Pose Flow: Efficient Online Pose Tracking*

Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, Cewu Lu
Shanghai Jiao Tong University
{yuliangxiu, ljf_likit, why2011btv, yhfang, lucewu}@sjtu.edu.cn

Abstract

Multi-person articulated pose tracking in complex unconstrained videos is an important and challenging problem. In this paper, going along the road of top-down approaches, we propose a decent and efficient pose tracker based on pose flows. First, we design an online optimization framework to build association of cross-frame poses and form pose flows. Second, a novel pose flow non maximum suppression (NMS) is designed to robustly reduce redundant pose flows and re-link temporal disjoint pose flows. Extensive experiments show our method significantly outperforms best reported results on two standard Pose Tracking datasets ([Iqbal *et al.*, 2017] and [Girdhar *et al.*, 2017]) by **13 mAP 25 MOTA** and **6 mAP 3 MOTA** respectively. Moreover, in the case of working on detected poses in individual frames, the extra computation of proposed pose tracker is very minor, requiring 0.01 second per frame only.

1 Introduction

Motivated by its extensive applications in human behavior understanding and scene analysis, human pose estimation has witnessed a significant boom in recent years. Mainstream research fields have advanced from pose estimation of single pre-located person [Newell *et al.*, 2016; Chu *et al.*, 2017] to multi-person pose estimation in complex and unconstrained scenes [Cao *et al.*, 2016; Fang *et al.*, 2016]. Beyond static human keypoints in individual images, pose estimation in videos has also emerged as a prominent topic [Song *et al.*, 2017; Zhang and Shah, 2015]. Furthermore, human pose trajectory extracted from the entire video is a high-level human behavior representation [Wang and Schmid, 2013; Wang *et al.*, 2015], naturally providing us with a powerful tool to handle a series of visual understanding tasks, such as Action Recognition [Chéron and Laptev, 2015; Zolfaghari *et al.*, 2017], Person Re-identification [Su *et al.*, 2017; Zheng *et al.*, 2017], Human-Object Interaction [Gkioxari *et al.*, 2017] and numerous downstream practical applications, e.g., video surveillance and sports video analysis.

*Yuliang Xiu and Jiefeng Li contributed equally to this paper

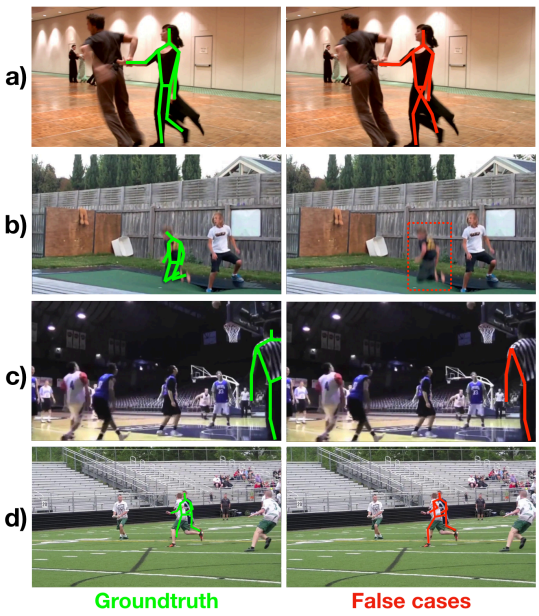

Figure 1: Failure cases of previous pose estimation methods, ground-truth in green and false cases in red. a) Ambiguous assignment. b) Missing detection. c) Human truncation. d) Human occlusion.

To this end, multi-person pose tracking methods are developed, whose dominant approaches can be categorized into top-down [Girdhar *et al.*, 2017] and bottom-up [Insafutdinov *et al.*, 2016a; Iqbal *et al.*, 2017]. Top-down methods first detect human bounding boxes in every frame, estimate human keypoints within each box independently, and then track human boxes over the entire video in terms of similarity between pairs of boxes in adjacent frames, and that is the reason why it is also referred to as Detect-and-Track method [Girdhar *et al.*, 2017]. By contrast, bottom-up methods first generate a set of joint detection candidates in every frame, construct the spatio-temporal graph, and then solve an integer linear program to partition this graph into sub-graphs that correspond to plausible human pose trajectories of each person. Both strategies have their own advantages and disadvantages.

