GLPose: Global-Local Representation Learning for Human Pose Estimation

YINGYING JIAO and HAIPENG CHEN*, Jilin University, China
RUNYANG FENG*, HAOMING CHEN, and SIFAN WU, Zhejiang Gongshang University, China
YIFANG YIN, Institute for Infocomm Research, A*STAR, Singapore
ZHENGUANG LIU, Zhejiang University, China

Multi-frame human pose estimation is at the core of many computer vision tasks. Although state-of-the-art approaches have demonstrated remarkable results for human pose estimation on static images, their performances inevitably come short when being applied to videos. A central issue lies in the visual degeneration of video frames induced by rapid motion and pose occlusion in dynamic environments. This problem, by nature, is insurmountable for a single frame. Therefore, incorporating complementary visual cues from other video frames becomes an intuitive paradigm. Current state-of-the-art methods usually leverage information from adjacent frames, which unfortunately place excessive focuses on only the temporally nearby frames. In this paper, we argue that combining global semantically similar information and local temporal visual context will deliver more comprehensive and more robust representations for human pose estimation. Towards this end, we present an effective framework, namely global-local enhanced pose estimation (**GLPose**) network. Our framework consists of a feature processing module that conditionally incorporates global semantic information and local visual context to generate a robust human representation and a feature enhancement module that excavates complementary information from this aggregated representation to enhance keyframe features for precise estimation. We empirically find that the proposed GLpose outperforms existing methods by a large margin and achieves new state-of-the-art results on large benchmark datasets.

CCS Concepts: • **Computing methodologies** → *Activity recognition and understanding*; **Computer vision**.

Additional Key Words and Phrases: Human pose estimation, feature aggregation, pose estimation, global-local representation

ACM Reference Format:
Yingying Jiao, Haipeng Chen, Runyang Feng, Haoming Chen, Sifan Wu, Yifang Yin, and Zhenguang Liu. 2022. GLPose: Global-Local Representation Learning for Human Pose Estimation. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1, Article 1 (January 2022), 19 pages. https://doi.org/10.1145/3519305

*Corresponding authors: Runyang Feng and Haipeng Chen

Authors' addresses: Yingying Jiao, jiaoyingy17@gmail.com; Haipeng Chen, chenhp@jlu.edu.cn, Jilin University, Changchun, China; Runyang Feng, runyang2019.feng@gmail.com; Haoming Chen, chenhaomingbob@gmail.com; Sifan Wu, wusifan2021@gmail.com, Zhejiang Gongshang University, Hangzhou, China; Yifang Yin, yin_yifang@i2r.a-star.edu.sg, Institute for Infocomm Research, A*STAR, Singapore, Singapore; Zhenguang Liu, liuzhenguang2008@gmail.com, Zhejiang University, Hangzhou, China.

1551-6857/2022/1-ART1 $15.00
https://doi.org/10.1145/3519305

1 INTRODUCTION

Estimating keypoint positions of each person from an image is a fundamental challenge in the area of computer vision. It has a wide spectrum of applications including human behavior understanding, action recognition, augmented reality, and surveillance tracking [15, 16, 26–29, 37, 42, 55]. Recent attempts address this problem by employing deep convolutional neural networks (CNNs), and have witnessed rapid advancements [5, 11, 30, 31, 44, 49]. Renewed detection frameworks [15, 23], potent visual backbone networks [19, 40], and large-scale benchmark datasets [1, 25] jointly push forward the performance boundary of human pose estimation.

Multi-frame human pose estimation has recently emerged as a rising challenge beyond recognizing human poses in static images. Benefiting from the technological advances in image-based pose estimation models and convolutional neural networks, the performance of human joints detectors is nearly saturated in regular scenarios (*e.g.*, single-person scenes with minimal occlusion). Unfortunately, the application of these models on video sequences featuring dynamic and complex environments suffers from performance diminishment and flicking. The central issue lies in the appearance degeneration of human subjects induced by the motion blur, video out-of-focus, and frequent pose occlusions that are imperceptible in static image setting. Static image pose estimators tend to fail in handling such cases. A video, on the other hand, contains more abundant visual information. Integrating available cues from other video frames thus becomes an intuitive paradigm for mitigation of visual degradation.

Several methods [4, 8, 28, 60] suggest incorporating temporal information from adjacent frames to compete against visual degeneration. [4] proposes PoseWarper to propagate pose annotations between a pair of frames, and further aggregate temporal pose information from neighboring frames. [28] presents a DCPose framework to aggregate spatial features over nearby frames and model their motion contexts. By leveraging relatively limited local visual context within a short temporal window, these methods have demonstrated superior performance. However, the richer global semantic cues in videos are neglected, which leads to their failures in the cases of contiguous pose occlusions. Another line of work instead considers capturing video-level evidence to bridge the degraded features. [7, 34, 38] propose to estimate optical flow between video frames, and employ the flow based motion field for temporally aligning features. These approaches produce promising results when the flow cues can be computed precisely. However, the computation of optical flow relies heavily on motion estimation, which inevitably suffers from image quality degradation. Thus in cases involving defocus or pose occlusions, the optical flows usually perform unsatisfactorily and hence are not that helpful for human pose estimation. An additional issue of the above approaches is that they directly perform feature aggregation either by concatenation or element-wise summation, and fails to discover valuable information for the key frame, which also limits their performance.

