> TIP-26494-2021 <

Multi-person Pose Tracking with Sparse Key-point Flow Estimation and Hierarchical Graph Distance Minimization

Yalong Jiang, *Member, IEEE*, Wenrui Ding, Hongguang Li and Zheru Chi, *Member, IEEE*

***Abstract*—In this paper, we propose a novel framework for multi-person pose estimation and tracking on challenging scenarios. In view of occlusions and motion blurs which hinder the performance of pose tracking, we proposed to model humans as graphs and perform pose estimation and tracking by concentrating on the visible parts of human bodies which are informative about complete skeletons under incomplete observations. Specifically, the proposed framework involves three parts: (i) A Sparse Key-point Flow Estimating Module (SKFEM) and a Hierarchical Graph Distance Minimizing Module (HGMM) for estimating pixel-level and human-level motion, respectively; (ii) Pixel-level appearance consistency and human-level structural consistency are combined in measuring the visibility scores of body joints. The scores guide the pose estimator to predict complete skeletons by observing high-visibility parts, under the assumption that visible and invisible parts are inherently correlated in human part graphs. The pose estimator is iteratively fine-tuned to achieve this capability; (iii) Multiple historical frames are combined to benefit tracking which is implemented using HGMM. The proposed approach not only achieves state-of-the-art performance on PoseTrack datasets but also contributes to significant improvements in other tasks such as human-related anomaly detection.**

***Index Terms*—Sparse Key-point Flow Estimating Module, Hierarchical Graph Distance Minimizing Module, Framework for Pose Estimation and Tracking, Anomaly Detection.**

I. INTRODUCTION

Multi-person pose tracking involves detecting the body joints of pedestrians and linking them over time by assigning consistent instance IDs. The advancements in large-scale datasets [1] [2] as well as deep learning models [3] [4] [5] [6] [7] [8] [9] [10] jointly contribute to the remarkable performance on both tasks. Existing methods are divided into two types [11]: top-down approaches which predict bounding boxes before detecting body joints [12], and bottom-up approaches [13] which localize independent body joints before grouping them into people.

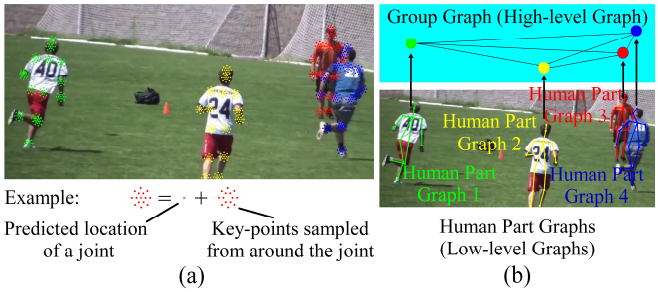

Fig. 1. Demonstration of the proposed hierarchical graph structures. (a) An input image with sampled key-points. (b) Hierarchical graph representations of people. Each node in a human part graph corresponds to one body joint. The feature representation of a body joint is obtained by sampling key-points from around the joint and concatenating the multi-level feature vectors of key-points, the locations for sampling are shown by the circles surrounding crosses in (a). Each node in a group graph denotes one person.

Top-down approaches take advantage of object detectors and obtain pose estimations within predicted bounding boxes. However, these approaches suffer from two disadvantages. Firstly, multiple entangled people are usually regarded as one by person detectors [14] [15] and top-down methods cannot deal with missed detections. Secondly, top-down approaches cannot perform well in videos due to some atypical types of challenges such as view-point variation and motion blur which occasionally lead to the failures of detectors. Even if positive results have been achieved by some bottom-up approaches [3] [16] in these scenarios, incomplete observations lead to the uncertainties regarding the identities of predicted joints.

Although temporal propagation [12] is leveraged for refining joint locations, multiple consecutive frames are all influenced because visual degradations last for long periods. As a matter of fact, pose tracking is still unsolved because of incomplete observations. In this paper we propose a novel approach for visibility-aware pose estimation and association. SKFEM and HGMM are proposed, leveraging two properties of human motion: region-level appearance consistency and human-level structural consistency both of which contribute to the measurement of visibility scores. SKFEM samples key-points around body joints and estimates key-points' motion, avoiding the redundancy produced by dense sampling. HGMM represents human bodies with graphs, as is shown by Fig. 1. The matching of graphs from consecutive frames produces visibility scores which are higher for the joints with

Manuscript received November 21, 2021, revised on April 19, 2024. This work is supported by the National Natural Science Foundation of China under Grant 62301020 and the Beijing Natural Science Foundation under grant 4234085. (*Corresponding authors: Yalong Jiang*)

Yalong Jiang, Wenrui Ding and Hongguang Li are with the Unmanned System Research Institute, Beihang University, Beijing 100191, China (e-mail: allenyljiang@outlook.com; ding@buaa.edu.cn; lihongguang@buaa.edu.cn).

Zheru Chi was with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. (e-mail: enzheru@polyu.edu.hk).

higher consistency. Visibility score benefits both pose estimation and tracking. Firstly, it facilitates the exclusion of occluded and ambiguous regions which are less informative about poses, facilitating the inference of complete poses with inter-joint graphical correlations. Secondly, by focusing the measurement of inter-human similarity on visible parts, more reliable tracking can be achieved. Besides, the propagation of predictions across frames contributes to refined joint detections which function in fine-tuning pose estimators.

Specifically, the representations of human part graphs involve both node appearances and the structures for organizing nodes. Each node in a human part graph represents one joint while each node in a group graph corresponds to one person. The visibility scores of nodes are optimized to exclude low-visibility nodes from graph similarity measurement, maintaining the consistencies in node appearances and graph structures across moments, as is shown in Section III-C and D. The advantages of the proposed approach come in three ways: (1) The proposed SKFEM facilitates the association of body joints from a local perspective, it associates human regions based on both appearance consistency and the heuristic that body joint locations do not undergo dramatic variations across a few intervals. It is complementary to HGMM module which associates entire graphs from a global perspective. The regions which show both appearance consistency and structural consistency are assigned high visibility scores because they can provide stable clues for pose estimation and tracking. (2) The pose estimator learns to infer complete human skeletons only with the observations on high-visibility parts, inherently leveraging inter-joint correlations. This is especially the case when multiple alike people are in close proximity, ambiguous and occluded regions have low visibility scores. (3) Inspired by [12], the predicted graphs, which are associated by SKFEM and HGMM, are propagated for refining joint detections.

Besides, analysis is also conducted on the reasons behind poor tracking performance. Human objects undergo frequent variations across time and each frame only captures a partial description of people. As a result, it is challenging to match people in the current frame with one historical frame correctly. In this paper we propose to maintain a list of historical frames with complementary descriptions of identities to achieve more robust matching. This scheme improves performance and the proposed framework outperforms current state-of-the-art methods on PoseTrack 2017, 2018 and 2021 datasets. Ablation studies are also conducted to demonstrate the effectiveness of each component. Additionally, the proposed framework also contributes to improvements in down-stream tasks such as anomaly detection, as will be shown in experiments. The contributions of this paper are summarized as follows: (1) SKFEM and HGMM are combined for associating objects and determining the visibility score of each joint. (2) Visibility-aware pose estimation is performed and the motion vectors from SKFEM function in propagating joint heatmaps across frames, refining pose estimations and fine-tuning pose estimators. (3) The combination of multiple historical frames contributes to better tracking performance.

The rest of the paper is organized as follows. Section II introduces related work. Methodology is discussed in Section III. Section IV shows the details of implementation as well as experimental results. Section V provides the conclusion.

II. Related Work

A. Human Pose Estimation and Tracking

Models for human pose estimation have achieved remarkable performance on datasets such as COCO Key-points Challenge [2]. Existing models can be classified into top-down [17] [13] [6] [4] and bottom-up approaches [3] [10]. The former predicts bounding boxes before conducting pose estimation. However, detectors usually fail on hard scenarios with highly occluded or entangled people. Even when the occluded people are partially detected, the predicted boxes are far from accurate and detection accuracy significantly drops. For instance, methods such as 3dhrnet [18] run 3d convolutional networks on each bounding box. Bottom-up methods predict body joints before assembling them into humans. The computational cost and inference time of bottom-up methods are nearly invariant to the number of people in an image. However, the limbs from different people are easily mixed due to entanglement, as will be shown in Fig. 6. Pose variations increase the difficulties in combining joints. To filter out invalid estimations, [19] is proposed to determine the validity of pose estimations.

Pose tracking [20] [21] differs from common object tracking [22] [23] in that body joints are not independent from each other but belong to structured human bodies. Existing methods for pose tracking detect joints in each frame or tracklet and then link the predictions over time [24] [14] [25] [26] [27]. Bottom-up tracking methods [28] [29] [16] construct graphs by connecting spatially and temporally correlated joints. Top-down approaches [14] [12] detect bounding boxes in frames or video clips before linking temporally correlated boxes [30]. However, occlusions easily lead to the missing of identities. Some approaches fixed the mistakes by evaluating temporal similarity using optical flow estimations [9] but brought huge computational burdens [31].

Current state-of-the-art methods [20] [18] [12] [26] address the challenges by combining the features from multiple instants. Although remarkable performance is achieved, the missing of visual features and residence of out-of-domain postures last for long periods, hindering the effectiveness of temporal feature fusion. To address the challenges, we introduce visibility-aware pose estimation. The joints with temporally consistent appearances and structures are assigned high visibility scores. By training pose estimators to focus on high-visibility parts, more accurate pose estimations can be achieved.

B. Multi-Object Tracking

Multi-Object Tracking (MOT) is the task of detecting objects from video frames and maintaining consistent instance IDs across frames [22] [32] [33]. The first type of methods for associating objects across frames are 2D approaches, such as

measuring the similarity in visual features [34], network flow [35], multiple hypothesis tracking [36] and conditional random field (CRF) [37]. The second type of methods combine 2D and 3D associations, a typical example is [23] [38] which takes advantage of 2D feature representations in creating tracklets before leveraging 3D feature representations to merge short tracklets into long sequences. The third type of methods for data association are based on 3D representations [39], scene flow vectors are averaged to describe the motion of 3D bounding boxes. Other relevant methods estimate the transformation matrices of 3D objects between frames [40] for aligning point clouds. Besides detection and association, segmentation masks have also been explored to conduct pixel-level matching and tracking [41] [42] [43] [44]. Segmentation-guided approaches such as [45] [46] [47] [48] have achieved remarkable performance.

However, pedestrians are different from typical rigid objects, the relative positions between body joints change with gestures. As a result, the tracking of humans can be regarded as tracking multiple correlated objects. In this paper we propose to represent multiple people in an image using graphs and match identities from consecutive frames by computing the similarities between graphs, the measurement of human graph similarities is based on body joint appearances and inter-joint connections.

C. Structured Data and Graph Neural Networks

Human body and human groups are highly structured data and can be encoded with graphs. Human body is an ensemble of nodes each of which corresponds to a semantic body joint [49]. Similarly, each person is also represented by a node in a high-level group graph. Edges denoting inter-node relations [50] [51] represent another piece of information in composing graphs. Tracking can be formulated as matching the graphs from consecutive frames. The similarity in human part graphs denotes the consistency in both human appearance and body structure while the similarity in group graphs denotes the consistency in humans' appearances and relative positions.