Currently top-down methods have largely outperformed bottom-up ones, since the absence of global viewpoint of

bottom-up approach causes ambiguous assignments of keypoints, as Figure 1 a) shows. Therefore, top-down methods may be a more promising direction. Following this direction, however, there are many obstacles. First, due to occlusion, truncation and frame degeneration (e.g. blurring), as shown in Figure 1 b) c) d) , pose estimation in individual frame can be unreliable. This requires associating cross-frame detected instances to share information and thus reduce uncertainty.

In this paper, we propose an efficient and decent method to achieve online pose tracking. The proposed method includes two novel techniques, namely Pose Flow Building and Pose Flow NMS. First, we associate the cross-frame poses that indicate the same person, i.e., iteratively constructing pose flow from pose proposals within a short video clip picked by a temporal video sliding window. Instead of employing greedy match, we design an elegant objective function to seek a pose flow with maximum overall confidence among potential flows. This optimization design helps to stabilize pose flows and associate discontinuous ones (due to missing detections). Second, unlike conventional schemes that apply NMS in frame-level, we propose pose flow NMS, that is, to take pose flow as a unit in NMS processing. In this way, temporal information will be fully considered in NMS process and thus stabilization can be largely improved. Our approach is general to different image-based multi-person pose estimation and takes minor extra computation. Given detected poses in individual frames, our method can track at 100 FPS.

Intensive experiments are conducted to verify the effectiveness of proposed framework. Two standard pose tracking datasets **PoseTrack Dataset** [Iqbal *et al.*, 2017] and **PoseTrack Challenge Dataset** [Andriluka and Iqbal, 2017] are used to benchmark our performance. Our proposed approach significantly outperforms the state-of-the-art method [Girdhar *et al.*, 2017], achieving 58.5% MOTA and 66.5% mAP in PoseTrack Challenge validation set, 50.9% MOTA and 62.9% mAP in testset.

2 Related Work

2.1 Multi-Person Pose Estimation in Image

In recent years, multi-person pose estimation in images has experienced large performance advancement. With respect to different pose estimation pipelines, relevant work can be grouped into graph decomposition and multi-stage techniques. Graph decomposition methods, such as DeeperCut [Insafutdinov *et al.*, 2016b], re-define the multi-person pose estimation problem as a partitioning and labeling formulation and solve this graph decomposition problem by an integer linear program. These methods' performance depends largely on strong parts detector based on deep visual representations and efficient optimization strategy. However, their body parts detector always performs vulnerably because of absence of global context and structural information. OpenPose [Cao *et al.*, 2016] introduces Part Affinity Fields (PAFs) to associate body parts with individuals in image, but ambiguous assignments still occur in crowds. To address this limitation, multi-stage pipeline [Fang *et al.*, 2016; Chen *et al.*, 2017] handles multi-person pose estimation problem by separating this task into human detection, single person pose estimation and post-processing stages. The main difference among dominant multi-stage frameworks lies in different choices of human detector and single person pose estimator network. With the remarkable progress of object detection and single person pose estimator over the past few years, the potentials of multi-stage approaches have been deeply exploited. Now multi-stage framework has been in the epicenter of the methods above, achieving the state-of-the-art performance in almost all benchmark datasets, e.g., MSCOCO and MPII.

2.2 Multi-Person Articulated Tracking in Video

Based on the multi-person pose estimation architectures described above, it is natural to extend them from still image to video. Pose estimation in single person videos has been explored extensively in the literature [Song *et al.*, 2017; Zhang and Shah, 2015]. These methods focus on using temporal smoothing constraints and matching features between adjacent frames to improve localization of keypoints. However, these architectures are not scalable for multi-person tasks, especially in unconstrained videos with unknown number of highly occluded persons. PoseTrack [Iqbal *et al.*, 2017] and ArtTrack [Insafutdinov *et al.*, 2016a] in CVPR'17 primarily introduce multi-person pose tracking challenge and propose a new graph partitioning formulation, building upon 2D DeeperCut [Insafutdinov *et al.*, 2016b] by extending spatial joint graph to spatio-temporal graph. Although plausible results can be guaranteed by solving this minimum cost multicut problem, hand-crafted graphical models are not scalable for long clips of different unseen types of scenes. It is worth noting that solving this sophisticated IP formulation requires tens of minutes per video, even implemented with state of the art solvers. Hence, we would like to embrace a more efficient and scalable top-down method: first detect persons in every frames, operate single person pose estimation on every detection, and then link them in terms of appearance similarity and temporal relationship between pairs of boxes. Yet some issues should be dealt with properly: how to filter redundant boxes correctly with fusion of information from adjacent frames, how to produce robust pose trajectories by leveraging temporal information, and how to connect human boxes with the same identity meanwhile keeping away from disturbance of scale variance. Although some latest work try to give their solution to these problems, such as 3D Mask R-CNN[Girdhar *et al.*, 2017] which is designed for correcting keypoints' location by leveraging temporal information in 3D human tubes, these work still did not exploit full potential of top-down architecture, just improving keypoint location task slightly but not employing pose flow as a unit. This is what our Pose Flow Builder and Pose Flow NMS focus on.