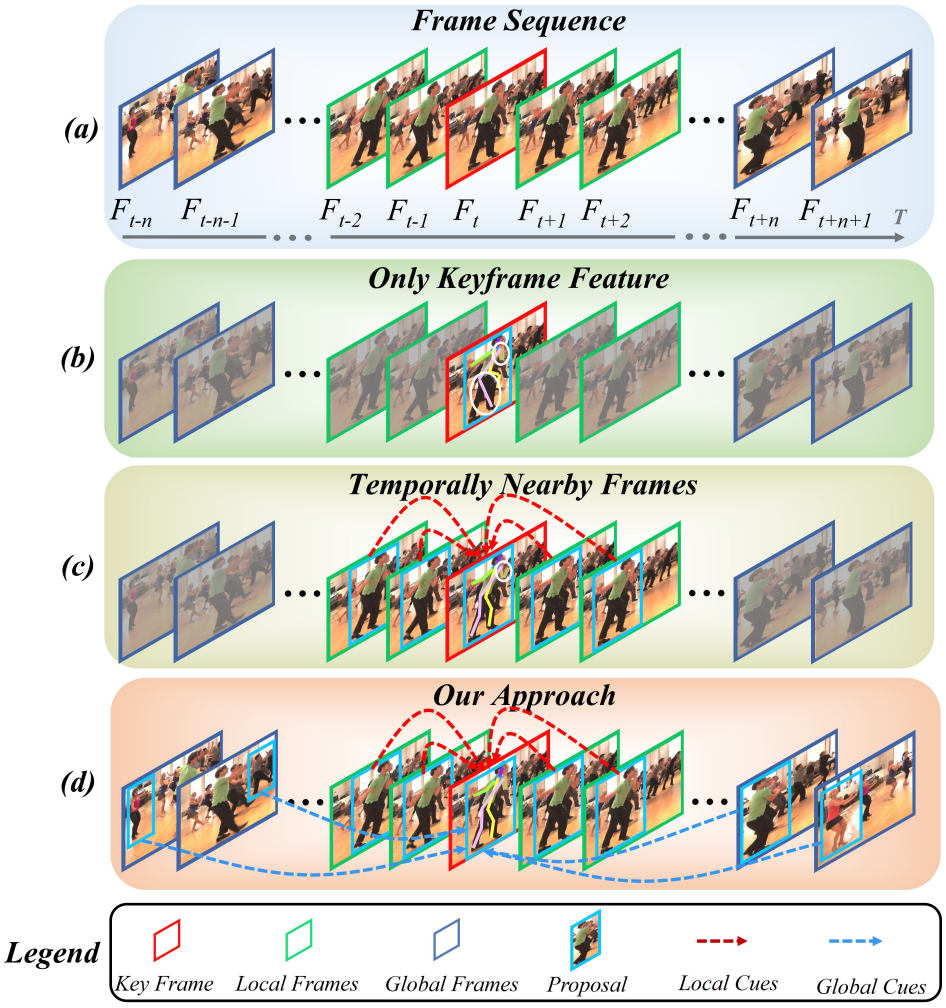

Fig. 1. An illustration of the motivation for our method. (a) Original frame sequence in the PoseTrack dataset. The goal is to detect the pose of the persons in the key frame F_t. We show three different schemes: (b) image based approaches ([40]) employ only the information of the key frame; (c) several recent methods ([4]) incorporate the temporally nearby frames into the pose estimation of the key frame; and (d) we propose to aggregate global semantic information and local visual context for effective pose estimation. By exploiting richer cues from both non-local and local frames, our method learns a comprehensive representation that is robust to challenging dynamic scenes. Inaccurate keypoint detections are highlighted with the white solid circles.

In this paper, we argue that learning a robust human representation requires looking at both *global* semantically similar information and *local* temporal visual context. Intuitively,

visual cues from neighboring frames are closely related to the current frame, which ensures the stability of feature aggregation. *On the other hand, any instances from other frames sharing highly semantic similarity (e.g., similar actions) with the current human subject might be useful, which provide discriminative and complementary information.* As illustrated in Fig. 1 (d), different human proposals might also contribute to the pose estimation of the current human subject. The composition of global semantical information and local visual context will yield more comprehensive and more robust representations. Starting from this concept, we present a global-local enhanced pose estimation (GLPose) network, which consists of a Feature Processing module and a Feature Enhancement module. Specifically, features of human subjects are first extracted from sampled video frames, and then go through our Feature Processing module for conditional aggregation. The aggregated features are then handed to the Feature Enhancement module to yield the enhanced discriminative keyframe features. Finally, a detection head [40, 51] is used to output heatmap estimates. Notably, unlike previous approaches, we design the Feature Enhancement module to further explore the valuable and complementary information for the key frame after obtaining the aggregated representations, delivering more accurate pose estimation results.

Our contributions are summarized as follows.

- We propose to incorporate global semantically similar information and local temporal visual context to deliver more comprehensive and more robust representations for human pose estimation.
- The proposed global-local enhanced pose estimation network is able to fully leverage video information by conditional feature aggregation and complementary feature enhancement, effectively competing against frame degradation.
- Our method achieves new state-of-the-art results on three benchmark datasets, namely PoseTrack2017, PoseTrack2018, and Sub-JHMDB.

2 RELATED WORK

2.1 Human Pose Estimation in Static Images

Conventional image-based human pose estimation approaches consider probabilistic graphical models or the pictorial structure models [12, 36, 41, 46, 48, 59] to represent relations between connected body parts. Notwithstanding the efficient inference, they rely heavily on hand-craft features and tend to fail in uncommon situations. More recently, deep convolutional neural networks have emerged as dominant solutions [2, 10, 24, 39, 43, 45, 57] due to their superior performance. One line of work [6] directly regresses skeletal keypoint coordinates from the input image, which has later been surpassed by the heatmap based methods [28, 40].

In general, heatmap estimation based methods broadly fall within two paradigms, top-down and bottom-up. *Bottom-up* paradigm first detects individual body joints and then assembles them into entire persons. [5] presents a bottom-up architecture that leverages part affinity field to represent pairwise relationships between body parts. Conversely, *top-down* paradigm first

detects persons in an image and then proceeds with single-person pose estimation on each individual. [31] designs a symmetric stacked hourglass architecture based on the successive steps of pooling and upsampling, incorporating features across all scales to yield a robust representation. [15] presents a regional multi-person pose estimation framework to tackle the problem of pose estimation within inaccurate human bounding boxes. A recent work in [40] proposes a HRNet architecture that maintains high resolution features through the whole process, achieving state-of-the-art results in static images.