Graph Convolutional Networks (GCNs) [52] [53] extends classical CNNs to non-Euclidean data while maintaining basic convolutional operations. GCNs are able to deal with any type of graph data from a spatial perspective or a spectral perspective [52]. The convolutional operations in GCNs are implemented by aggregating the features of each target node with the attributes of its neighboring nodes using kernel weights. GCNs find a wide spectrum of applications in modeling structural data [54]. A great deal of studies on computer vision have taken advantage of GCNs in modeling the spatial relationships among objects or temporal relations in sequences [55]. A representative application of GCN in modeling human dynamics and actions is the ST-GCNN [55] which represents human joints as nodes and explored the spatial and temporal edges in generating embeddings and describing human actions. Different from existing approaches such as [56], we build hierarchical graphs, including human part graphs and group graphs, for pose estimation and tracking. The advantages over only using human part graphs will be demonstrated in experiments. Besides, the proposed approach for measuring inter-graph distances does not require training.

Furthermore, the number of visible nodes in each identity varies across frames due to occlusions. However, GCNs with fixed input shapes cannot deal with the variances in input node numbers. Our proposed approach focuses graph matching on high-visibility nodes and models the variances in input node numbers using the variations in visibility scores.

D. Dense Optical Flow Estimation

Deep learning models [43] [31] [57] have achieved significant improvements in dense optical flow estimation. However, the estimations involve the motion vectors on both foreground and background regions which bring interferences. Noisy estimations of motion parameters lead to the difficulties in characterizing the motion of human joints. Besides, huge computation burden also hinders the application. In this paper we propose an efficient SKFEM for accurately estimating the motion vectors on a sparse set of foreground key-points.

III. Methodology

The proposed framework is shown in Fig. 2. The steps of pose estimation, visibility score estimation and pose refinement are shown in Fig. 2(a). Fig. 2(b) and Fig. 2(c) present more details about SKFEM and HGMM.

Firstly, a Body Joint Localizing module (BJLM) is leveraged to localize human joints. Then key-points are sampled from around body joints and are associated with SKFEM. Each key-point is one pixel. In each pair of frames, N pixels are sampled from the former frame and $2N$ from the latter. $2N$ samples cover a broader region and cater for inter-frame joint offsets. SKFEM is proposed to estimate key-point correspondences across frames and find from the latter frame a subset of samples which describe the same contents as those in the former frame. SKFEM avoids the associations between alike but spatially remote objects from adjacent frames by restricting potential pairs to be nearer than radii.

HGMM module associates the human graphs from the former frame with their counterparts from the latter frame. The appearance similarity between the $i-th$ joint in Frame k and the $j-th$ joint in Frame $k+1$ is determined by the average matching error on the subset of samples around Joint j which are associated with the counterparts around Joint i by SKFEM. Sample-based appearances are complementary to skeleton-based structures in HGMM for measuring similarity. The structural similarity between Joint i and Joint j is assessed based on whether they share the same semantic and whether their semantically connected neighbors are detected.

To reduce the influences of incomplete observations on graph matching, a visibility score is introduced for each pair of joints. The visibility score of i and j takes a non-negative value, it is normalized to $[0,1]$ and is determined by their similarities in terms of both appearances and structures. Low visibility scores correspond to the joints with significant appearance variations or temporally inconsistent connections with neighbors. The regions with low-visibility joints lack in

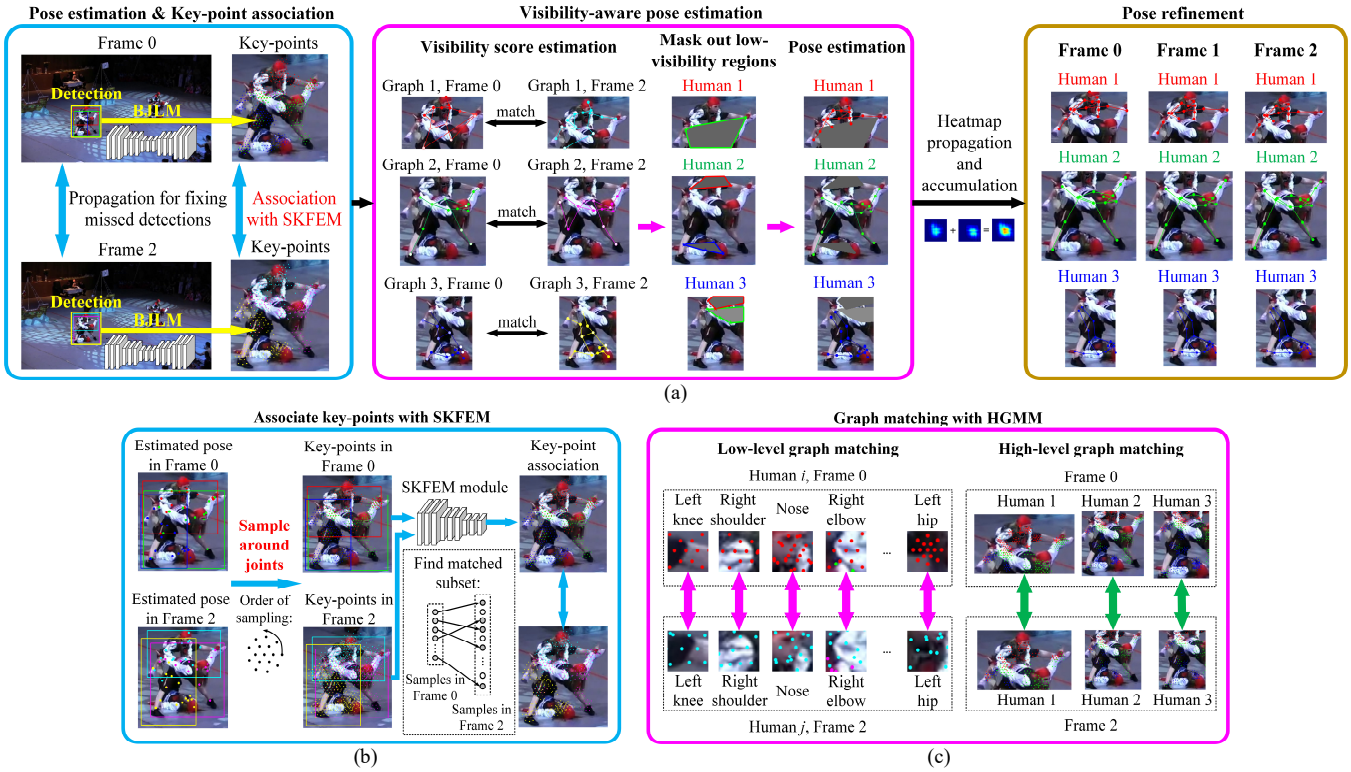

Fig. 2. The proposed framework. (a) Pose estimation and key-point association: A Body Joint Localizing Module (BJLM) localizes body joints inside bounding boxes in Frame 0 and 2. Propagation is leveraged for fixing missed detections before conducting pose estimation. Then key-points are sampled around body joints. SKFEM is proposed to estimate the movements of key-points by associating the key-points from both frames. Visibility-aware pose estimation: One human-part graph is built for each human. The graphs from Frame 0 and those from Frame 2 are matched by HGMM which predicts visibility scores. The estimated visibility score of each pair of associated joints is shown by the saturation of the circle centering at them. In each bounding box, a convex hull masks out low-visibility regions. The masks are propagated to Frame 1 where BJLM predicts skeletons by only observing high-visibility regions, only Frame 1 is masked for pose estimation. Pose refinement: The predicted heatmaps at Frame 0 and 1 are propagated to Frame 2 for accumulation and refinement. The same operations are conducted from Frame 0 and 2 to Frame 1, and from Frame 1 and 2 to Frame 0. SKFEM's predicted motion vectors are leveraged for propagation. (b) Details about sampling and key-point association. The range and number for sampling are both larger in Frame 2 than those in Frame 0 to cater for joint offsets. SKFEM associates the samples in Frame 2 with those which describe the same contexts in Frame 0. (c) Hierarchical structures for graph matching.

informative features.

Based on visibility score estimations, the proposed method trains the pose estimator to infer human skeletons by only observing high-visibility regions, the pose estimator leverages the graphical correlations between different joints. Finally the heatmaps from different frames are propagated and accumulated for pose refinement. Pose estimation and association are conducted in a sliding-window fashion with temporal stride 3, 3 frames are processed at each time.

A. Body Joint Localizing Module (BJLM)

Although existing state-of-the-art models for body joint localization are equipped with effective structures [5] [7], they still cannot exhibit robustness to complex lightening conditions or occlusions. The lack of sufficient training data leads to failures. In this section we propose to tackle this problem by training BJLM to predict the locations of low-visibility joints based on their high-visibility neighbors in the same human body. Temporal consistency is also leveraged for refining joint locations and fine-tuning BJLM. The BJLM is built upon HRNet-48 [7] which has four stages and predicts heatmaps indicating the locations of body joints. The intermediate activations from different stages are resized and concatenated with input images, providing multi-level feature representations of SKFEM module's input key-points. The training of BJLM is conducted iteratively. In each iteration, it is fine-tuned on the refined locations of joints provided by the previous iteration. Pose refinement is shown in Fig. 2(a).

B. Sparse Key-point Flow Estimating Module (SKFEM)

Before pose estimation, the bounding boxes from former and latter frames are matched based on the similarity (Intersection over Union, IoU) [9] between them, the unmatched boxes are firstly enlarged by 25% [12] along both dimensions and then propagated to the other frame to fix missed detections. This is based on the assumption that the spatial range of an individual does not undergo dramatic changes across adjacent frames. In comparison with existing models such as dense optical flow estimators which evaluate the motion vectors on all pixels most of which describe backgrounds, SKFEM samples key-points from around body joints and estimates the motion vectors of key-points. To cover potential locations of joints, the region of sampling around each joint is circular with radius determined by the

variance of the joint's heatmap in the former frame. The heatmaps are from the second-to-last layer of BJLM. The number of samples around each joint in the latter frame exceeds that in the former frame to cater for inter-frame joint offsets, as will be shown in Table X. The ranges of sampling in the latter frame exceed those in the former frame accordingly, as will be shown by the sampling strategy in Section IV-B.

The multi-level features of surrounding samples represent each joint. Denote $\mathbf{F} \in \mathbb{R}^{M \times H \times W}$ the concatenation of an input image with intermediate feature-maps along channel axis where M, H, W denote the number of feature channels, height and width of a concatenated tensor. Sampling is conducted on $\mathbf{F}$ and motion estimation is performed on all key-points simultaneously. For frames with more identities, the samples around each body joint become sparser and vice versa. The number of samples collected from all human bodies in one frame is fixed to N, $N = N_1 \times N_1$. SKFEM's structure is shown in Fig. 3(b) and Table I. SKFEM is composed of three stages: feature extraction, point correlation and flow vector generation.

TABLE I
STRUCTURE OF SKFEM

Stage	Layer Name	Stride	Radius	Micro-Layer Settings
Stage 1	Micro-Layer1	1	1	[64, 64, 128]
	Micro-Layer2	2	1	[128, 128, 256]
Stage 2	Micro-Layer3	21	10	[256, 256, 256]
	Micro-Layer4	2	2	[256, 256, 512]
	Micro-Layer5	2	4	[512, 512, 1024]
Stage 3	Micro-Layer6	0.5	4	[256, 256, 512]
	Micro-Layer7	0.5	2	[256, 256, 512]
	Micro-Layer8	0.5	1	[256, 256, 256]
	Micro-Layer9	1	1	[256, 256, 256]
	Layer10	-	-	2

Denote i the index of an input sample with $f_i \in \mathbb{R}^M$ and $x_i \in \mathbb{R}^2$ representing its features and coordinates in images. They are concatenated to produce a vector with length $(M+2)$, M represents the dimension of each input pixel's feature vector, 2 denotes 2-dimensional coordinates in images. Then the vectors are concatenated along the second axis to obtain a matrix with shape $N \times (M+2)$ which is re-organized to $N_1 \times N_1 \times (M+2)$ to maintain the relative positions between the N samples in a 2-dimensional space, as is shown in Fig. 3(a). An $N_1 \times N_1 \times (M+2)$ input tensor is the combination of samples collected from all human bodies in one frame. The two input tensors of SKFEM in Fig. 3(b) are with shapes $N_1 \times N_1 \times (M+2)$ and $2N_1 \times N_1 \times (M+2)$. In Micro-Layer1 and Micro-Layer9, the operations are shown by Eq. (1).