2.3 Multi-Object Tracking

Multi-Object Tracking (MOT) is a deeply explored traditional visual topics. Recent approaches primarily focus on Tracking-by-Detection pipeline, which either operates on online linking detections over time. [Kim *et al.*, 2015; Choi, 2015] or grouping detections into tracklets and then merge them into tracks [Bing Wang, 2015]. There are merit and demerit in both methods. Group-and-Merge method can

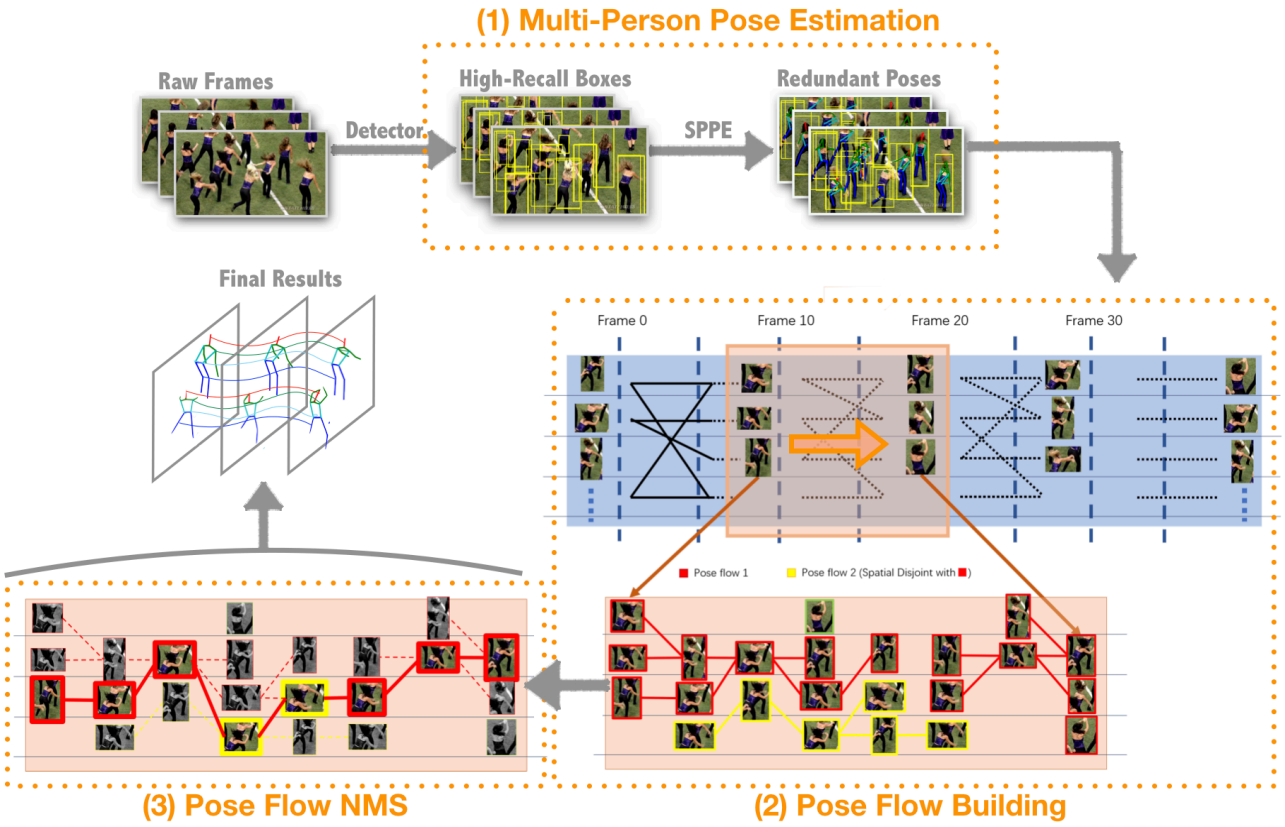

Figure 2: Overall Pipeline. The whole pipeline contains: 1) Pose Estimation. 2) Pose Flow Building. 3) Pose Flow NMS. First, we estimate multi-person poses. Second, we build pose flows by maximizing overall confidence and purify them by Pose Flow NMS. Finally, reasonable multi-pose trajectories can be obtained.

only take effect when there is no requirement for online tracking. Nevertheless, due to its global optimization mechanism, when long term tracking is expected, its performance can surpass most online trackers. On the other hand, some online trackers just simplify this tracking problem as a maximum weight bipartite matching problem and solve it with greedy or Hungarian Algorithm. Nodes of this bipartite graph are human bounding boxes in two adjacent frames. This configuration did not take pose information into account, which is essential in tracking occasional truncated human. To address this limitation, meanwhile maintaining its efficiency, we put forward a new pose flow generator, which combines Pose Flow Building and Pose Flow NMS. Furthermore, we re-design two kinds of ORB based similarity criteria, inspired by [Tang *et al.*, 2016].

3 Our Proposed Approach