2.2 Human Pose Estimation in Videos

For human pose estimation in videos, a principle problem is to utilize the abundant temporal information (*e.g.*, local temporal consistency and global semantic cues) to improve the accuracy of joints detectors. Some previous approaches propose to integrate temporal information from neighboring frames into pose estimation of current frame. [4] presents a PoseWarper which is able to incorporate spatiotemporal pose information from adjacent frames in the inference stage. [28] proposes a DCPose framework, which encodes keypoint context over consecutive frames into localized search scopes and further resample heatmap within this range according to the pose residuals. These methods perform temporal aggregation across local adjacent frames while disregarding global semantic information, which potentially limits their performance.

Another line of work focus on capturing video-level information to facilitate human pose estimation. [34, 38] propose to compute optical flow between video frames and employs flow based representation for feature calibration. However, motion blur and video defocus hinder optical flow computation which translates to performance drop. [2, 30] propose to employ convolutional LSTM to model the sequential spatiotemporal features, achieving good results on single person scenes. [47] presents a 3D-HRNet that incorporates temporal dimension into the original HRNet [40], performing successive feature aggregation for pose tracking and estimation. Unlike previous methods, our global-local enhanced pose estimation (GLPose) network combines global semantic information and local visual context to provide more robust representations for human pose estimation, achieving state-of-the-art performance.

3 OUR APPROACH

Motivation In order to cope with video frame degradation for multi-frame human pose estimation, feature aggregation is an effective solution, in which there are two main issues: (1) selecting *suitable features* for aggregation; and (2) adopting a *valid strategy* to aggregate multiple features. Several previous approaches [4, 28, 56] intuitively incorporate the *temporally nearby frames*. However, contiguous frame degeneration often occurs over a wide time window, which limits the effectiveness of local cues. Moreover, existing methods directly perform *concatenation or element-wise summation* among features for aggregation, without highlighting the information that is overlooked by the key frame. To tackle the above problems, we propose

to aggregate both global semantically similar information and local temporal visual context to generate more robust representations. We further mine discriminative information from this aggregated representation to enhance the features of key frame, delivering effective feature aggregation.

Method Overview The pipeline of our proposed global-local enhanced pose estimation (GLPose) network is illustrated in Fig. 2. To improve pose estimation for person i in key frame $F_t = f_t^i$, we simultaneously incorporate local temporal information from $F_l = \{f_{t-k}^i, ..., f_{t-1}^i, f_{t+1}^i, ..., f_{t+k}^i\}$ and global semantic cues from $F_g = \{f_n^j\}_{n\in N}^{j\in J}$. Note that N represents the number of video frames, symbol J denotes the number of human proposals within a video frame, and subscripts i and j denote i^{th} and j^{th} person, respectively. Specifically, we first extract the features of frames $\{F_t, F_l, F_g\}$ and feed them into our Feature Processing module, which outputs the spatiotemporal aggregation $\mathcal{S}_i$. The aggregated representation $\mathcal{S}_i$ and the keyframe features are then processed through our Feature Enhancement module, which delivers the final enhanced features $\mathcal{F}_i$. $\mathcal{F}_i$ is then handed to a detection head to yield the final estimated heatmaps H_i. In what follows, we introduce the proposed key modules in detail.

3.1 Feature Processing Module

Ideally, incorporating features within the tracklet of person i will produce optimal representations. However, the ground truth association for human proposals across frames is not available in the test stage. We instead propose the Feature Processing module that combines temporally proximate information within *short tracklet* F_l and semantically similar cues from *non-local frames* F_g to yield a robust representation. There are three key procedures: feature extraction, similarity definition, and conditional feature aggregation.

Feature Extraction We first employ a shared feature extract network ϕ to extract the visual features $\{V_t^i, V_l^i, V_g^i\}$ from frames $\{F_t, F_l, F_g\}$, respectively:

$$\mathbf{V}_z^i = \phi(F_z), \quad z = \{t, l, g\}. \tag{1}$$

In order to obtain more representative features for subsequent computations, we implement the network ϕ through current state-of-the-art human joints detector HRNet-W48 [40]. The extracted features are then leveraged for similarity definition and conditional aggregation.

Similarity Definition We compute the similarity between reference frames $\{F_l, F_g\}$ and key frame F_t to guide the following aggregation. For the localized tracklet F_l, each frame in F_l shares the identical person with key frame F_t, and thus they are superficially similar. As a result, intuitively leveraging such low-identified feature similarity to guide feature aggregation is problematic. To address this issue, we explicitly consider the temporal distance between frames F_l and F_t. In other words, we assign higher weights to the frames that are temporally nearer to the key frame F_t, and vice versa. This computation can be expressed as follows:

$$\mathbf{W}_l^i = \mathcal{T}(F_l, F_t), \tag{2}$$

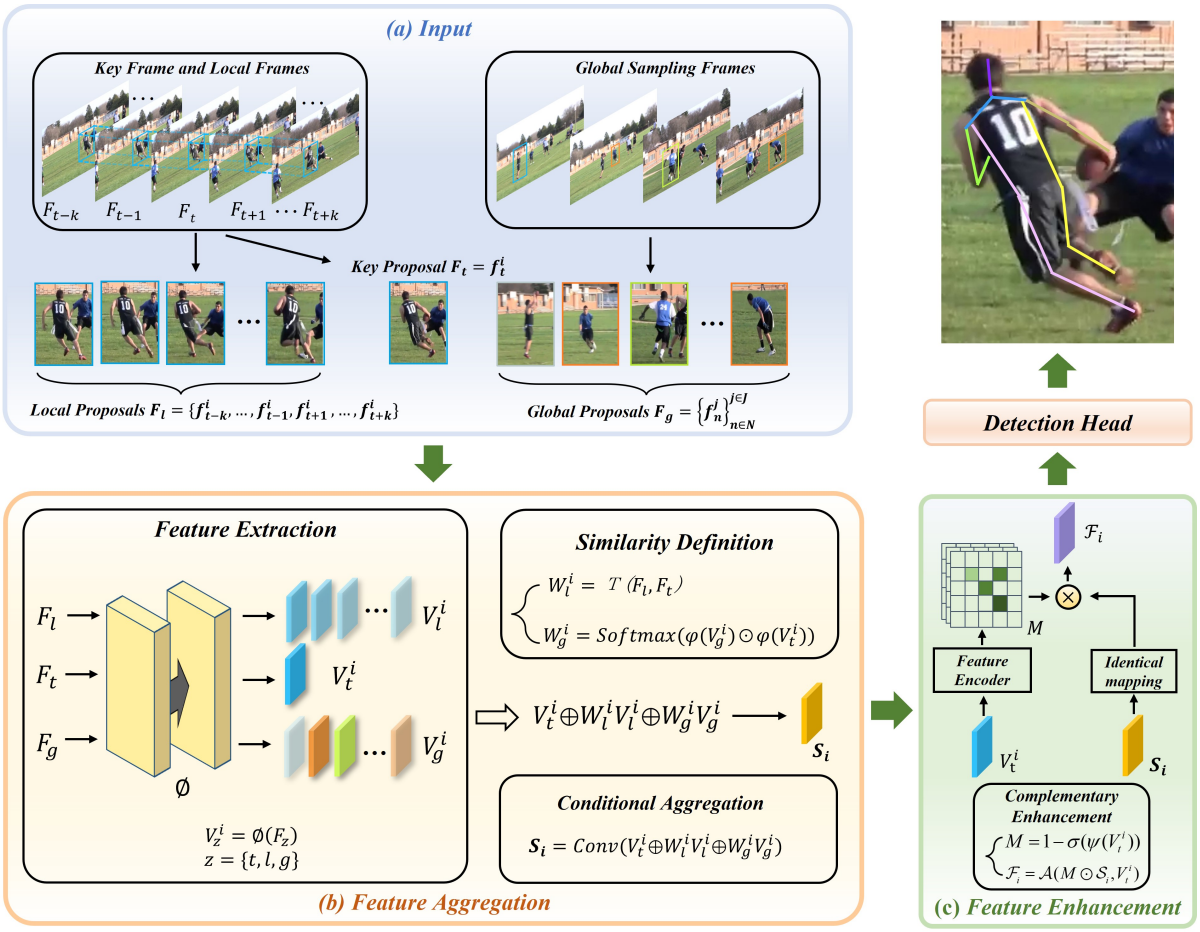

Fig. 2. Overall pipeline of our GLPose network. The goal is to detect the pose of person i in the keyframe F_t. We first sample the local tracklet F_l and non-local similar frames F_g, and feed them into the Feature Processing module, which performs conditional feature aggregation and outputs $\mathcal{S}_i$. The aggregated features $\mathcal{S}_i$ is then processed by the Feature Enhancement module, which provides the enhanced keyframe features $\mathcal{F}_i$. Finally, a detection head is leveraged to output the pose estimation.

where $\mathcal{T}(\cdot)$ denotes the distance function, and $\mathbf{W}_l^i$ denotes the similarity between F_l and F_t.

Conversely, the reference human proposals within non-local frames F_g are not temporally associated to the key frame F_t, and we treat the feature semantical similarity between V_g^i and V_t^i as guidance for aggregation. In particular, given features V_g^i and V_t^i, their semantical similarity $\mathbf{W}_g^i$ can be computed by the following procedures. (i) Feature tensors V_g^i and V_t^i are fed into a shared feature embedding layer, which translates the feature space and outputs an embedding for each frame. (ii) The frame level embeddings are then employed to compute corresponding

similarity:

$$\mathbf{W}_g^i = Softmax(\varphi(V_g^i) \odot \varphi(V_t^i)), \tag{3}$$

where $\varphi(\cdot)$ and $\odot$ indicate the feature embedding layer and Hadamard matrix product operation, respectively. It is worth mentioning that the similarity between features V_g^i and V_t^i is built upon the frame level, which provides a more robust guidance for feature aggregation than pixel-level models ([7, 34]).

Conditional Feature Aggregation Given visual features V_t^i, V_l^i, V_g^i and corresponding similarity $\mathbf{W}_l^i$ and $\mathbf{W}_g^i$, the feature aggregation conditioned on semantical similarity is defined as:

$$V_t^i \oplus \mathbf{W}_l^i V_l^i \oplus \mathbf{W}_g^i V_g^i \xrightarrow[\text{Basic Blocks}]{\text{stack of}} \mathcal{S}_i, \tag{4}$$

where $\oplus$ denotes the concatenation operation. We modify a residual structure from [19] to implement the Basic Blocks in Eq. 4. By conditionally aggregating global semantic information and local visual context across multiple frames, the new representation $\mathcal{S}_i$ contains more abundant information that allows effectively competing against appearance degeneration such as pose occlusion and motion blur.

3.2 Feature Enhancement Module

Despite allowing for incorporation of richer features within the Feature Processing module, *straightforwardly weighted aggregation across multiple features still has difficulties in perceiving information that is valuable for the pose estimation of key frame* F_t. Previous approaches [9, 51] directly output final estimations according to such features, incurring undesirable results. Instead, we propose the Feature Enhancement module that purposefully excavates discriminative cues from the aggregated representation $\mathcal{S}_i$ to enhance the keyframe features V_t^i.

The architecture of our Feature Enhancement module is illustrated in Fig. 2. Specifically, we first encode the keyframe features V_t^i into a weight matrix, which reveals the activeness of each pixel location within feature maps V_t^i. The weight matrix is then used as a mask to mine the information that is overlooked by the key frame from $\mathcal{S}_i$. Finally, keyframe features and complementary information are aggregated to yield the enhanced features $\mathcal{F}_i$. This computation can be defined as follows:

$$\mathcal{F}_i = \mathcal{A}(\underbrace{\big(1 - \sigma(\psi(V_t^i))\big)}_{\text{weight matrix}} \odot \mathcal{S}_i, V_t^i), \tag{5}$$

where $\psi(\cdot)$, σ, and $\mathcal{A}$ denote feature encoder, Sigmoid function, and aggregation transformation, respectively. After obtaining the final enhanced features $\mathcal{F}_i$ of key frame, we adopt a detection head to output the heatmaps $\mathbf{H}_i$ of person i. Note that the detection head is implemented with a regular 3×3 convolution layer.