In an input tensor, the samples are arranged with relative positions unchanged. For instance, if Pixel i lies on the left top of j in an original image, then i still lies on the left top of j in input tensors. The detailed structure of SKFEM is shown

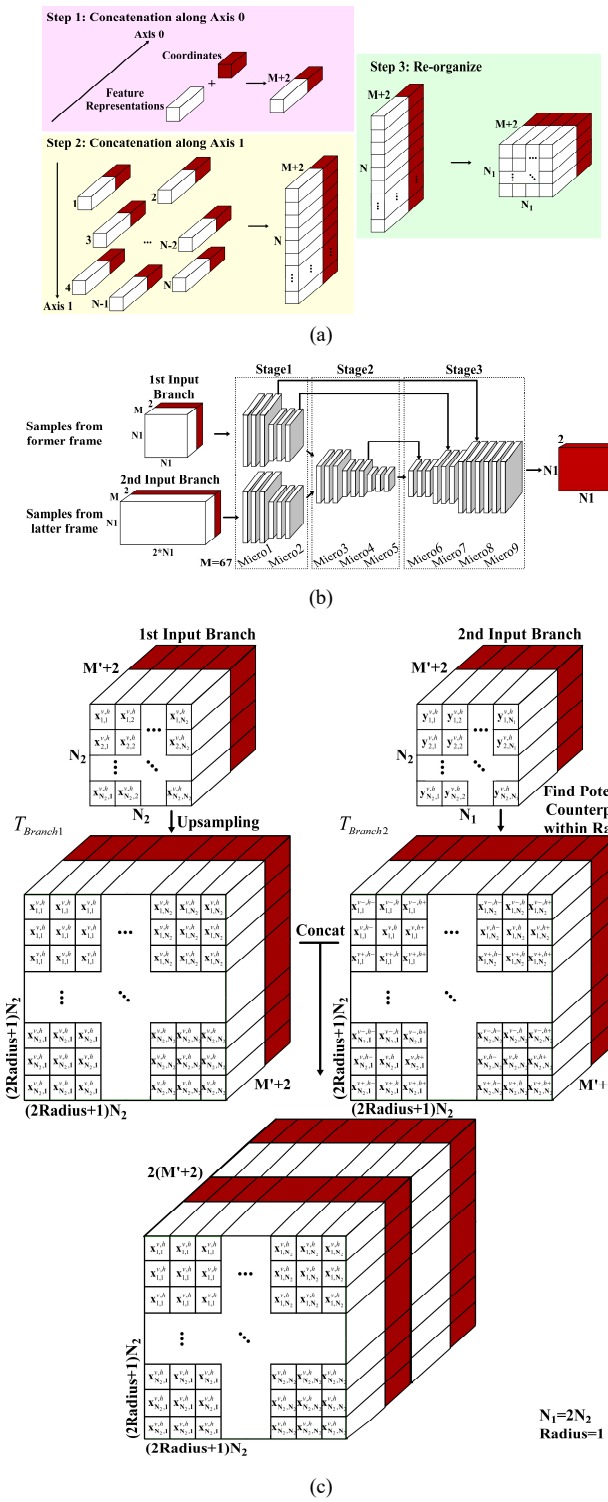

Fig. 3. (a) The structure corresponding to Eq. (1). (b) The structure of SKFEM. The two inputs represent the two groups of points which are sampled from two frames and corresponding feature-maps. (c) The structure denoted by Eq (2).

in Table I. Each micro-layer consists of three convolutional layers. The Micro-Layer Settings show the numbers of output channels of the convolutional layers. For each location j in

the input tensor, a binary mask with 0/1 elements and size $(2\times Radius+1)\times(2\times Radius+1)$ accompanies the convolutional kernel to avoid huge computational burdens brought by large kernels, a neighbor i with horizontal distance $\left|x_i^h-x_j^h\right|$ and vertical distance $\left|x_i^v-x_j^v\right|$ to j both below or equal to *Radius* has value 1 in the mask and others have value 0 which disables computations on the location during training and testing. x_i^h and x_i^v are the horizontal and vertical components of x_i. The "Stride" in Micro-Layer1 is 1 so that the number of output samples equals to that of input samples, the coordinates of output samples satisfy $x_1',...,x_l',...,x_N'=x_1,...,x_i,...,x_N$. The max-pooling operation shares both kernel size and $0/1$ masks with the convolutional layer, the maximum operation ignores the locations with mask value 0. The output of Micro-Layer 1 also concatenates $f_1',...,f_l',...,f_N'$ with $x_1',...,x_l',...,x_N'$.

$$\begin{pmatrix} f_1' \\ \vdots \\ f_l' \\ \vdots \\ f_N' \end{pmatrix} = F_{\substack{MaxPool \\ Stride=1}}\left(F_{\substack{Conv \\ kernel\_size \\ =2Radius+1}}\left(F_{Reshape}\left(F_{\substack{Concat \\ Axis1}}\begin{pmatrix} F_{\substack{Concat \\ Axis0}}(f_1,x_1), \\ \vdots \\ F_{\substack{Concat \\ Axis0}}(f_i,x_i), \\ \vdots \\ F_{\substack{Concat \\ Axis0}}(f_N,x_N) \end{pmatrix}\right)\right)\right) \quad (1)$$

In Micro-Layer2, the operations differ from Eq. (1) only in stride being 2 in horizontal and vertical directions of the first convolutional layer. For instance, $x_1',...,x_l',...,x_{N/4}'$ are sampled from $x_1,...,x_i,...,x_N$ in both axes with stride 2 in the first input branch. Micro-Layer4 and Micro-Layer5 are similar to Micro-Layer 2. In Micro-Layer3, the two input tensors are with shapes $N_2\times N_2\times(M'+2)$ and $N_2\times N_1\times(M'+2)$. $M'=256$ denotes the number of input feature channels. In the first branch, the tensor is up-sampled by factor $2\times Radius+1$ on both horizontal and vertical dimensions to produce a $(2\times Radius+1)N_2\times(2\times Radius+1)N_2\times(M'+2)$ tensor. If the coordinate y_i of an input sample of the second branch has $\left|x_i^h-y_i^h\right|$ and $\left|x_i^v-y_i^v\right|$ both below or equal to *Radius* to a sample x_i from the first branch, the two samples are potential counterparts. In the second branch, the candidate counterparts of each of the $N_2\times N_2$ samples in the first branch are ensembled in a $(2\times Radius+1)^2$ block, producing a tensor with shape $(2\times Radius+1)N_2\times(2\times Radius+1)N_2\times(M'+2)$. The two processed tensors are concatenated, obtaining a $(2\times Radius+1)N_2\times(2\times Radius+1)N_2\times 2(M'+2)$ tensor. Fig. 3(c) shows an example with $Radius=1$, the coordinates are shown in rectangles. The same mask as Micro-Layer 1 is applied. In Eq. (2) to Eq. (4), f and x denote the features and coordinates in the first branch, g and y denote those in the second branch. $y_{i,1},...,y_{i,(2Radius+1)^2}$ denote the neighboring coordinates within horizontal and vertical ranges of *Radius* to x_i. $x_{a,b}^{v,h}$ in Fig. 3(c) denotes $x_{(a-1)*N_2+b}$ with vertical and horizontal components shown by v and h, respectively. $x_{a,b}^{v-,h-},x_{a,b}^{v-,h},x_{a,b}^{v-,h+},x_{a,b}^{v,h-},x_{a,b}^{v,h},x_{a,b}^{v,h+},x_{a,b}^{v+,h-},x_{a,b}^{v+,h},x_{a,b}^{v+,h+}$ are eight spatial neighbors to pixel $x_{a,b}^{v,h}$ in images. If any of them resides in $y_{(a-1)*N_2+b,1},...,y_{(a-1)*N_2+b,(2Radius+1)^2}$, corresponding mask value is set to 1. $f_1',...,f_l',...,f_{N/4}'$ are concatenated with $x_1',...,x_l',...,x_{N/4}'$ which equal to input coordinates $x_1,...,x_i,...,x_{N/4}$ in input.

$$T_{Branch1}=F_{Upsample}\left(F_{Reshape}\left(F_{\substack{Concat \\ Axis1}}\begin{pmatrix} F_{\substack{Concat \\ Axis0}}(f_1,x_1), \\ F_{\substack{Concat \\ Axis0}}(f_i,x_i), \\ \mathrm{M} \\ F_{\substack{Concat \\ Axis0}}(f_{N/4},x_{N/4}) \end{pmatrix}\right)\right) \quad (2)$$

$$T_{Branch2}= F_{Reshape}\left(F_{\substack{Concat \\ Axis1}}\begin{pmatrix} F_{\substack{Concat \\ Axis0}}(g_{1,1},y_{1,1}),...,F_{\substack{Concat \\ Axis0}}\left(g_{1,(2Radius+1)^2},y_{1,(2Radius+1)^2}\right), \\ F_{\substack{Concat \\ Axis0}}(g_{i,1},y_{i,1}),...,F_{\substack{Concat \\ Axis0}}\left(g_{i,(2Radius+1)^2},y_{i,(2Radius+1)^2}\right), \\ \mathrm{M} \\ F_{\substack{Concat \\ Axis0}}(g_{N/4,1},y_{N/4,1}),...,F_{\substack{Concat \\ Axis0}}\left(g_{N/4,(2Radius+1)^2},y_{N/4,(2Radius+1)^2}\right) \end{pmatrix}\right) \quad (3)$$

$$\left(f_1',...,f_{N/4}'\right)^{\mathrm{T}}=F_{MaxPool,\,Stride=1}\left(F_{\substack{Conv \\ kernel\_size=2Radius+1, \\ Stride=2Radius+1}}\left(F_{\substack{Concat \\ Axis0}}\left(T_{Branch1},T_{Branch2}\right)\right)\right) \quad (4)$$

In Micro-Layer6, 7 and 8, "Stride=0.5" means that up-sampling on both axes is conducted following convolution and according to Eq. (5) where $w(x_i,x_l)$ is inversely proportional to distance. The radii in the micro-layers are 4, 2 and 1, respectively. Convolutional operations are same as (1). Layer 10 is a 1×1 convolutional layer with 2 output channels.

$$f_l'=\sum_{i\left|\left\|x_i-x_l'\right\|\le Radius\right.}w(x_i,x_l')f_i,\ i\in\{1,...,N_{input}\},\ l\in\{1,...,4N_{input}\} \quad (5)$$

N_{input} is the number of input samples.

The outputs of SKFEM, as shown by the red tensor in Fig. 3(b), describe the 2-dimensional motion vectors of the sampled key-points in the first frame which is shown by the first input branch in Fig. 3(b). Denote $X=\{x_1,x_2,...,x_N\}$ and $Y=\{y_1,y_2,...,y_{2N}\}$ as the coordinates of samples in the two input branches in Fig. 3(b). $d_i,i=1,...,N$ and $d_i^*,i=1,...,N$ are the predicted flow vectors and ground truth, respectively. $X'=\{x_1+d_1,x_2+d_2,...,x_N+d_N\}$ denote the coordinates of points propagated from first branch to the second one. The point flow estimation is conducted in a forward pass and a backward pass with loss function in Eq. (6):

$$L(U,V)=\frac{1}{N}\sum_{i=1}^{N}\left(\left\|d_i-d_i^*\right\|^2+\left\|d_i+d_i'\right\|^2\right) \quad (6)$$

where $d_i',i=1,...,N$ denote the flow vectors of the samples with coordinates X' in the latter frame, the samples are associated to those in the former frame.