In this section, we present our pose tracking framework. As mentioned before, pose flow means a set of pose indicating the same person instance in different frames. Our framework includes two steps: Pose Flow Building and Pose Flow NMS. First, we build pose flow by maximizing overall confidence along the temporal sequence. Second, we reduce redundant pose flows and relink disjoint pose flows by Pose Flow NMS. The overall pipeline shows in Figure 2.

3.1 Preliminary

In this section, we introduce some basic metrics and tools that will be used in our framework.

Intra-Frame Pose Distance Intra-frame Pose distance is defined to measure the similarity between two poses P_1 and P_2 in a frame. We adopt the image pose distance defined in [Fang *et al.*, 2016]. The soft matching function is defined as

$$K_{Sim}(P_1, P_2|\sigma_1) = \begin{cases} \sum_n tanh\frac{c_1^n}{\sigma_1} \cdot tanh\frac{c_2^n}{\sigma_1} & \text{if } p_2^n \text{ is within } B(p_1^n) \\ 0 & otherwise \end{cases} \quad (1)$$

where $B(p_1^n)$ is the box of p_1^n and its size follows standard PCK metric [Mykhaylo Andriluka, 2014]. c_1 and c_2 are the keypoint scores of p_1 and p_2. The $tanh$ function is to suppress the low score keypoints. If both poses are confident with high scores, the function output is closed to 1.

The spatial distance among keypoints are written as

$$H_{Sim}(P_1, P_2|\sigma_2) = \sum_n exp[-\frac{(p_1^n - p_2^n)^2}{\sigma_2}] \quad (2)$$

The final distance combining Eqs. 1 and 2 is written as

$$\begin{aligned} & d_f(P_1, P_2) \\ & = K_{Sim}(c_1, c_2|\sigma_1)^{-1} + \lambda H_{Sim}(p_1, p_2|\sigma_2)^{-1} \end{aligned} \quad (3)$$

where $\Lambda = \{\sigma_1, \sigma_2, \lambda\}$. Note that these parameters can be determined in a data-driven manner.