Ultimately, the above procedure is performed for each individual person i. By adequately aggregating the auxiliary features and purposefully digging the effective cues in them, the final estimated pose heatmaps are more spatially accurate. We also demonstrate the effectiveness of our Feature Enhancement module in the ablation experiments (Sec. 4.3).

3.3 Implementation Details

Frame Sampling Given an input video, we first employ an object detector (Cascaded-RCNN) to detect human bounding boxes for each person in all frames. To obtain the localized short tracklet F_l, the bounding box of person i in the key frame F_t is enlarged 25% and then utilized to crop the identical person in a temporal window $[t-k, ..., t-1, t+1, ..., t+k]$. For the non-local frames F_g, we first sample 20 human proposals in the video level, then rank them according to the semantical similarity between them and the human proposal of the key frame, and finally select the top n frames for feature aggregation.

Network Structures We leverage the HRNet-W48 [40] pretrained on the COCO and PoseTrack dataset as our feature extractor. In fact, HRNet-W48 usually serves as an out-of-box human joints detector which directly gives the final pose heatmaps, whilst we adopt the output of *HRNet-Stage3-Branch0* as the visual features $\{V_z^i\}_{z=t,l,g}$. The shared feature embedding layer φ in Similarity Definition is implemented with a Global Average Pooling (GAP) layer, which adaptively performs average pooling over an input signal composed of several input planes. Three basic blocks are employed for conditional feature aggregation.

Loss Function The standard pose estimation loss function [40, 52] is adopted as our cost function. Training aims to reduce the total Euclidean or L2 distance between the prediction and the ground truth heatmaps for all joints. The loss function is defined as:

$$L = \frac{1}{M} \sum_{m=1}^{M} v_m \times \|G(m) - P(m)\|^2, \tag{6}$$

where $G(m)$, $P(m)$, v_m denote the ground truth heatmap, prediction heatmap, and visibility of joint m. During the training phase, the total number of joints is set to $M = 15$. The ground truth heatmaps are generated by a 2D Gaussians centered on the positions of the joints.

4 EXPERIMENTS

In this section, we present our experiments on three widely used benchmark datasets PoseTrack2017, PoseTrack2018, and Sub-JHMDB. We first introduce the detailed experimental settings in Sec. 4.1, including datasets and parameter settings. We then compare our GLPose method with state-of-the-art methods in terms of quantitative results and visual results (Sec. 4.2). Finally, we conduct extensive ablation experiments in Sec. 4.3 to examine the effectiveness of each proposed component in our method.

Table 1. Quantitative comparisons with state-of-the-art methods on the PoseTrack2017 validation set.

Method	Backbone	Additional Training Data	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
Dataset: **PoseTrack2017 Validation** set.										
PoseTracker [17]	ResNet-3D	COCO	67.5	70.2	62.0	51.7	60.7	58.7	49.8	60.6
PoseFlow[54]	-	MPII Pose + COCO	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
JointFlow[13]	-	-	-	-	-	-	-	-	-	69.3
FastPose[58]	-	-	80.0	80.3	69.5	59.1	71.4	67.5	59.4	70.3
SimpleBaseline[52]	ResNet-50	COCO	79.1	80.5	75.5	66.0	70.8	70.0	61.7	72.4
SimpleBaseline[52]	ResNet-152	COCO	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
STEmbedding[22]	4-stage Stacked Hourglass	-	83.8	81.6	77.1	70.0	77.4	74.5	70.8	77.0
HRNet[40]	HRNet-W48	COCO	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
MDPN[18]	SimpleBaseline	MPII Pose + COCO	85.2	88.5	83.9	77.5	79.0	77.0	71.4	80.7
Dynamic[56]	HRNet-W48	COCO	88.4	88.4	82.0	74.5	79.1	78.3	73.1	81.1
PoseWarper[4]	HRNet-W48	COCO	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
GLPose (Ours)	HRNet-W48	COCO	**88.1**	**88.9**	**84.1**	**78.1**	**83.5**	**81.5**	**74.2**	**83.1**

4.1 Experimental Settings

Datasets PoseTrack is a large-scale benchmark video dataset for human pose estimation and articulated tracking. The PoseTrack2017 dataset contains 514 videos with 16,219 pose annotations, in which 250 videos are for training and 50 videos are for validation. The PoseTrack2018 dataset increases the number of videos to 1,138 and contains 153,615 pose annotations. These are split into 593 and 170 videos respectively for training and validation. In training videos, dense annotations for 30 center frames of a video are provided. In validation videos, human poses are annotated every four frames. Both datasets label 15 joints, with an additional annotation label for joint visibility. We benchmark our model on PoseTrack2017 and PoseTrack2018 datasets with the metric of average precision (**AP**). Note that only the visible joints are used for performance evaluation. The Sub-JHMDB dataset includes 319 videos for a total of 11, 200 frames. The annotations are provided for 15 joints but only for visible joints. Following previous works [30, 32, 60], we perform three different data splits for this dataset, each with a training to testing ratio of 3 : 1, and report the mean accuracy over the three splits.

Parameter Settings Our GLPose is implemented on PyTorch. During training, we incorporate data augmentation strategies including random rotation $[-45°, 45°]$, random scaling $[0.65, 1.35]$, truncation and horizontal flip. Input image size is fixed to 384×288. We utilize a total of 8 local frames ($k = 4$) and 6 non-local frames ($n = 6$) for feature aggregation. All subsequent weight parameters are randomly initialized from a Gaussian distribution with $\mu = 0$ and $\sigma = 0.001$, while bias parameters are initialized to 0. We employ the Adam optimizer

Table 2. Quantitative comparisons with state-of-the-art methods on the PoseTrack2018 validation set.

