars \( M \) respectively correspond to location shifts and intensity fluctuations of each pixel in \( \tilde{z}_{t+\delta}^i \) 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 \( \tilde{z}_{t+\delta}^i \), the kernel sampling offsets \( O \), and the modulated scalars \( M \) as inputs, the modulated deformable convolution outputs the fine-tuned feature \( \tilde{z}_{t+\delta}^i \):

\[
(\tilde{z}_{t+\delta}^i, O, M) \xrightarrow{\text{modulated deformable convolution}} \tilde{z}_{t+\delta}^i.
\]

(3)

To anticipate the discussion of the mutual information loss, we would like to point out that the key frame
ture \( z_i^t \) 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 \( \tilde{z}_i^{t+\delta} \).

**Heatmap Generation** Ultimately, we aggregate over all final aligned supporting frame features \{\( \tilde{z}_i^{t+\delta} \ | \ \delta \in \mathcal{N} \)} and the key frame feature \( z_i^t \) via element-wise addition to obtain the enhanced feature \( \tilde{z}_i^t \). \( \tilde{z}_i^t \) is fed to a detection head to produce pose heatmap estimations \( \hat{H}_i^t \). We implemented the detection head using a stack of 3x3 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 \( z_i^{t+\delta} \) and the key frame feature \( z_i^t \). 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 \( \tilde{z}_i^t \).

**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 \( v_1 \) and \( v_2 \):

\[
\mathcal{I}(v_1; v_2) = \mathbb{E}_{p(v_1, v_2)} \left[ \log \frac{p(v_1, v_2)}{p(v_1)p(v_2)} \right],
\tag{4}
\]

where \( p(v_1, v_2) \) is the joint probability distribution between \( v_1 \) and \( v_2 \), while \( p(v_1) \) and \( p(v_2) \) are their marginals.

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

\[
\max \mathcal{I}(y_i^t; \tilde{z}_i^t \mid z_i^t),
\tag{5}
\]

where \( y_i^t \) represents the label, and \( \mathcal{I}(y_i^t; \tilde{z}_i^t \mid z_i^t) \) denotes the amount of task-relevant information in the enhanced feature \( \tilde{z}_i^t \), complementary to \( \{i.e., \) excluding\} the information from the key frame feature \( z_i^t \). 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:

\[
\mathcal{I}(y_i^t; \tilde{z}_i^t \mid z_i^t) = \mathcal{I}(y_i^t; \tilde{z}_i^t) - \mathcal{I}(\tilde{z}_i^t; z_i^t) + \mathcal{I}(\tilde{z}_i^t; z_i^t \mid y_i^t),
\tag{6}
\]

where \( \mathcal{I}(y_i^t; \tilde{z}_i^t) \) measures the relevance of the label \( y_i^t \) and feature \( \tilde{z}_i^t \), \( \mathcal{I}(\tilde{z}_i^t; z_i^t) \) indicates the dependence between the two features \( \tilde{z}_i^t \) and \( z_i^t \), and \( \mathcal{I}(\tilde{z}_i^t; z_i^t \mid y_i^t) \) represents the *task-irrelevant information* in both \( \tilde{z}_i^t \) and \( z_i^t \). 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:

\[
\mathcal{I}(y_i^t; \tilde{z}_i^t \mid z_i^t) \rightarrow \mathcal{I}(y_i^t; \tilde{z}_i^t) - \mathcal{I}(z_i^t; \tilde{z}_i^t). \tag{7}
\]

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

\[
\min \left[ \mathcal{I}(y_i^t; \tilde{z}_i^{t+\delta} \mid \tilde{z}_i^t) + \mathcal{I}(y_i^t; \tilde{z}_i^t \mid \tilde{z}_i^t) \right]. \tag{8}
\]