The matching error between two joints is computed by averaging the matching error between their surrounding key-points which are associated by SKFEM. The joints which are not associated by SKFEM are with similarity values of 0 in graph matching. SKFEM leverages the assumption that the pose of an individual does not undergo dramatic and abrupt changes across a very few frame intervals, plausible graph matching is achieved under this assumption.

In the ground truth preparation for training SKFEM, we firstly sample key-points from around all ground truth joints in

each frame. The pairing relations between samples are consistent with the labeled identities of joints. The ground truth motion vectors on samples surrounding body joints are the samples' offsets across frames.

C. Hierarchical Graph Distance Minimizing Module (HGMM)

In this section we propose to represent the humans in images with hierarchical graphs. In a high-level graph, each node corresponds to one person and the edges connecting nodes represent the relations between identities. Low-level graphs are composed of nodes each of which represents one body joint, the edges correspond to inter-joint connections.

The high-level and low-level graphs are with similar structures. For a graph $\mathbf{G}_{level,o}$ with $N_{level,o}$ nodes ($level=0$ for low-level and $level=1$ for high-level), the feature representations involve the feature vectors of nodes $\mathbf{V}_{level,o}=\{\mathbf{v}_{level,o,1},\mathbf{v}_{level,o,2},...,\mathbf{v}_{level,o,N_{level,o}}\}$ as well as the structure representations $\mathbf{C}_{level,o}$ of edges, $\mathbf{V}_{level,o}\in\mathbf{G}_{level,o}$, $\mathbf{C}_{level,o}\in\mathbf{G}_{level,o}$. $\mathbf{C}_{level,o}(i,h);i=1,...,N_{level,o};h=1,...,N_{level,o}$ denotes the connection between the $i-th$ and $h-th$ nodes in the $o-th$ graph:

$$\mathbf{C}_{level,o}=\begin{bmatrix}\mathbf{C}_{level,o}(1,1) & \mathbf{C}_{level,o}(1,2) & \cdots & \mathbf{C}_{level,o}(1,N_{level,o})\\ \mathbf{C}_{level,o}(2,1) & \ddots & & \mathbf{C}_{level,o}(2,N_{level,o})\\ \vdots & & \ddots & \vdots\\ \mathbf{C}_{level,o}(N_{level,o},1) & \mathbf{C}_{level,o}(N_{level,o},2) & \cdots & \mathbf{C}_{level,o}(N_{level,o},N_{level,o})\end{bmatrix} \quad (7)$$

$o=1,...,O$ and O denotes the number of graphs. In a low-level graph, $\mathbf{C}_{0,o}(i,h)$ is set to 1 if the $i-th$ and $h-th$ joints are semantically connected and 0 otherwise. In a high-level graph, the value of $\mathbf{C}_{1,o}(i,h)$ is the reciprocal of the spatial distance between the centers of the $i-th$ and $h-th$ humans. The semantic meanings of nodes are demonstrated in Fig. 4(a) and Section IV-A, the table in Fig. 4(b) indicates the edges in a graph, black dots denote the semantic connections between nodes. Note that an edge $\mathbf{C}_{level,o}(i,h)$ is zero if node i or h is invisible, even though they are semantically connected.

D. Visibility-Aware Pose Estimation

In this part, visibility scores are introduced to measure the consistency in joints' appearances and their connections with neighbors across frames. As is illustrated in Fig. 2(a) where visibilities are shown by the saturation of dots, the bounding box of each human includes not only the centering human's contexts but also those from others. For instance, the human marked with red dots and the one with light blue dots describe the same identity at two instants, some joints such as shoulders and noses show consistent appearances according to SKFEM's predictions. However, even if other joints such as knees have consistent connections with neighbors, they are not associated by SKFEM due to abrupt changes in appearances and spatial locations. So knees are assigned low visibility scores.

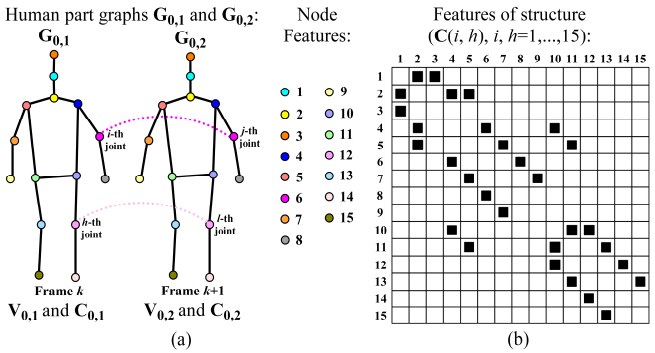

Fig. 4. Demonstration of human part graphs.

The divergences in appearances or connections are led by incomplete observations. As is shown in Fig. 2(a), the graph with red dots is partially occluded by the one with green dots. In the region where low-visibility red dots lie, green dots show higher visibility scores because the region provides informative clues for describing the green graph instead of the red one.

In the determination of visibility scores, firstly we provide an analysis on low-level graphs. $\mathbf{v}_{0,o,i}\in\mathbb{R}^{Q},i=1,...,N_o$ and $\mathbf{C}_{0,o}(i,h),i,h=1,...,N_o$ define two feature spaces (appearance feature space and structure feature space) where similarity can be measured. N_o is the number of nodes in Graph o. It is common sense that the matching of people from consecutive frames should exclude inaccurately localized body joints which may lead to the association between the joints with different semantics and the resulting increase in matching error. As a result, we apply the approach [58] for evaluating the discrepancies in structure space where each human part graph is formalized as $(\mathbf{C}_{0,o},\mathbf{p})\in\mathbb{R}^{N_o\times N_o}\times\Sigma_{N_o}$ with $\mathbf{C}_{0,o}$ encoding the connections between body joints and $\mathbf{p}$ encoding the weights of joints, $\Sigma_{N_o}=\{\mathbf{p}\in\mathbb{R}^{+}_{N_o};\sum_i p_i=1\}$. To concentrate on informative human regions, lower weights are assigned to joints with higher matching error. In appearance space, two graphs are defined as $(\mathbf{V}_{0,1},\mathbf{p})\in\mathbb{R}^{N_1\times Q}\times\Sigma_{N_1}$ and $(\mathbf{V}_{0,2},\mathbf{q})\in\mathbb{R}^{N_2\times Q}\times\Sigma_{N_2}$, each element in $\mathbf{p}$ indicates the visibility score of one node in $\mathbf{V}_{0,1}$, the same for $\mathbf{q}$ and $\mathbf{V}_{0,2}$. The discrepancy in appearance space is measured by:

$$Dist(\mathbf{V}_{0,1},\mathbf{V}_{0,2})=\min_{\mathbf{W}}\sum_j\sum_i L(\mathbf{v}_{0,1,i},\mathbf{v}_{0,2,j})w_{i,j} \quad (8)$$

where $\mathbf{W}\in\mathbb{R}^{N_1\times N_2}$ denotes the coupling between two graphs, satisfying $\mathbf{W}\mathbf{1}_{N_2}=\mathbf{p}$ and $\mathbf{W}^{\mathrm{T}}\mathbf{1}_{N_1}=\mathbf{q}$. The reason for leveraging a pair-wise occlusion matrix $\mathbf{W}$ instead of $\mathbf{p}$ and $\mathbf{q}$ for each joint in $\mathbf{G}_{0,1}$ and $\mathbf{G}_{0,2}$ individually lies in the fact that if a body joint is visible in one frame but invisible in the other, the matching error on the joint should not be included in the similarity evaluation. $L(\ )$ in Eq. (8) and Eq. (9) is introduced in Alg. 1, $L(\mathbf{v}_{0,1,i},\mathbf{v}_{0,2,j})$ is obtained by averaging the matching error on the subset of samples around Joint i which are associated with those around Joint j by SKFEM. The discrepancy between two graphs in structure feature space is measured by (9) with i,h,j,l shown in Fig. 4:

$$Dist(\mathbf{C}_{0,1},\mathbf{C}_{0,2})=\min_{\mathbf{W}}\sum_{i,j}\sum_{h,l}L(\mathbf{C}_{0,1}(i,h),\mathbf{C}_{0,2}(j,l))w_{i,j}w_{h,l} \quad (9)$$

$L(\mathbf{C}_{0,1}(i,h),\mathbf{C}_{0,2}(j,l))$ measures whether the connections between pairs of joints are consistent across the two graphs. If both joint pairs (i,j) and (h,l) each of which share the same

Algorithm 1 Obtain Visibility Scores

Input: ℓ_2 distance $L\left(\mathbf{v}_{level,1,i},\mathbf{v}_{level,2,j}\right)$ between the node features $\mathbf{v}_{level,1,i}, i\in\left[1,N_{level,1}\right]$, $\mathbf{v}_{level,2,j}, j\in\left[1,N_{level,2}\right]$ from graphs $\mathbf{G}_{level,1}$ and $\mathbf{G}_{level,2}$ where $N_{level,1}=N_{level,2}=15$, the structures $\mathbf{C}_{level,1},\mathbf{C}_{level,2}$ of $\mathbf{G}_{level,1}$ and $\mathbf{G}_{level,2}$

Output: Visibility scores in $\mathbf{W}\in\mathbb{R}^{N_{level,1}\times N_{level,2}}$

$\mathbf{W}\leftarrow$ Diagonal matrix, diagonal values being $1/N_{level,1}$

$i_{iter}\leftarrow 1$ // Index of iteration

$\langle,\rangle\leftarrow$ Matrix scalar product with Frobenius norm

while $i_{iter}\le N_{iter}$ **do**

$\mathbf{G}\leftarrow\nabla_{\mathbf{W}}Dist\left(\mathbf{G}_{level,1},\mathbf{G}_{level,2}\right)$ // See Eq. (10)

$\tilde{\mathbf{W}}\leftarrow$ Minimize $\left\langle\mathbf{G},\tilde{\mathbf{W}}\right\rangle$ with respect to $\tilde{\mathbf{W}}$ // See [71]

$c_{\mathbf{C}_{level,1},\mathbf{C}_{level,2}}(\mathbf{W})\leftarrow\mathbf{C}_{level,1}\mathbf{C}_{level,1}\mathbf{W}\mathbf{1}_{N_{level,1}}\mathbf{1}^{\mathrm{T}}_{N_{level,1}}$

$c_{\mathbf{C}_{level,1},\mathbf{C}_{level,2}}(\mathbf{W})\leftarrow c_{\mathbf{C}_{level,1},\mathbf{C}_{level,2}}(\mathbf{W})+\mathbf{1}_{N_{level,2}}\mathbf{1}^{\mathrm{T}}_{N_{level,2}}\mathbf{W}\left(\mathbf{C}_{level,2}\mathbf{C}_{level,2}\right)^{\mathrm{T}}$

$c_{\mathbf{C}_{level,1},\mathbf{C}_{level,2}}(\mathbf{W})\leftarrow c_{\mathbf{C}_{level,1},\mathbf{C}_{level,2}}(\mathbf{W})-2\mathbf{C}_{level,1}\mathbf{W}\mathbf{C}^{\mathrm{T}}_{level,2}$

$a\leftarrow -2\gamma\left\langle\mathbf{C}_{level,1}\tilde{\mathbf{W}}\mathbf{C}^{\mathrm{T}}_{level,2},\tilde{\mathbf{W}}\right\rangle$