Inter-frame Pose Distance Inter-frame pose distance is to measure distance between a pose P_1 in one frame and another pose P_2 in the next frame. We need to import temporal matching to measure how likely two cross-frame poses indicate the same person. We denote p_1^i and p_2^i as the i^{th} keypoints of pose P_1 and P_2 respectively. Bounding boxes surrounding p_1^i and p_2^i are extracted and denoted as B_1^i and B_2^i. The box size is 10% person bounding box size according to the standard PCK [Mykhaylo Andriluka, 2014]. We evaluate the similarity of B_1^i and B_2^i by implementing ORB matching [Rublee *et al.*, 2011]. Given m_i feature points extracted from B_1^i, we can find n_i matching points in B_2^i. Obviously, the matching percentage $\frac{n_i}{m_i}$ can indicate the similarity of B_1^i and B_2^i. Therefore the inter-frame pose distance between P_1 and P_2 can be expressed as:

$$d_c(P_1, P_2) = \sum_i \frac{n_i}{m_i} \tag{4}$$

3.2 Multi-Person Pose Estimation

We adopt RMPE [Fang *et al.*, 2016] as our multi-person pose estimator, which uses Faster R-CNN[Ren *et al.*, 2015] as human detector and Pyramid Network[Yang *et al.*, 2017] as single person pose estimator. Our pipeline is ready to adopt different human detectors and pose estimators.

Data Augmentation To facilitate RMPE to perform better in presence of serious truncation and occlusion of humans, we propose a improved deep proposal generator (iDPG) as a data augmentation scheme. iDPG produces a large amount of truncated human proposals using random-crop strategy during training. Empirically, random-crop algorithm will crop normal human into quarter or half man. Thus, those random-crop proposal will be used as augmented training data. We observe the improvement of RMPE, when it applies into video frame.

3.3 Pose Flow Building

We firstly preform frame by frame pose estimation. Pose flows are built by associating poses that indicates the same person across frames. The straight-forward method is to connect them by matching closest pose in the next frame, given metric $d_c(P_1, P_2)$. However, this greedy scheme would be less effective due to recognition error and false alarm of frame-level pose detection. On the other hand, if we apply graph-cut model in spatial and temporal domains, it will lead to heavy computation and non-online solution. Therefore, in this paper, we propose an efficient and decent method for high quality pose flow building. We denote P_i^j as the i^{th} pose at j^{th} frame and its candidate association set as

$$\begin{aligned} \mathcal{T}(P_i^j) = \{P | d_c(P, P_i^j) \leq \epsilon\}, \\ s.t. P \in \Omega_{j+1} \end{aligned} \tag{5}$$

where Ω_{j+1} is the set of pose at $(j+1)^{th}$ frame. In our paper, we set $\epsilon = \frac{1}{25}$ by cross-validation. $\mathcal{T}(P_i^j)$ means possible corresponding pose set in next frame for P_i^j. Without lose of generality, we discuss tracking for P_i^t and consider pose flow building from t^{th} to $(t+T)^{th}$ frames. To optimize pose selection, we maximize the following objective function

$$\begin{aligned} F(t, T) = \max_{Q_t, \ldots, Q_{t+T}} \sum_{i=t}^{t+T} s(Q_i), \\ s.t.\ Q_0 = P_i^t, \\ s.t.\ Q_i \in \mathcal{T}(Q_{i-1}) \end{aligned} \tag{6}$$

where $s(P)$ is a function that outputs confidence score of P, defined as

$$s(P) = s_{box}(P) + mean(s_{pose}(P)) + \max(s_{pose}(P)), \tag{7}$$

where $s_{box}(P)$, $mean(s_{pose}(P))$ and $\max(s_{pose}(P))$ are score of human box, mean score and max score of all keypoints within this human proposal, respectively. The optimum $\{Q_t, \ldots, Q_{t+T}\}$ is our pose flow for P_i^t from t^{th} to $(t+T)^{th}$ frame.

Analysis We regard the sum of confidence scores ($\sum_{i=t}^{t+T} s(Q_i)$) as objective function. This design helps us resist many uncertainties. When a person is highly occluded or blurred, its score is quite low because the model is not confident about it. But we can still build a pose flow to compensate it, since we look at the overall confidence score of a pose flow, but not single frame. Moreover, the sum of confidence score can be calculated online. That is, $F(t, T)$ can be determined by $F(t, T-1)$ and $s(Q_T)$.

Solver Eq. 6 can be solved in an online manner, since it is a standard dynamic programming problem. At $(u-1)^{th}$ frame, we have m_{u-1} possible poses and record m_{u-1} optimum pose trajectories (with sum of scores) to reach them. At u^{th} frame, we compute the optimum path to m_u possible poses based on previous m_{u-1} optimum pose trajectories. Accordingly, m_u trajectories are updated. $F(u)$ is the sum of scores of best pose trajectories.