The terms \( \mathcal{I}(y_i^t; \tilde{z}_i^{t+\delta} \mid \tilde{z}_i^t) \) and \( \mathcal{I}(y_i^t; \tilde{z}_i^t \mid \tilde{z}_i^t) \) respectively measure the vanishing task-relevant information in \( \tilde{z}_i^{t+\delta} \) and \( \tilde{z}_i^t \) during feature alignment. They serve to facilitate the nondestructive propagation of information. Simultaneously minimizing these two terms would prevent excessive information loss in \( \tilde{z}_i^{t+\delta} \) and \( \tilde{z}_i^t \) 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:

\[
\mathcal{I}(y_i^t; \tilde{z}_i^{t+\delta} \mid \tilde{z}_i^t) \rightarrow \mathcal{I}(y_i^t; \tilde{z}_i^{t+\delta}) - \mathcal{I}(\tilde{z}_i^{t+\delta}; \tilde{z}_i^t),
\tag{9}
\]

\[
\mathcal{I}(y_i^t; \tilde{z}_i^t \mid \tilde{z}_i^t) \rightarrow \mathcal{I}(y_i^t; \tilde{z}_i^t) - \mathcal{I}(\tilde{z}_i^t; \tilde{z}_i^t). \tag{9}
\]

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

\[
\mathcal{L}_{MI} = \mathcal{I}(y_i^t; \tilde{z}_i^t) + \mathcal{I}(y_i^t; \tilde{z}_i^t \mid \tilde{z}_i^t) - \alpha \cdot \mathcal{I}(y_i^t; \tilde{z}_i^t \mid \tilde{z}_i^t), \tag{10}
\]

where \( \alpha \) 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.
3.3. Training Objective

Our training objective consists of two parts. (1) We adopt the heatmap estimation loss function $L_H$ to supervise the learning of final pose estimates:

$$ L_H = \left\| \hat{H}_i - H_i \right\|^2_2, \quad (11) $$

where $\hat{H}_i$ and $H_i$ 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:

$$ L_{\text{total}} = L_H + \beta \cdot L_{\text{MI}}, \quad (12) $$

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 593 for training, 170 for validation, and 375 for testing. Both datasets are annotated with 15 joints, with 593 for training, 170 for validation, and 375 for testing.

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 $384 \times 288$. We perform data augmentation including random rotation $[-45^\circ, 45^\circ]$, random scaling $[0.65, 1.35]$, random truncation, and horizontal flipping. The predefined window $N$ 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-
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 8th, 19th, and 60th 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 (↑ 1.6) for wrists and an mAP of 77.0 (↑ 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.

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 (↑ 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.

**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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Global Transformation</th>
<th>Local Calibration</th>
<th>MI Loss</th>
<th>Wrist</th>
<th>Ankle</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRNet [52]</td>
<td></td>
<td></td>
<td></td>
<td>73.3</td>
<td>74.5</td>
<td>77.3</td>
</tr>
<tr>
<td>(a)</td>
<td>✓</td>
<td></td>
<td></td>
<td>76.1</td>
<td>74.3</td>
<td>82.9</td>
</tr>
<tr>
<td>(b)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>79.7</td>
<td>76.0</td>
<td>84.0</td>
</tr>
<tr>
<td>(c)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.0</td>
<td>77.0</td>
<td>84.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Supp. Frame Window $N$</th>
<th>Head</th>
<th>Shoulder</th>
<th>Elbow</th>
<th>Wrist</th>
<th>Hip</th>
<th>Knee</th>
<th>Ankle</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N = [-1]$</td>
<td>88.1</td>
<td>89.2</td>
<td>83.9</td>
<td>78.0</td>
<td>83.5</td>
<td>80.7</td>
<td>73.4</td>
<td>82.8</td>
</tr>
<tr>
<td>$N = [-1.1]$</td>
<td>89.1</td>
<td>88.5</td>
<td>84.8</td>
<td>79.0</td>
<td>84.2</td>
<td>82.3</td>
<td>74.9</td>
<td>83.9</td>
</tr>
<tr>
<td>$N = [-2]$</td>
<td>89.3</td>
<td>89.8</td>
<td>85.3</td>
<td>78.5</td>
<td>84.2</td>
<td>82.6</td>
<td>76.2</td>
<td>84.5</td>
</tr>
<tr>
<td>$N = [-2,-1]$</td>
<td>89.6</td>
<td>90.1</td>
<td>86.3</td>
<td>80.0</td>
<td>84.6</td>
<td>83.4</td>
<td>77.0</td>
<td>84.8</td>
</tr>
</tbody>
</table>

Table 7. Impact of modifying the supporting frame window.

[1]https://github.com/Pose-Group/FAMI-Pose
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 \( N \) 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 \( N \) 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 \( N = \{-1\} \) to 83.9, 84.5, 84.8 at \( N = \{-2, -1, 1\} \), \( N = \{-2, -1, 1, 2\} \), 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


11015