$b\leftarrow\left\langle(1-\gamma)\sum_j\sum_i L\left(\mathbf{v}_{level,1,i},\mathbf{v}_{level,2,j}\right)w_{i,j}+\gamma c_{\mathbf{C}_{level,1},\mathbf{C}_{level,2}}(\mathbf{W}),\tilde{\mathbf{W}}\right\rangle$

$b\leftarrow b-2\gamma\left\langle\mathbf{C}_{level,1}\tilde{\mathbf{W}}\mathbf{C}^{\mathrm{T}}_{level,2},\mathbf{W}\right\rangle$

$c\leftarrow(1-\gamma)\left\langle L\left(\mathbf{v}_{level,1,i},\mathbf{v}_{level,2,j}\right)_{i,j},\mathbf{W}\right\rangle+\gamma\left\langle c_{\mathbf{C}_{level,1},\mathbf{C}_{level,2}}(\mathbf{W}),\mathbf{W}\right\rangle$

If $a>0$ **then**

$\tau\leftarrow\min\left(1,\max\left(0,-b/2a\right)\right)$

end

If $a\le 0$ **then**

If $a+b<0$ **then**

$\tau\leftarrow 1$

end

If $a+b\ge 0$ **then**

$\tau\leftarrow 0$

end

end

$\mathbf{W}\leftarrow\mathbf{W}+\tau\tilde{\mathbf{W}}$ // Line-search algorithm [71]

$i_{iter}\leftarrow i_{iter}+1$

end

return $\mathbf{W}$

semantic are detected, the structure is consistent. This is because a bone is partially invisible if one of its ending joints is not detected. $\mathbf{W}$'s entries sum to 1 and the minimization with respect to $\mathbf{W}$ in (8) places higher weights on visible nodes because they provide more meaningful and discriminative clues about joints. The discrepancy between two graphs considering both appearances and structures is expressed as (10):

$$Dist\left(\mathbf{G}_{0,1},\mathbf{G}_{0,2}\right)=(1-\gamma)Dist\left(\mathbf{V}_{0,1},\mathbf{V}_{0,2}\right)+\gamma Dist\left(\mathbf{C}_{0,1},\mathbf{C}_{0,2}\right) \quad (10)$$

where γ is the weight of structural similarity in the evaluation of graph similarity. We propose to minimize $Dist\left(\mathbf{G}_{0,1},\mathbf{G}_{0,2}\right)$ with respect to the entries in $\mathbf{W}$ according to Alg. 1.

To avoid the matching between nodes with different semantic meanings in low-level graphs, each discrepancy $L\left(\mathbf{v}_{0,1,i},\mathbf{v}_{0,2,j}\right), i\neq j$ which correspond to an off-diagonal entry in $\mathbf{W}$ is added with a penalty term which is twice the maximal discrepancy between joint pairs with same semantics. The penalty terms are not used by high-level graphs because nodes cannot be distinguished by semantics. The diagonal entries in $\mathbf{W}$ achieved with Alg. 1 are the visibility scores. The appearance discrepancy between each pair of nodes in a high-level graph is determined by the overall matching error in Eq. (10) between human-part graphs. Higher weights are assigned to visible people. Only one high-level graph resides in an image. Finally, the visibility score of each joint in a low-level graph is multiplied with the score of the corresponding human in high-level graph. The visibility scores in low-level graphs function in masking out uninformative regions, as is shown by Fig. 2(a).

Upon achieving visibility scores, BJLM learns to detect human skeletons by only observing high-visibility regions, without being influenced by the ambiguous regions with low-visibility joints. Denote the joints which are not estimated based on the current bounding box as alien joints. For instance, the red dots are native ones in the red box in Fig. 2(a), the green and blue dots in the red box are alien joints, contributing to two alien sets each of which produces a convex hull. For each alien set, we iteratively remove joints until two conditions are satisfied: (1) The average visibility score of the set exceeds the average visibility of all native joints in the bounding box. (2) The mean visibility of the native dots which lie in the convex hull is lower than the mean visibility score of all native dots in the bounding box. The remaining alien joints in each set contribute to a convex hull which is masked by filling in the mean color of all pixels in that hull.

As is shown by Fig. 2, pose estimations are conducted in Frame 0 and Frame 2 before achieving the visibility scores which produce masks on Frame 2. According to the associated key-points in Frame 0 and Frame 2, a set of pixels in Frame 1 which correspond to the associated key-points are obtained through interpolation. The key-points in Frame 2 and the interpolated pixels are leveraged in computing a homography matrix for mapping the low-visibility region masks from Frame 2 to Frame 1. Finally, pose estimations are performed on the masked Frame 1. Different from Frame 0 and 2, the postures in Frame 1 are estimated based on the learned inherently relations between high-visibility joints and low-visibility ones. In this way, BJLM is not influenced by the variations in low-visibility regions. The sum of corresponding estimations in the three frames produces more reliable results. Fig. 5 shows the usage of visibility scores.

E. Tracking and Fine-tuning of BJLM Module

High-level graphs are leveraged in tracking. Inter-node similarities are inversely proportional to inter-human matching error which is obtained in low-level graph matching. The human pairs which are not associated by SKFEM are assigned extremely large matching error. Besides, structural similarity is measured with the relative distances between humans in Eq. (7). $\mathbf{W}\in\mathbb{R}^{N_1\times N_2}$ is optimized with Alg. 1. For a frame pair with N_1 and N_2 humans, the max value in each row of $\mathbf{W}$ shows the association between the object with row index in the former frame and the one with column index in the latter.

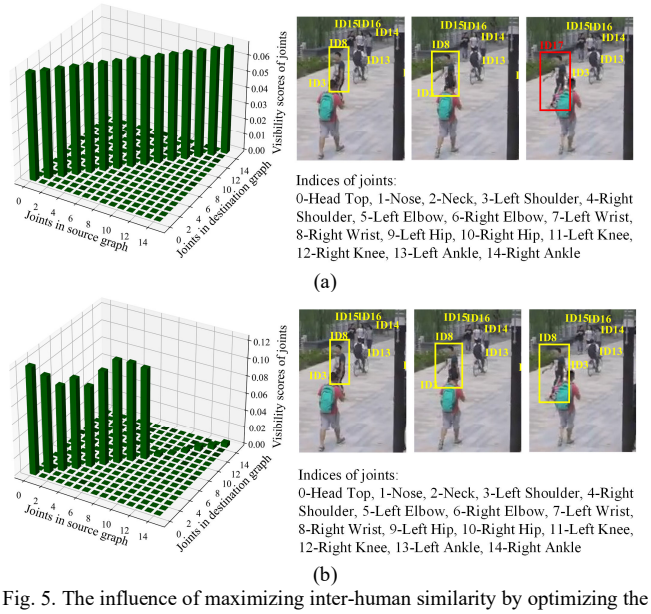

Fig. 5. The influence of maximizing inter-human similarity by optimizing the visibility scores of human joints. (a) Without optimization on visibility scores, different body joints share the same weight. The tracking of the eighth identity fails because the lower body is occluded and the computation of graph distance still involves the consideration of occluded joints; (b) Visible points which show higher similarity across frames are assigned higher weights (visibility scores) and vice versa. The occluded joints are ignored to maintain the similarity of the same identity across frames.

Algorithm 2 Fine-tune BJLM

Input: Video, pre-trained BJLM, SKFEM, HGMM
Output: Fine-tuned BJLM
$m \leftarrow 1$ // F_m denotes the $m-th$ frame in Video
$M \leftarrow$ Temporal length of Video
while $m \le M-2$ **do**
  $J_m, J_{m+2} \leftarrow BJLM(F_m), BJLM(F_{m+2})$ // J_m – Joints in F_m
  $S_m, S_{m+2} \leftarrow Sample(J_m), Sample(J_{m+2})$ // S_m – Key-points in F_m
  foreach Graph u in F_m **do**
    foreach Graph v in F_{m+2} **do**
      $L(\mathbf{v}_{0,u,i}, \mathbf{v}_{0,v,j}) \leftarrow SKFEM(S_m, S_{m+2}),\ i, j \in [1,15]$
      $\mathbf{C}_{0,u}, \mathbf{C}_{0,v} \leftarrow$ Structures of Graph u and Graph v
      $\mathbf{W} \leftarrow$ The visibility scores estimated with $L(\mathbf{v}_{0,u,i}, \mathbf{v}_{0,v,j}), \forall i, j$ and $\mathbf{C}_{0,u}, \mathbf{C}_{0,v}$ using Alg. 1
      $\tilde{F}_{m+1} \leftarrow Mask(F_{m+1}, \mathbf{W})$ // Mask out low-vis regions
      $\tilde{J}_{m+1} \leftarrow BJLM(\tilde{F}_{m+1})$ // $\tilde{J}_{m+1}$ – Joints in $\tilde{F}_{m+1}$
      $\hat{J}_m, \hat{J}_{m+1}, \hat{J}_{m+2} \leftarrow Refine(J_m, \tilde{J}_{m+1}, J_{m+2})$
    end
  end
  $m \leftarrow m+3$
End
Fine-tune BJLM with refined $\hat{J}_m, \forall m \in [1, M]$
return BJLM

Based on the associations between key-points and humans in Frame 0, 1 and 2, one homography matrix is obtained for human-level heatmap propagation between each pair of frames. Heatmaps are propagated in this way. In implementations, we propagate the heatmaps of frames with indices u_1 and u_2, $u_1, u_2 \in \{0,1,2\}, u_1, u_2 \neq s$ to Frame $s \in \{0,1,2\}$. The propagated heatmaps are added with Frame s's heatmap and the peaks in the accumulated map determine the refined skeletons, as is shown by Fig. 2(a). The refined joint coordinates are leveraged in fine-tuning BJLM iteratively, leveraging both masked Frame 1 and unmasked Frame 0 and 2. The procedures are shown by Alg. 2 where $L(\mathbf{v}_{0,u,i}, \mathbf{v}_{0,v,j})$ is obtained by averaging the matching error on the subset of sampled key-points around Joint i which are associated with those around Joint j by SKFEM. $SKFEM(\ )$ denotes the association of key-points across frames. $Mask(\ )$ denotes the masking out of low-visibility regions based on estimated visibility scores, as is illustrated in Section III-D. $BJLM(\ )$ denotes the localization of body joints using BJLM, $Sample(\ )$ is the procedure for sampling around joints. $Refine(\ )$ shows the refinement of joint locations through heatmap propagation.

In the matching of humans in Frame T with those in Frame $T+1$, additional results from frames $T-K+1,...,T-1, K \ge 1$ are also introduced. Specifically, the humans in Frame $T+1$ are associated with those in each of the historical frames for once. The resulting $\mathbf{W}$'s are averaged. The trajectories which are not associated to historical ones are taken as new ones. In the matching with humans in frames $T-K+1,...,T-1$, γ=0 because the distances between humans vary significantly across time. Historical frames are only used in tracking.

IV. Experiments

In this section, both quantitative and qualitative results are provided to show the effectiveness of the proposed approach.

A. Dataset and Evaluation Metric

The PoseTrack dataset [21] [59] [60] is a large-scale benchmark for human pose estimation and tracking in videos. It is composed of challenging videos where crowds of people perform a wide range of activities, some human objects suffer from heavy occlusions and/or are with complex poses. Experiments are conducted on both the PoseTrack 2017, 2018 and 2021 datasets. PoseTrack 2017 contains 250 videos for training, 50 videos for validation and 214 for test, PoseTrack 2018 has 593 training videos, 170 validation videos and 375 for test. PoseTrack 2021 dataset [1] shares the same videos but annotates the identities which are not annotated in PoseTrack 2018. All humans with visible heads are labeled. Each person is annotated with 15 body joints: head top, nose, neck, left and right shoulders, elbows, wrists, hips, knees and ankles, and is assigned a person id which is consistent across the entire video. The test server is closed so we only evaluate on validation sets.