Stop Criterion and Confidence Unification

We process video frame-by-frame with Eq. 6 until it meets a stop criterion. Our criterion doesn't simply check confidence score in a single frame, but looks at more frames to resist sudden occlusion and frame degeneration (e.g. motion blur). Therefore, a pose flow stops at u when $F(t, u+r) - F(t, u) < \gamma$, where γ is determined by cross validation. It means the sum of scores within the following r frames is very small. Only in this way, we can make sure a pose flow really stops. In our paper, we set $r = 3$. After a pose flow stops, we refresh all keypoint confidence by averaging confidence scores. We believe pose flow should be the basic block and should use single confidence value to represent it. We call this process as confidence unification.

3.4 Pose Flow NMS

We hope our NMS can be preformed in spatio-temporal domain instead of individual frame processing. That is, we take poses in a pose flow as a unit in NMS processing, which can reduce errors by both spatial and temporal information. The

key step is to determine the distance of two pose flows that indicate the same person.

Pose Flow Distance Given two pose flows $\mathcal{Y}_a$ and $\mathcal{Y}_b$, we can extract their temporal overlapping sub-flows. The sub-flows are denoted as $\{P_a^1, \ldots, P_a^N\}$ and $\{P_b^1, \ldots, P_b^N\}$, where N is the number of temporal overlapping frames. That is, P_a^i and P_b^i are two poses in the same frame. The distance between $\mathcal{Y}_a$ and $\mathcal{Y}_b$ can be calculated as,

$$d_{PF}(\mathcal{Y}_a, \mathcal{Y}_b) = median[\{d_f(P_a^1, P_b^1), \ldots, d_f(P_a^N, P_b^N)\}] \tag{8}$$

where $d_f(\cdot)$ is the intra-frame pose distance defined in Eq. 3. Median metric can be more robust to resist some miss-detection due to occlusion and motion blur.

Pose Flow Merging Given $d_{PF}(\cdot)$, we can perform NMS scheme as conversional pipeline. First, the pose flow with maximum confidence score (after confidence unification) is selected as reference pose flow. Making use of $d_{PF}(\cdot)$, we group pose flows closed to reference pose flow. Thus, pose flows in the group will be merged into one more robust pose flow to represent the group. This new pose flow (pose flow NMS result) is called representative pose flow. The 2D coordinate of i^{th} keypoint $\mathbf{x}_{t,i}$ and confidence score $s_{t,i}$ of representative pose flow in t^{th} frame are computed by,

$$\hat{\mathbf{x}}_{t,i} = \frac{\sum_j s_{t,i}^j \mathbf{x}_{t,i}^j}{\sum s_{t,i}^j} \quad and \quad \hat{s}_{t,i} = \frac{\sum_j s_{t,i}^j}{\sum \mathbb{1}(s_{t,i}^j)} \tag{9}$$

where $\mathbf{x}_{t,i}^j$ and $s_{t,i}^j$ are the 2D coordinate and confidence score of i^{th} keypoint in j^{th} pose flow in the group in t^{th} frame. If j^{th} pose flow does not have any pose at t^{th} frame, we set $s_{t,i}^j = 0$. In Eq. 9, $\mathbb{1}(s_{t,i}^j)$ outputs 1, if input is non-zero, otherwise it outputs 0. This merging step not only can reduce redunant pose flow, but also re-link some disjoint pose flows into a longer and completed pose flow.

We redo this process until all pose flows are processed. This process is computed in sliding temporal window (the window length is $L = 20$ in our paper). Therefore, it is an online process. The whole pipeline shows in Figure 3.

4 Experiments and Results

We evaluate our framework on **PoseTrack** [Iqbal *et al.*, 2017] dataset and **PoseTrack Challenge** [Andriluka and Iqbal, 2017] dataset.

4.1 Evaluation and Datasets

For comparison with both state-of-the-art top-down and bottom-up approaches, we evaluate our framework on **PoseTrack** and **PoseTrack Challenge** dataset separately. PoseTrack Dataset was introduced in [Iqbal *et al.*, 2017], which is used to evaluate the spatio-temporal graph-cut method. Labeled frames in this dataset come from consecutive unlabeled adjacent frames of MPII Multi-Person Pose dataset[Mykhaylo Andriluka, 2014]. These selected videos contain multiple persons and cover a wide variety of activities of complex cases, such as scale variation, body truncation, severe occlusion and motion blur. For fair comparison, we train our RMPE on 30 training videos and test it on the rest 30 testing videos like PoseTrack graph-cut framework [Iqbal *et al.*, 2017]. Table 1 presents tracking results in PoseTrack dataset, and pose estimation results show in Table 4. It shows that our method outperforms best reported results graph-cut approach by 13.5 mAP and 25.4 MOTA.