Method	Backbone	Additional Training Data	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
Dataset: **PoseTrack2018 Validation** set.										
TML++ [20]	OpenPose	COCO	-	-	-	60.2	-	-	56.9	67.8
STAF [35]	VGG	-	-	-	-	64.7	-	-	62.0	70.4
AlphaPose [15]	8-stage Stacked Hourglass	COCO	63.9	78.7	77.4	71.0	73.7	73.0	69.7	71.9
MDPN [18]	SimpleBaseline	MPII Pose + COCO	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 [56]	HRNet-W48	COCO	80.6	84.5	80.6	74.4	75.0	76.7	71.9	77.9
PoseWarper [4]	HRNet-W48	COCO	79.9	86.3	**82.4**	77.5	79.8	78.8	73.2	79.7
GLPose (Ours)	HRNet-W48	COCO	**84.0**	**86.9**	**82.4**	**77.6**	**80.4**	**79.3**	**73.8**	**80.8**

Fig. 3. Visualization of the results of our GLPose on PoseTrack2017 and PoseTrack2018 datasets. Various challenging dynamic scenes are involved: multiple persons, severe pose occlusions, and fast motion.

for parameter updates. The basic learning rate is set to $1e-3$, which is reduced to $1e-4$, $1e-5$, and $1e-6$ at the 6^{th}, 12^{th}, and 18^{th} epochs, respectively. The training process is terminated at 20 epochs. We train our model on 4 Nvidia GeForce 2080Ti GPUs. We adopt the identical settings for both PoseTrack2017 and PoseTrack2018.

4.2 Comparison with State-of-the-art Approaches

Results on the PoseTrack2017 Dataset We first evaluate our model on PoseTrack2017 validation set with the widely adopted average precision (**AP**) protocol. Quantitative results including the APs for each joint as well as the mAP for all joints are reported in Table 1. We benchmark our GLPose network against 11 current state-of-the-art methods, PoseTracker [17], PoseFlow [54], JointFlow [13], FastPose [58], SimpleBaseline (ResNet-101, ResNet-152) [52], STEmbedding [22], HRNet [40], MDPN [18], Dynamic [56], and PoseWarper [4]. Surprisingly, our model achieves a remarkable 83.1 mAP on the validation set, consistently outperforming existing methods. The performance improvement in challenging joints is also encouraging: we obtain an mAP of 74.2 for the ankle and an mAP of 78.1 for the wrist. We also present some visual results of our method in Fig. 3, which demonstrate the effectiveness of our method on degraded frames within dynamic environments.

Results on the PoseTrack2018 Dataset We then evaluate our model on the PoseTrack2018 dataset, and tabulate the AP results of validation set in Table 2. As shown in this table, the proposed GLPose achieves the best performance, delivering a 1.1 mAP boost over the previous state-of-the-art method [4]. We reach a final average accuracy of 80.8 mAP, and obtain an accuracy of 77.6 for the wrist and an accuracy of 73.8 for the ankle. Some visual results are depicted in Fig. 3.

Results on the Sub-JHMDB Dataset To further evaluate the proposed method, we compare GLPose with existing methods in the Sub-JHMDB dataset. The results on the test set are tabulated in Table 3. As shown in the table, the current state-of-the-art method MotionAdaptive has achieved an impressive accuracy of 94.7 mAP, while our model is able to achieve the best performance of 95.1 mAP. We also attain a 98.9 mAP for the head joint and a 98.1 mAP for the shoulder joint. The visual results are provided in Fig. 4.

Visual Comparison with Existing Methods To verify the generalization of our approach in dynamic environments, we illustrate in Fig. 5 the side-by-side comparisons of the proposed GLPose with state-of-the-art approaches. Each column shows different challenging scenarios including *rapid motion*, *pose occlusions*, *self-occlusions*, and *video defocus*, whereas each row gives the pose estimates of different approaches. We can observe that our method yields more robust and accurate results on such challenging situations. HRNet-W48 [40] is designed for static images and inherently has difficulties in capturing temporal nearby context or global semantic information, leading to suboptimal results. On the other hand, PoseWarper [4] performs spatiotemporal pose aggregation at the inference stage but only leveraging temporal information from nearby frames, ignoring global semantic cues. Our GLPose jointly incorporates global semantically similar information and local temporal visual context, establishing new state-of-the-arts for multi-frame human pose estimation.

Table 3. Quantitative comparisons with state-of-the-art methods on the Sub-JHMDB test set.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Avg**
Dataset: **Sub-JHMDB Test** set.								
Part Models [33]	79.0	60.3	28.7	16.0	74.8	59.2	49.3	52.5
Joint Action [53]	83.3	63.5	33.8	21.6	76.3	62.7	53.1	55.7
Pose-Action [21]	90.3	76.9	59.3	55.0	85.9	76.4	73.0	73.8
CPM [50]	98.4	94.7	85.5	81.7	97.9	94.9	90.3	91.9
Thin-slicing Net [38]	97.1	95.7	87.5	81.6	98.0	92.7	89.8	92.1
LSTM PM [30]	98.2	96.5	89.6	86.0	98.7	95.6	90.0	93.6
DKD(ResNet-50) [32]	98.3	96.6	90.4	87.1	99.1	96.0	92.9	94.0
K-FPN(ResNet-18) [60]	94.7	96.3	95.2	90.2	96.4	95.5	93.2	94.5
K-FPN(ResNet-50) [60]	95.1	96.4	95.3	91.3	96.3	95.6	92.6	94.7
MotionAdaptive [14]	98.2	97.4	91.7	85.2	99.2	96.7	92.2	94.7
GLPose-Split 1 (Ours)	99.1	98.5	95.0	92.2	99.0	90.0	93.2	95.5
GLPose-Split 2 (Ours)	98.7	97.5	91.1	87.8	98.4	90.6	93.8	94.3
GLPose-Split 3 (Ours)	99.0	98.3	94.7	92.4	98.9	88.7	94.8	95.6
GLPose (Ours)	98.9	98.1	93.6	90.8	98.7	89.8	93.9	**95.1**

Fig. 4. Visualization of the results of our GLPose on the Sub-JHMDB dataset.