The metrics for evaluation involve average precision (AP) [61] for human pose estimation and multi-object tracking accuracy (MOTA) for tracking [21] [62]. The metrics are evaluated on each joint before averaging. The computation of AP is based on Object Key-point Similarity (OKS) [2]: $OKS = \sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0) / \sum_i \delta(v_i > 0)$ where d_i is the

Euclidean distance between the $i-th$ prediction and ground truth, s denotes object scale and k_i is a constant for the $i-th$ joint. AP averages precision values at 10 thresholds of OKS from 0.50 to 0.95. Additionally, Higher Order Tracking Accuracy (HOTA) [63] is also introduced for evaluating the performance on PoseTrack 21 dataset. An additional dataset, APT-36K [64], is introduced to show the generalizability of the proposed approach. APT-36K contains 30 different animal species. It has 36,000 annotated frames with 53,006 annotated animal instances from 2,400 video clips. 17 body joints are labeled for each animal instance, including two eyes, one nose, one neck, one tail, two shoulders, elbows, knees, hips, and four paws. AP is also leveraged for evaluation based on OKS.

Besides the datasets for tracking, the proposed framework is also leveraged in video anomaly detection. The Human-related (HR) ShanghaiTech Campus dataset [65] [66] is the most comprehensive and realistic datasets for this task. It involves a wide spectrum of anomaly types. The metric for evaluation is Area Under the Curve (AUC) [65]. Due to sophistication of human motions and field-of-views, existing models struggle to get adequate performance on tracking. There are 330 training videos and 101 test videos whose abnormal events are related to humans. The performance of existing anomaly-detection models is limited by their capability in tracking. As a result, experiments will be conducted to show the improvements in anomaly detection contributed by our approach for tracking.

B. Implementation Details

We leverage YOLOv3 [67] as human detector. The BJLM for localizing human joints is the one proposed in [7] where the last layer generates output heatmaps indicating the locations of human joints. BJLM is composed of four stages each of which has parallel multi-resolution subnetworks. The "HRNet-W48" setting [7] is adopted. It is pre-trained on the COCO dataset [2]. During training, each human bounding box is resized to 384×288 and data augmentation is applied with random rotation in range $([-45°,45°])$ and random scaling in range $([0.65,1.35])$. Adam optimizer [68] is applied with initial learning rate 1e-3 which drops to 1e-4 and 1e-5 at the 170th and 200th epochs, respectively. Training ends at the 210th epoch. Upon pre-training, BJLM is trained on the PoseTrack datasets with initial learning rate 1e-4 which drops to 1e-5 at the 10th epoch and 1e-6 at the 15th epoch. Training lasts for 20 epochs. The $\mathbf{F}\in\mathbb{R}^{M\times H\times W}$ in Section III-B is the concatenation of a RGB image with the up-sampled 64- channel output from the first layer of BJLM, $M=67$, $N=1024$.

The heatmaps of joints have the shape being similar to gaussian kernels. Sampling is conducted in a circular region around each joint, with the radius of outer circle being equal to the variance of the joint's heatmap. The radii and numbers of samples on different circles compose two arithmetic progressions. The samples in each circle are distributed evenly, as is shown in Fig. 2(b). The ground truth flow vectors are generated based on the associations between ground truth body joint locations of same identities across frames. The training of SKFEM is conducted for 250 epochs with stochastic gradient descent (SGD). The initial learning rate is 1e-1 with momentum 0.9. The learning rate is lowered by a factor of 10 at the 75th, 150th and 200th epochs. Weight decay is 1e-4. The runtime of SKFEM is 52ms per frame, as compared with 125ms of the most commonly used model for dense optical flow estimation [31] on NVIDIA GTX 1080 GPU and 116ms of [57]. The HGMM for estimating visibility scores does not require annotations. Experiments will be performed to show that SKFEM contributes more to tracking than [31]. The procedure for running the overall framework has been demonstrated in Alg. 2 and Fig. 2. In Step 1, BJLM is trained on COCO and PoseTrack. The fine-tuning of BJLM is conducted for multiple iterations. In each iteration, BJLM is fine-tuned for 2 epochs. In default settings, BJLM is fine-tuned for 3 iterations, $K=5$ historical frames with temporal stride 2 between neighboring historical frames are used for tracking, as will be detailed in Section IV-E.

C. Comparisons with Existing Methods

The proposed framework is compared with current state-of-the-art methods on both key-point localization and tracking on the PoseTrack 2017 and 2018 datasets. As is shown in Table II to Table V, the proposed framework outperforms state-of-the-art methods on both tasks by leveraging SKFEM and HGMM for estimating motion vectors and visibility scores.

The improvements in pose estimation are contributed by the approach for fine-tuning BJLM, according to Alg. 2. Fig. 6 subjectively shows the proposed approach under visual degradations and resulting incomplete detections. Intersection over Union (IoU) is applied to match the initial detections in former and latter frames. A threshold of 0.5 is set to determine whether the bounding boxes can be matched because 99% of boxes from the same human have an IoU being larger than 0.5 across neighboring frames on the training set. The boxes in former and latter frames in Fig. 6 cannot be matched. Then the boxes are propagated for obtaining complete detections.

TABLE II
COMPARISON OF KEY-POINT LOCALIZATION (AP, %) ON POSETRACK 2018 VALIDATION SET

Method	Head	Sho	Elb	Wri	Hip	Kne	Ank	Avg
15 Key-points [26]	-	-	-	-	-	-	-	81.6
DCPose [12]	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
Combing Detection and Track [18]	84.9	87.4	84.8	79.2	77.6	79.7	75.3	81.5
Learning Dynamics [14]	85.1	87.7	85.3	80.0	81.1	81.6	77.2	82.7
PoseTrans [28]	88.9	90.3	87.4	81.8	83.5	85.5	80.6	85.7
FAMI-Pose [20]	85.5	87.7	84.2	79.2	81.4	81.1	74.9	82.2
Ours (BJLM finetuned for 3 iterations)	87.9	87.7	84.9	79.5	81.4	80.0	75.8	82.8
Ours (BJLM finetuned for 7 iterations)	**89.2**	**90.7**	**88.5**	**82.6**	**84.9**	**86.7**	**81.2**	**86.5**

TABLE III
COMPARISON OF JOINT TRACKING (MOTA, %) ON POSETRACK 2018 VALIDATION SET

Method	Head	Sho	Elb	Wri	Hip	Kne	Ank	Avg
15 Key-points [26]	-	-	-	-	-	-	-	66.6
Combing Detect and Track [18]	74.2	76.4	71.2	64.1	64.5	65.8	61.9	68.7
Learning Dynamics [14]	74.3	77.6	71.4	64.3	65.6	66.7	61.7	69.2
Ours without group graph	77.2	76.9	73.2	65.7	68.3	65.9	64.5	70.6
Ours	**77.4**	**77.5**	**74.2**	**66.8**	**70.1**	**66.1**	**65.2**	**71.4**

TABLE IV
COMPARISON OF KEY-POINT LOCALIZATION (AP, %) ON POSETRACK 2017 VALIDATION SET

Method	Head	Sho	Elb	Wri	Hip	Kne	Ank	Avg
STEmbedding [29]	83.8	81.6	77.1	70.0	77.4	74.5	70.8	77.0
FastPose [70]	80.0	80.3	69.5	59.1	71.4	67.5	59.4	70.3
FlowTrack [9]	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
HRNet [7]	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
DCPose [12]	88.0	88.7	84.1	78.4	83.0	81.4	74.2	82.8
Combing Detect and Track [18]	89.4	89.7	85.5	79.5	82.4	80.8	76.4	83.8
FAMI-Pose [20]	89.6	90.1	86.3	80.0	84.6	83.4	77.0	84.8
Learning Dynamics [14]	90.9	90.7	86.0	79.2	83.8	82.7	78.0	84.9
Ours	**92.9**	**91.8**	**86.8**	**82.4**	**85.3**	**82.8**	**79.8**	**86.3**

TABLE V
COMPARISON OF JOINT TRACKING (MOTA, %) ON POSETRACK 2017 VALIDATION SET

Method	Head	Sho	Elb	Wri	Hip	Kne	Ank	Avg
STEmbedding [29]	78.7	79.2	71.2	61.1	74.5	69.7	64.5	71.8
FastPose [70]	-	-	-	-	-	-	-	63.2
FlowTrack [9]	73.9	75.9	63.7	56.1	65.5	65.1	53.5	65.4
LightTrack [25]	74.1	76.2	63.8	56.3	65.7	65.3	53.4	65.6
Combing Detect and Track [18]	80.5	80.9	71.6	63.8	70.1	68.2	62.0	71.6
Learning Dynamics [14]	82.0	83.1	73.4	63.5	72.3	71.3	63.5	73.4
Ours	**83.8**	**82.9**	**73.9**	**66.5**	**72.0**	**69.3**	**64.9**	**73.7**

The performance of the proposed approach on PoseTrack 21 dataset is shown in Table VI. The configurations of CorrTrack and Tracktor are the same as [1]. If a pose from the former frame cannot be matched with later ones, the corresponding track is deactivated. As is shown in Section III-E, K historical frames are leveraged for storing the historical

TABLE VI
COMPARISON OF KEY-POINT LOCALIZATION AND JOINT TRACKING (AP AND MOTA, %) ON POSETRACK 2021

Method	mAP (Average AP)	MOTA	HOTA
CorrTrack w. ReID [24]	72.7	63.8	52.7
Tracktor w. correspondences [72]	73.6	61.6	48.9
Ours	**76.9**	**64.7**	**54.1**

TABLE VII
INFLUENCE OF CHANGING THE TRACKING APPROACH ON THE PERFORMANCE OF ANOMALY DETECTION (AUC, %)

	HR-ShanghaiTech		
Method	Reconstruction	Prediction	Rec. + Pred.
ED-RNN/G+L [65]	69.9	72.2	71.3
Learning Regularity [65]	74.4	74.5	75.4
Tracking with our proposed method	**83.4**	**79.5**	**84.2**

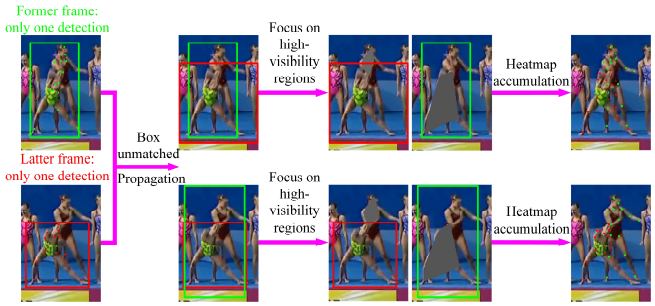

Fig. 6. An example of leveraging the proposed approach for pose estimation under incomplete detections. The procedures in Fig. 2 are applied.

appearances to deal with temporary disappearance and re-entering of humans. Upon the disappearance of some tracks, the similarity between a human in a new frame and an identity from historical frames is computed, if the similarity exceeds the average same-human similarity, the human is determined to be re-entering. Else it is started as a new track. Same-human similarity is obtained from the statistics about the similarity between different observations of the same human, it is computed on training data.

Besides pose tracking, the proposed method is also evaluated on the ShanghaiTech Campus dataset [66] for human-related anomaly detection. We replace the first part of a representative anomaly detection model [65] with the proposed tracking module and keep the second part unchanged. The comparison is demonstrated in Table VII.