Method	[Iqbal *et al.*, 2017]	Ours
Rcll↑	63.0	**65.9**
Prcn↑	64.8	**83.2**
MT↑	775	**949**
ML↓	**502**	623
IDs↓	431	**202**
FM↓	5629	**3358**
MOTA↑	28.2	**53.6**
MOTP↑	55.7	**56.4**

Table 1: PoseTrack Dataset results

Method	**mAP Total**	**MOTA Total**	**MOTP Total**	**Rcll**	**Prcn**
Our Full Architecture	**66.5**	**58.3**	67.8	87.0	70.3
w/o PF	65.9	53.9	62.2	69.7	87.4
w/o PF-NMS	61.2	55.8	68.0	64.0	90.3

Table 2: Ablation comparison. "w/o PF" means the use of naive boxIoU matching instead of Eq. 6. "w/o PF-NMS" means the use of frame-by-frame Pose-NMS instead of our Pose Flow NMS.

PoseTrack Challenge Dataset is released in [Andriluka and Iqbal, 2017]. Selected and annotated like PoseTrack Dataset, it contains more videos. The testing dataset evaluation includes three tasks, but we only join tracking related task, namely, Task2-multi-person pose estimation and Task3-Pose tracking. Task2 is evaluated using the mean average precision (mAP) metric and Task3 is using multi-object tracking accuracy (MOTA) metric. MOTA metric penalizes false positives equally regardless of pose scores, so we drop low-score keypoints before generating results. We empirically determine this threshold in a data-driven manner on validation set. Tracking results of validation set and test set of Posetrack Challenge Dataset are presented in Table 3. We found our method can achieve better results on validation and comparable results in test. Some representative results are shown in Figure. 4.

Time Performance Our proposed pose tracker is based on resulting poses in individual frames. That is, it is ready to apply in different multi-person pose estimators. The extra computation by our pose tracker is very minor, requiring 0.01 second per frame only. Therefore, it will not be the bottleneck of whole system, in terms of testing speed.