Fig. 5. Visual comparisons of c) our GLPose method against a) HRNet-W48 [40] and b) PoseWarper [4] on the challenging dynamic environments. Each column from left to right denotes the Pose Occlusions, Rapid-Motion, Self-Occlusions, and Video Defocus. Inaccurate results are highlighted with the red solid circles.

4.3 Ablation Experiments

We perform extensive ablation experiments focused on investigating the influence of each component in the proposed GLPose framework. The first key component of our GLPose is the Feature Processing module that incorporates global semantic information and local visual context, and we mainly examine the impact of the number of local frames and non-local frames within this module. The second component is the Feature Enhancement module that mines discriminative cues from the aggregated representation to enhance the features of key frame. We would like to point out that we ablate each component to evaluate its contribution to the complete network.

Feature Processing Module We first investigate the effects of using different number of temporally nearby frames (*i.e.*, local frames) for feature aggregation. The estimation results are presented in Table 4. We experiment with five different settings: $k = 0$, $k = 1$, $k = 2$, $k = 3$, and the default setting $k = 4$. From the Table 4, we observe that the gradual improvement of mAP with increasing number of local frames, from 82.2 $\rightarrow$ 82.6 $\rightarrow$ 82.7 $\rightarrow$ 83.0 $\rightarrow$ 83.1. This is in line with our intuition, (i) temporally nearby frames provide more credible contextual information, which ensures the *stability* of feature aggregation; (ii) more local frames provide richer temporal visual cues, delivering robust human representations for pose estimation.

Table 4. Ablation experiments of different components in our GLPose. "r/m X" refers to removing X module in the network. We focus on two aspects: Feature Processing module and Feature Enhancement module. The complete GLPose consistently achieves the best results which are highlighted. Note that all ablation experiments are performed on the PoseTrack2017 dataset.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
GLPose, complete, $k = 4$**,** $n = 6$	88.1	88.9	84.1	78.1	83.5	81.5	74.2	**83.1**
Ablation Analysis of the **Feature Processing module**								
r/m local frames $k = 0$	87.6	88.3	83.3	77.1	82.9	80.5	73.2	82.2
$k = 1$, a total of 2 local frames	87.7	88.5	83.4	77.4	83.3	81.0	73.8	82.6
$k = 2$, a total of 4 local frames	87.8	88.8	83.4	77.3	83.3	80.9	73.8	82.7
$k = 3$, a total of 6 local frames	88.1	88.9	83.8	77.8	83.5	81.4	74.1	83.0
$k = 4$, a total of 8 local frames	88.1	88.9	84.1	78.1	83.5	81.5	74.2	**83.1**
r/m non-local frames, $n = 0$	87.6	88.6	83.9	76.7	83.3	80.8	73.9	82.5
$n = 1$	87.8	88.6	83.6	77.6	83.4	81.2	74.0	82.7
$n = 2$	88.0	88.7	83.7	77.7	83.4	81.3	73.6	82.7
$n = 3$	87.7	88.9	83.7	77.8	83.3	81.3	74.0	82.8
$n = 4$	88.0	88.9	83.9	78.0	83.5	81.5	74.0	83.0
$n = 5$	88.0	88.9	83.8	78.0	83.4	81.6	74.2	83.0
$n = 6$	88.1	88.9	84.1	78.1	83.5	81.5	74.2	**83.1**
Ablation Analysis of the **Feature Enhancement module**								
r/m Feature Enhancement	87.8	88.6	83.5	77.4	83.3	81.0	73.8	82.6
MLP (Detection Head)	87.8	88.8	83.7	77.6	83.4	81.0	73.9	82.7
3x3 Conv (Detection Head)	88.1	88.9	84.1	78.1	83.5	81.5	74.2	**83.1**
5x5 Conv (Detection Head)	88.0	88.9	83.8	78.0	83.6	81.3	74.0	82.9

We then study the contribution of the non-local frames F_g. We experiment with decreasing the number of non-local frames, where n is set to 6, 5, 4, 3, 2, 1, and 0. The results in Table 4 reflect a performance reduction with a decrease in n, and the mAP diminishes from 83.1 for $n = 6$ to 83.0, 83.0, 82.8, 82.7, 82.7, 82.5 at $n = 5, 4, 3, 2, 1, 0$, respectively. We also observe that incorporating non-local frames could improve accuracy by 0.6 mAP. This is not contrary to our expectation, *i.e.*, by aggregating semantically similar information (*e.g.*, analogical action

information), the Feature Processing module is able to access more available cues that are favorable for the pose estimation of the key frame.

Feature Enhancement Module We also explore the contribution of the proposed Feature Enhancement module. In the empirical study of this module, we explore the influences of the feature enhancement and the detection head. (1) For the feature enhancement, we remove complementary enhancement for the keyframe features from this module, and obtain the final pose estimates directly using the detection head (3×3 convolution layer by default): $\mathcal{S}_i \xrightarrow[\text{convolution}]{3\times3} \mathbf{H}_i$. As shown in Table 4, the mAP result falls from 83.1 to 82.6. This significant performance drop upon the *removal of the Feature Enhancement module* highlights the important role of this component in providing discriminative information for the key frame. Through the mining of complementary cues from aggregated representation $\mathcal{S}_i$ to effectively enhanced keyframe features, we obtain more spatially accurate pose heatmaps. (2) For the detection head, we investigate its network structure and experiment with three settings: Multilayer Perceptron (MLP), 3×3 convolution, and 5×5 convolution. Specifically, we adopt a channel-wise MLP which takes the features $\mathcal{S}_i$ as input to output the keypoint heatmaps $\mathbf{H}_i$, and obtain the accuracy of 82.7 mAP. In contrast, GLPose achieves an 83.1 mAP with 3×3 convolution and an 82.9 mAP with 5×5 convolution. We conjecture that adopting a medium receptive field in the detection head is more suitable for the conversion of the enhanced features to the final heatmaps. Therefore, we implement the detection head by using the 3×3 convolution.