The experiments on APT-36K are also conducted with HRNet [7] according to the setting in [64]. BJLM is pre-trained on COCO before training on APK-36K for 210 epochs. The initial learning rate is 5e-4 and is decreased by a factor of 10 at the 170th and 200th epochs. The results are shown in Table VIII. The tracker is SwinTrack [69], as is used by [64].

D. Ablation Study

In this section, experiments are conducted to demonstrate

TABLE VIII
PERFORMANCE OF JOINT LOCALIZATION AND TRACKING (AP, %) ON APT-36K

Approach	Joint Localization	Tracking
BJLM without fine-tuning	77.4	74.9
BJLM fine-tuned according to Alg. 2	**78.5**	**75.7**

TABLE IX
ADVANTAGE OF 3-FRAME SCHEME ON KEY-POINT LOCALIZATION (AP, %), MEASURED ON POSETRACK 2018 VALIDATION SET

METHOD	AP
Ours with only the pose estimations on masked Frame 1, for fine-tuning BJLM and testing	81.7
Ours with the pose estimations on Frame 0 and Frame 2, for fine-tuning BJLM and testing	82.5
Our pose estimations on 3 frames	**82.8**

TABLE X
CONTRIBUTION OF VISIBILITY SCORES AND THE NUMBER OF SAMPLING TO KEY-POINT LOCALIZATION (AP, %) AND JOINT TRACKING (MOTA, %), MEASURED ON POSETRACK 2018 VALIDATION SET

METHOD	AP	MOTA
Ours without visibility scores and masks	79.9	67.9
DCPose [12]	80.9	-
Ours with visibility scores replaced by joint confidences	81.2	68.3
Ours with N samples in former frame, $1.5N$ ones in latter frame	**82.1**	**71.0**
Ours with N samples in former frame, $3N$ ones in latter frame	**82.8**	**71.4**
Ours with N samples in former frame, $2N$ ones in latter frame	**82.8**	**71.4**

the contributions of different components. The influences on performance brought by iteratively fine-tuning BJLM in a feedback fashion are also shown. Firstly, we show the advantage of the 3-frame scheme for pose estimation in Fig. 2 through comparing with other variants, as is shown in Table IX. In the first row of Table IX, the pose estimations on masked Frame 1 in Fig. 2 are propagated to Frame 0 and 2 as final results, ignoring heatmap accumulation. It can be seen that the fusion of results from different frames with diverged contexts is useful. In the second row of Table IX, we conduct pose estimation on masked Frame 0 and Frame 2 after obtaining and propagating low-visibility region masks. It can be seen that conducting pose estimations on 3 frames outperforms the variant on 2 frames, because the slight differences between Frame 1 and 2 bring complementary clues.

Besides, Table X shows the contributions of visibility scores. In the first row of Table X, only heatmap accumulation is performed based on the results in unmasked Frame 0, 1 and 2. It can be seen that visibility scores and masks are useful because they facilitate the framework to focus on visible joints and ignore occluded parts, the performance of tracking can be improved in this way. It can be seen from the third row that the proposed visibility scores outperform the joint confidences from BJLM. It is also shown in Table X that sampling twice as many key-points in the latter frame is enough to cover potential joint offsets which are addressed in Section III.

TABLE XI
COMPARISON OF SKFEM WITH FLOWNET 2.0 [31] (AP AND MOTA, %) ON POSETRACK 2018 VALIDATION SET

Key-point Localization (AP, %)								
Method	Head	Sho	Elb	Wri	Hip	Kne	Ank	Avg
FlowNet 2.0 in replace of SKFEM	85.7	87.6	85.1	79.3	78.4	80.2	76.3	82.2
Ours	**87.9**	**87.7**	**84.9**	**79.5**	**81.4**	**80.0**	**75.8**	**82.8**
Joint Tracking (MOTA, %)								
Method	Head	Sho	Elb	Wri	Hip	Kne	Ank	Avg
FlowNet 2.0 in replace of SKFEM	76.6	78.6	72.9	66.1	66.5	67.8	65.9	70.9
Ours	**77.4**	**77.5**	**74.2**	**66.8**	**70.1**	**66.1**	**65.2**	**71.4**

Runtime		
Model	Flownet 2.0	SKFEM
# param. (M)	162.49	62.72
Runtime (ms)	123.89	51.76

The comparison between applying SKFEM in estimating the motion vectors of samples around body joints with that of applying FlowNet 2.0 [31] in performing dense optical flow estimation is demonstrated in Table XI and it can be inferred that SKFEM contributes more to both tasks than dense optical flow estimation. If the proposed SKFEM is replaced by FlowNet 2.0, performance drops. The same set of points are sampled from the predictions of FlowNet 2.0 as in SKFEM, Alg. 2 is conducted to produce the results in Table XI. It can be seen from Table XI that the average time consumption of SKFEM is only 42% of FlowNet2.0.

The computation of graph distances is based on Eq. (10) where lower γ values correspond to a higher weight on appearance similarity and higher γ values correspond to a higher weight on structural similarity. Table XII demonstrates the influences of γ. $\gamma_{Group\ Graph}$ and $\gamma_{Human\ Part\ Graph}$ denote the weights in group graphs and human part graphs, respectively.

In Table II and XIII the influence of the number of iterations in Alg. 2 for fine-tuning BJLM is demonstrated. Table XIII and the last two rows in Table II show that more iterations enhance performance. However, due to the limitation in time, we select the number of iterations to be 3.

E. Combining Historical Data for Better Tracking

Due to the quick changes of human appearances in videos, each frame may only contain a partial description of identities. Table XIV shows the contributions brought by introducing

Fig. 7. Subjective results demonstrating the effectiveness of the proposed framework for multi-person pose estimation and tracking. (a) Pose estimation and tracking under crowded cases. (b) Pose estimation and tracking under motion blur. (c) Failure cases under severe visual degradation, as is shown in red ellipses.

TABLE XII
INFLUENCES OF γ ON JOINT TRACKING (MOTA, %), EVALUATED ON POSETRACK 2018 VALIDATION SET

$\gamma_{Group\ Graph}$								
0.1			0.5			0.9		
$\gamma_{Human\ Part\ Graph}$			$\gamma_{Human\ Part\ Graph}$			$\gamma_{Human\ Part\ Graph}$		
0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
70.6	54.2	30.1	**71.4**	57.5	34.6	70.9	56.1	32.7

historical frames to tracking, "Number" means the number of historical frames for matching while "Stride" means the stride for sampling historical frames. It can be inferred from Table XIV that an appropriate number of frames and a proper stride contribute to improvements. Fig. 7 provides some subjective results demonstrating the effectiveness of the framework.

TABLE XIII
INFLUENCE OF THE NUMBER OF ITERATIONS FOR FINE-TUNING BJLM ON JOINT TRACKING (MOTA, %)

Number of Iterations	Head	Sho	Elb	Wri	Hip	Kne	Ank	Avg
1	75.6	75.5	72.1	64.9	68.3	64.2	63.5	69.5
3	77.4	77.5	74.2	66.8	70.1	66.1	65.2	71.4
5	77.9	77.7	73.9	67.2	70.7	66.5	65.4	71.7
7	78.0	77.8	74.2	67.6	71.1	66.8	65.5	71.9

TABLE XIV
INFLUENCE OF THE NUMBER OF HISTORICAL FRAMES AND STRIDE ON JOINT TRACKING (MOTA, %), EVALUATED ON POSETRACK 2018 VALIDATION SET

Method	Head	Sho	Elb	Wri	Hip	Kne	Ank	Avg
Number=1 Stride=1	76.1	77.8	72.4	66.2	66.5	67.3	63.9	70.4
Number=3 Stride=1	76.5	78.3	73.1	66.3	66.7	67.5	64.2	70.7
Number=5 Stride=1	76.7	78.6	73.2	66.6	66.8	67.9	64.5	71.1
Number=5 Stride=2	**77.1**	**79.4**	**73.6**	**66.9**	**67.2**	**68.5**	**64.6**	**71.4**

F. Limitations and Future Work

As is shown by the red ellipses in Fig. 7(c), when a human detection undergoes severe visual degradations, the predictions are not reliable. The predicted joints with confidence values below 0.5 are not shown.

The proposed approach differs from existing state-of-the-art ones [12] [20] in that it discovers the fact that visual degradations last for multiple moments and consecutive frames are all influenced. As a result, this approach not only combines the complementary predictions from consecutive frames but also learns to predict complete skeletons by only observing high-visibility parts. It inherently leverages the invariant relations between visible and invisible regions. In the future, more invariant properties will be explored to recover complete skeletons by only observing limited parts under severe degradations. The proposed approach can potentially be combined with fusion-based approaches such as [12] [20] to achieve better performance.

The proposed approach for representing objects by evaluating the visibility scores of different parts and leveraging partial observations in inference can also be applied in the recognition of incompletely observed graphical objects which are embedded with invariant relations between different parts. Besides, the proposed SKFEM can be applied in motion estimation because it performs sparse motion estimations and reduces the redundancy in dense flow estimations.

V. CONCLUSION

In this paper, we propose a novel approach for multi-person pose estimation and tracking in videos. SKFEM and HGMM are built for estimating human motion across consecutive frames and organizing groups of humans with hierarchical graph structures. The visibility scores of body joints are predicted based on the consistency in appearances and structures which are measured by SKFEM and HGMM. BJLM learns to estimate poses by focusing on high-visibility parts, it is fine-tuned iteratively. By assigning higher weights to visible parts in representing human appearances, the performance of pose estimation is improved. Besides, the combination of historical frames also benefits tracking. The SKFEM achieves a higher accuracy than the models for dense optical flow estimation while running at a higher speed. Sufficient experiments have been conducted to demonstrate the effectiveness of the overall approach as well as each component.

REFERENCES

[1] A. Doering, D. Chen, S. Zhang, B. Schiele and J. Gall, "PoseTrack21: A Dataset for Person Search, Multi-Object Tracking and Multi-Person Pose Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2022, pp. 20963-20972.

[2] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug, 2014, pp. 3–19.

[3] D. Shi, X. Wei, L. Li, Y. Ren and W. Tan, "End-to-End Multi-Person Pose Estimation with Transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2022, pp. 11069-11078.

[4] Y. Luo, Z. Xu, P. Liu, Y. Du and J. Guo, "Multi-Person Pose Estimation via Multi-Layer Fractal Network and Joints Kinship Pattern," *IEEE Trans. Image Process.,* vol. 28, no. 1, pp. 142-155, Jan. Jan, 2019.

[5] F. Sun, T. Kong, W. Huang, C. Tan, B. Fang and H. Liu, "Feature Pyramid Reconfiguration With Consistent Loss for Object Detection," *IEEE Trans. Image Process.,* vol. 28, no. 10, pp. 5041-5051, Oct, 2019.

[6] H. Fang, S. Xie, Y. Tai and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct, 2017, pp. 2334-2343.

[7] K. Sun, B. Xiao, D. Liu and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2019, pp. 5693-5703.

[8] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 43, no. 10, pp. 3349-3364, Oct, 2020.

[9] B. Xiao, H. Wu and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug, 2018, pp. 466-481.

[10] A. Newell, Z. Huang and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *Proc. Adv. Neural Inf. Process. Syst. (NIPS),* vol. 30, Dec, 2017,.

[11] Y. Wang, M. Li, H. Cai, W. M. Chen and S. Han, "Lite Pose: Efficient Architecture Design for 2D Human Pose Estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2022, pp. 13126-13136.

[12] Z. Liu, H. Chen, R. Feng, S. Wu, S. Ji, B. Yang and X. Wang, "Deep Dual Consecutive Network for Human Pose Estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June, 2021, pp. 525-534.

[13] N. Xue, T. Wu, G. S. Xia and L. Zhang, "Learning Local-Global Contextual Adaptation for Multi-Person Pose Estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2022, pp. 13065-13074.

[14] Y. Yang, Z. Ren, H. Li and C. Zhou, "Learning Dynamics via Graph Neural Networks for Human Pose Estimation and Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Jun. 2021, pp. 8074-8084.

[15] B. Singh, M. Najibi and L. S. Davis, "SNIPER: Efficient multi-scale training," *Proc. Adv. Neural Inf. Process. Syst. (NIPS),* vol. 31, Dec, 2018.

[16] Y. Raaj, H. Idrees, G. Hidalgo and Y. Sheikh, "Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2019, pp. 4620-4628.

[17] Z. Kan, S. Chen, C. Zhang, Y. Tang and Z. He, "Self-Correctable and Adaptable inference for generalizable human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5537-5546.

[18] M. Wang, J. Tighe and D. Modolo, "Combining detection and tracking for human pose estimation in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2020, pp. 11088-11096.

[19] Q. Liu, A. Kortylewski and A. L. Yuille, "PoseExaminer: Automated Testing of Out-of-Distribution Robustness in Human Pose and Shape Estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 672-681.

[20] Z. Liu, R. Feng, H. Chen, S. Wu, Y. Gao, Y. Gao and X. Wang, "Temporal Feature Alignment and Mutual Information Maximization for video-based human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2022, pp. 11006-11016.

[21] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall and B. Schiele, "PoseTrack: A benchmark for human pose estimation and tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2018, pp. 5167-5176.

[22] D. Stadler and J. Beyerer, "Improving Multiple Pedestrian Tracking by Track Management and Occlusion Handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10958-10967.

[23] L. Jonathon, F. Tobias and L. Ba, "Track to Reconstruct and Reconstruct to Track," in *arXiv preprint*, Sep, 2019, arXiv:1910.00130.

[24] U. Rafi, A. Doering, B. Leibe and J. Gall, "Self-supervised keypoint correspondences for multiperson pose estimation and tracking in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2020, pp. 36-52.

[25] G. Ning and H. Huang, "LightTrack: A Generic Framework for Online Top-Down Human Pose Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun, 2020, pp. 1034-1035.

[26] M. Snower, A. Kadav, F. Lai and H. P. Graf, "15 keypoints is all you need," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2020, pp. 6738-6748.

[27] Y. Xiu, J. Li, H. Wang, Y. Fang and C. Lu, "Pose Flow: Efficient online pose tracking," in *arXiv preprint*, Sep, 2018, arXiv:1802.00977.

[28] W. Jiang, S. Jin, W. Liu and C. Qian, "PoseTrans: A Simple Yet Effective Pose Transformation Augmentation for Human Pose Estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 643-659.

[29] S. Jin, W. Liu, W. Ouyang and C. Qian, "Multi-person articulated tracking with spatial and temporal embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2019, pp. 5664-5673.

[30] R. Girdhar, G. Gkioxari and L. Torresani, "Detect-and-Track: Efficient Pose Estimation in Videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2018, pp. 350-359.

[31] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy and T. Brox, "FlowNet2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2017, pp. 2462-2470.

[32] T. Xu, Z. Feng, X. Wu and J. Kittler, "Learning Adaptive Discriminative Correlation Filters via Temporal Consistency Preserving Spatial Feature Selection for Robust Visual Object Tracking," *IEEE Trans. Image Process.,* vol. 28, no. 11, pp. 5596-5609, Nov, 2019.

[33] Y. Tian, A. Dehghan and M. Shah, "On detection, data association and segmentation for multi-target tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 41, no. 9, pp. 2146-2160, Sep, 2019.

[34] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 1-21.

[35] L. Zhang, L. Yuan and R. Nevatia, "Global data association for multiobject tracking using network flows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2008, pp. 1-8.

[36] C. Kim, F. Li, A. Ciptadi and J. M. Rehg, "Multiple hypothesis tracking revisited," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec, 2015, pp. 4696-4704.

[37] A. Ošep, W. Mehner, M. Mathias and B. Leibe, "Combined image and world-space tracking in traffic scenes," in *Proc. IEEE ICRA*, May, 2017, pp. 1988-1995.

[38] A. Geiger, M. Lauer, C. Wojek, C. Stiller and R. Urtasun, "3d traffic scene understanding from movable platforms," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 36, no. 5, pp. 1012-1025, May, 2014.

[39] G. Wang, X. Wu, Z. Liu and H. Wang, "Hierarchical attention learning of scene flow in 3d point clouds," *IEEE Trans. Image Process.,* vol. 30, pp. 5168-5181, 2021.

[40] D. Mitzel and B. Leibe, "Taking mobile multi-object tracking to the next level: People, unknown objects, and carried items," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug, 2012, pp. 566-579.

[41] J. Luiten, P. Voigtlaender and B. Leibe, "Premvos: Proposal-generation, refinement and merging for video object segmentation," in *Proc. ACCV.*, Dec, 2018, pp. 565-580.

[42] X. Liu, C. R. Qi and L. J. Guibas, "FlowNet3D: Learning Scene Flow in 3D Point Clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2019, pp. 529-537.

[43] T. W. Hui and X. Tang, "A Lightweight Optical Flow CNN - Revisiting Data Fidelity and Regularization," in *arXiv preprint*, May, 2019, arXiv:1903.07414v1.

[44] H. Mittal, B. Okorn and D. Held, "Just Go with the Flow: Self-Supervised Scene Flow Estimation," in *arXiv preprint*, Dec, 2019, arXiv:1912.00497.

[45] M. Heo, S. Hwang, S. W. Oh, J. Y. Lee and S. J. Kim, "VITA: Video Instance Segmentation via Object Token Association," in *arXiv preprint*, 2022, arXiv:2206.04403.

[46] M. Heo, S. Hwang, S. W. Oh, J. Y. Lee and S. J. Kim, "Integrating Pose and Mask Predictions for Multi-person in Videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2022, pp. 2657-2666.

[47] G. Nam, M. Heo, S. W. Oh, J. Y. Lee and S. J. Kim, "Polygonal Point Set Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5569-5578.

[48] D. Wang and S. Zhang, "Contextual Instance Decoupling for Robust Multi-Person Pose Estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2022, pp. 11060-11068.

[49] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen and L. Shao, "Hierarchical Human Parsing with Typed Part-Relation Reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2020, pp. 8929-8939.

[50] S. Park, B. X. Nie and S. C. Zhu, "Attribute and-or grammar for joint parsing of human pose, parts and attributes," *IEEE Trans. Pattern Anal. Mach. Intell.,* p. 1555–1569, 7 40 Jul, 2018.

[51] T. Vayer, L. Chapel, R. Flamary, R. Tavenard and N. Courty, "Optimal Transport for structured data with application on graphs," in *arXiv preprint* , May, 2018, arXiv:1805.09114.

[52] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *arXiv preprint* , Apr, 2017, arXiv:1609.02907.

[53] F. Gama, A. G. Marques, G. Leus and A. Ribeiro, "Convolutional Neural Network Architectures for Signals Supported on Graphs," *IEEE Trans. Image Process.,* vol. 67, no. 4, pp. 1034-1049, Feb, 2016.

[54] V. Cherukuri, T. Guo, S. J. Schiff and V. Monga, "Deep MR Brain Image Super-Resolution Using Spatio-Structural Priors," *IEEE Trans. Image Process.,* vol. 29, pp. 1368-1383, Jun, 2020.

[55] S. Yan, Y. Xiong and D. Lin , "Spatial temporal graph convolutional networks for skeleton-based action," *Proc. AAAI Conf. Artif. Intell. (AAAI),* vol. 1, p. 32, Feb, 2017.

[56] S. Jin, W. Liu and P. Luo, "Differentiable Hierarchical Graph Grouping for Multi-Person Pose Estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 718-734.

[57] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug, 2020, pp. 402-419.

[58] G. Peyré, M. Cuturi and J. Solomon, "Gromov-Wasserstein averaging of kernel and distance matrices," in *Proc. ICML*, Jun. 2016, pp. 2664-2672.

[59] "Posetrack 2017: Leather board.," 2017. [Online]. Available: https://posetrack.net/leaderboard.php.

[60] "Posetrack 2018: Leather board.," 2018. [Online]. Available: https://posetrack.net/workshops/eccv2018/posetrack_eccv_2018_results.html.

[61] M. R. Ronchi and P. Perona, "Benchmarking and error diagnosis in multi-instance pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct, 2017, pp. 369-378.

[62] A. Milan, L. L. Taixe, I. Reid, S. Roth and K. Schindler, "MOT16: A benchmark for multi-object tracking," in *arXiv preprint* , Mar, 2016, arXiv:1603.00831.

[63] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé and B. Leibe, "Higher Order Tracking Accuracy," *International journal of computer vision,* vol. 129, pp. 548-578, 2020.

[64] Y. Yang, J. Yang, Y. Xu, J. Zhang, L. Lan and D. Tao, "Apt-36k: A large-scale benchmark for animal pose estimation and tracking," in *arXiv preprint*, 2023, arXiv:2206.05683.

[65] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour and S. Venkatesh, "Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2019, pp. 11996-12004.

[66] W. Luo, W. Liu and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct, 2017, pp. 341-349.

[67] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," in *arXiv preprint*, 2018, arXiv:1804.02767.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv preprint*, Dec, 2014, arXiv:1412.6980.

[69] L. Lin, H. Fan, Z. Zhang, Y. Xu and H. Ling, "Swintrack: A simple and strong baseline for transformer tracking," in *arXiv preprint*, 2021, arXiv:2112.00995.

[70] J. Zhang, Z. Zhu, W. Zou, P. Li, Y. Li, H. Su and G. Huang, "Fastpose: Towards real-time pose estimation and tracking via scale-normalized multi-task networks," in *arXiv preprint* , Aug, 2019, arXiv:1908.05593.

[71] S. Ferradans, N. Papadakis, G. Peyré and J. F. Aujol, "Regularized discrete optimal transport," *SIAM Journal on Imaging Sciences,* vol. 73, no. 3, pp. 1853-1882, Jan, 2014.

[72] P. Bergmann, T. Meinhardt and L. Leal-Taixe, "Tracking without bells and whistles," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 941-951.

Yalong Jiang received the Ph.D. degree in the Department of Electronic and Information Engineering, Hong Kong Polytechnic University. Since April 2020, he has been with Beihang University, where he is now an assistant professor in the Unmanned System Institute. His research interests include pattern recognition, computer vision, and machine learning.

Wenrui Ding received the doctorate degree in electrical and information engineering from Beihang University. She is currently in charge of information transmission and processing data link in the Unmanned System Research Institute in Beihang University. Her research interests include the command and control of aerial vehicles and image processing.

Hongguang Li received Ph.D. degree in aerospace science and technology from Beihang University, Beijing, China. He currently works at the institute of unmanned systems of Beihang University in China. His research interests include intelligent image processing in unmanned systems.

Zheru Chi received the BEng and MEng degrees from Zhejiang University, in 1982 and 1985, respectively, and the PhD degree from the University of Sydney, in March 1994, all in electrical engineering. He worked as a senior research assistant/research fellow in the Laboratory for Imaging Science and Engineering, University of Sydney, from April 1993 to January 1995. Between February 1995 and April 2021, he was with the Department of Electronic and Information Engineering, Hong Kong Polytechnic University, as an assistant/associate professor. He was an associate editor of the IEEE Transactions on Fuzzy Systems between 2008 and 2010, and is currently an editor of the International Journal of Information Acquisition. His research interests include image processing, pattern recognition, and computational intelligence. He has published more than 250 technical papers.