5 CONCLUSION

In this paper, we present a global-local enhanced pose estimation network for multi-frame human pose estimation. We design a Feature Processing module that aggregates global semantically similar information and local temporal visual context to yield more comprehensive and more robust representations. Our Feature Enhancement module further mine discriminative cues from the aggregated representation to enhance the features of key frame, endowing it with the ability to cope with challenging dynamic environments. Extensive experiments show that our GLPose achieves state-the-art performance on PoseTrack2017, PoseTrack2018 and Sub-JHMDB datasets.

REFERENCES

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5167–5176.

[2] Bruno Artacho and Andreas Savakis. 2020. UniPose: Unified Human Pose Estimation in Single Images and Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7035–7044.

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

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

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. 2016. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 4733–4742.

[7] Shuning Chang, Li Yuan, Xuecheng Nie, Ziyuan Huang, Yichen Zhou, Yupeng Chen, Jiashi Feng, and Shuicheng Yan. 2020. Towards Accurate Human Pose Estimation in Videos of Crowded Scenes. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4630–4634.

[8] James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman. 2016. Personalizing human video pose estimation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 3063–3072.

[9] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. 2020. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10337–10346.

[10] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. 2020. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5386–5395.

[11] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. 2017. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1831–1840.

[12] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool. 2013. Human pose estimation using body parts dependent joint regressors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3041–3048.

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

[14] Zhipeng Fan, Jun Liu, and Yao Wang. 2021. Motion Adaptive Pose Estimation From Compressed Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11719–11728.

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

[16] Zan Gao, Yuxiang Shao, Weili Guan, Meng Liu, Zhiyong Cheng, and Shengyong Chen. 2021. A Novel Patch Convolutional Neural Network for View-based 3D Model Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2699–2707.

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

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

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 770–778.

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

[21] Umar Iqbal, Martin Garbade, and Juergen Gall. 2017. Pose for action-action for pose. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 438–445.

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

[23] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. 2021. Human Pose Regression with Residual Log-likelihood Estimation. *arXiv preprint arXiv:2107.11291* (2021).

[24] Kyaw Zaw Lin, Weipeng Xu, Qianru Sun, Christian Theobalt, and Tat-Seng Chua. 2018. Learning a Disentangled Embedding for Monocular 3D Shape Retrieval and Pose Estimation. *arXiv preprint arXiv:1812.09899* (2018).

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

[26] Meng Liu, Leigang Qu, Liqiang Nie, Maofu Liu, Lingyu Duan, and Baoquan Chen. 2020. Iterative local-global collaboration learning towards one-shot video person re-identification. *IEEE Transactions on Image Processing* 29 (2020), 9360–9372.

[27] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM International Conference on Multimedia*. 843–851.

[28] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. 2021. Deep Dual Consecutive Network for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 525–534.

[29] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Shouling Ji, Qi Liu, Shijian Lu, and Li Cheng. 2021. Investigating Pose Representations and Motion Contexts Modeling for 3D Motion Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2021), 1–16. https://doi.org/10.1109/TPAMI.2021.3139918

[30] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. 2018. Lstm pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5207–5215.

[31] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*. Springer, 483–499.

[32] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. 2019. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6942–6950.
[33] Dennis Park and Deva Ramanan. 2011. N-best maximal decoders for part models. In *2011 International Conference on Computer Vision*. IEEE, 2627–2634.
[34] Tomas Pfister, James Charles, and Andrew Zisserman. 2015. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 1913–1921.
[35] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. 2019. 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*. 4620–4628.
[36] Benjamin Sapp, Alexander Toshev, and Ben Taskar. 2010. Cascaded models for articulated pose estimation. In *European Conference on Computer Vision*. Springer, 406–420.
[37] Luca Schmidtke, Athanasios Vlontzos, Simon Ellershaw, Anna Lukens, Tomoki Arichi, and Bernhard Kainz. 2021. Unsupervised Human Pose Estimation through Transforming Shape Templates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2484–2494.
[38] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. 2017. Thin-slicing network: A deep structured model for pose estimation in videos. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 4220–4229.
[39] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. 2019. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5674–5682.
[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5693–5703.
[41] Min Sun, Pushmeet Kohli, and Jamie Shotton. 2012. Conditional regression forests for human pose estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3394–3401.
[42] Yi Tan, Yanbin Hao, Xiangnan He, Yinwei Wei, and Xun Yang. 2021. Selective Dependency Aggregation for Action Classification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 592–601.
[43] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6619–6628.
[44] Alexander Toshev and Christian Szegedy. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[45] Ali Varamesh and Tinne Tuytelaars. 2020. Mixture Dense Regression for Object Detection and Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13086–13095.
[46] Fang Wang and Yi Li. 2013. Beyond physical connections: Tree models in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 596–603.
[47] Manchen Wang, Joseph Tighe, and Davide Modolo. 2020. Combining detection and tracking for human pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11088–11096.
[48] Yang Wang and Greg Mori. 2008. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *European Conference on Computer Vision*. Springer, 710–724.
[49] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[50] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 4724–4732.
[51] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. 2019. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9217–9225.
[52] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 466–481.
[53] Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. 2015. Joint action recognition and pose estimation from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1293–1301.
[54] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. 2018. Pose flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977* (2018).
[55] Xun Yang, Meng Wang, and Dacheng Tao. 2017. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing* 27, 2 (2017), 791–805.
[56] Yiding Yang, Zhou Ren, Haoxiang Li, Chunluan Zhou, Xinchao Wang, and Gang Hua. 2021. 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*. 8074–8084.
[57] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. 2020. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7093–7102.

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

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

[60] Yuexi Zhang, Yin Wang, Octavia Camps, and Mario Sznaier. 2020. Key Frame Proposal Network for Efficient Pose Estimation in Videos. In *European Conference on Computer Vision*. Springer, 609–625.