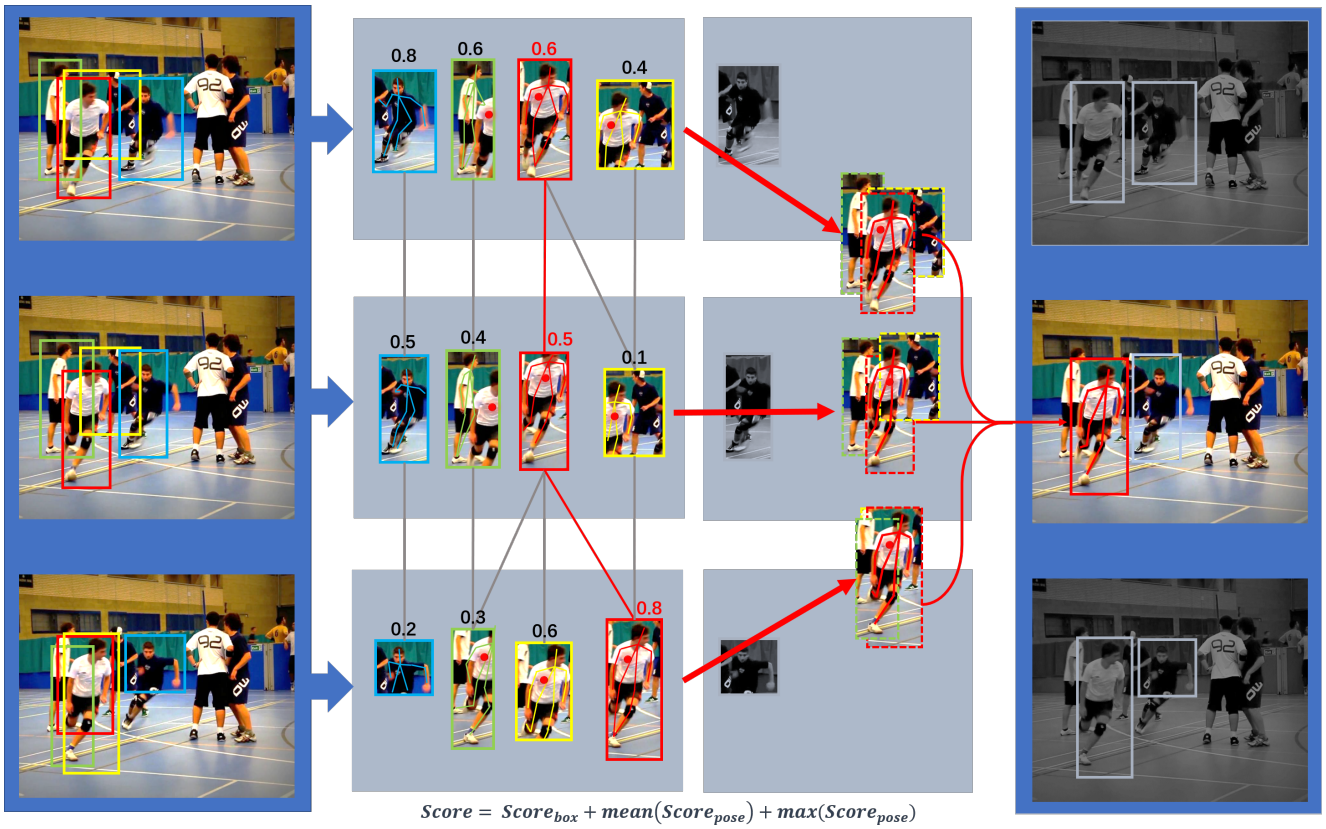

Figure 3: The whole pipeline of Pose Flow NMS.

Method	Dataset	MOTA Head	MOTA Shou	MOTA Elb	MOTA Wri	MOTA Hip	MOTA Knee	MOTA Ankl	MOTA Total	MOTP Total	Rcll	Prcn
[Girdhar *et al.*, 2017]	validation	61.7	65.5	57.3	45.7	54.3	53.1	45.7	55.2	61.5	88.1	66.5
Ours		59.8	67.0	59.8	51.6	60.0	58.4	50.5	**58.3**	67.8	87.0	70.3
[Girdhar *et al.*, 2017]	testset	-	-	-	-	-	-	-	**51.9**	-	-	-
Ours		52.0	57.4	52.8	46.6	51.0	51.2	45.3	51.0	16.9	78.9	71.2

Table 3: Multi-person pose tracking results on PoseTrack Challenge dataset

Method	Dataset	Head mAP	Shoulder mAP	Elbow mAP	Wrist mAP	Hip mAP	Knee mAP	Ankle mAP	Total mAP
[Iqbal *et al.*, 2017]	PoseTrack	56.5	51.6	42.3	31.4	22.0	31.9	31.6	38.2
Ours		64.7	65.9	54.8	48.9	33.3	43.5	50.6	**51.7**
[Girdhar *et al.*, 2017]	PoseTrack Challenge(valid)	67.5	70.2	62	51.7	60.7	58.7	49.8	60.6
Ours		66.7	73.3	68.3	61.1	67.5	67.0	61.3	**66.5**
[Girdhar *et al.*, 2017]	PoseTrack Challenge(test)	-	-	-	-	-	-	-	59.6
Ours		64.9	67.5	65.0	59.0	62.5	62.8	57.9	**63.0**

Table 4: Multi-person pose estimation results on all PoseTrack dataset

4.2 Training and Testing Details

In this paper, we use ResNet152-based Faster R-CNN as human detector. Due to the absence of human proposal annotations, we generate human boxes by extending human keypoints boundary 20% along both height and width directions, which are used for fine-tuning human detector. In the phrase of SPPE training, we employed online hard example mining (OHEM) to deal with hard keypoints like hips and ankles. For each iteration, instead of sampling the highest B/N losses in mini-batch, k highest loss hard examples are selected. After selection, the SPPE update weights only from hard keypoints. These procedures increase slight computation time, but notably improve estimation performance of hips and ankles.

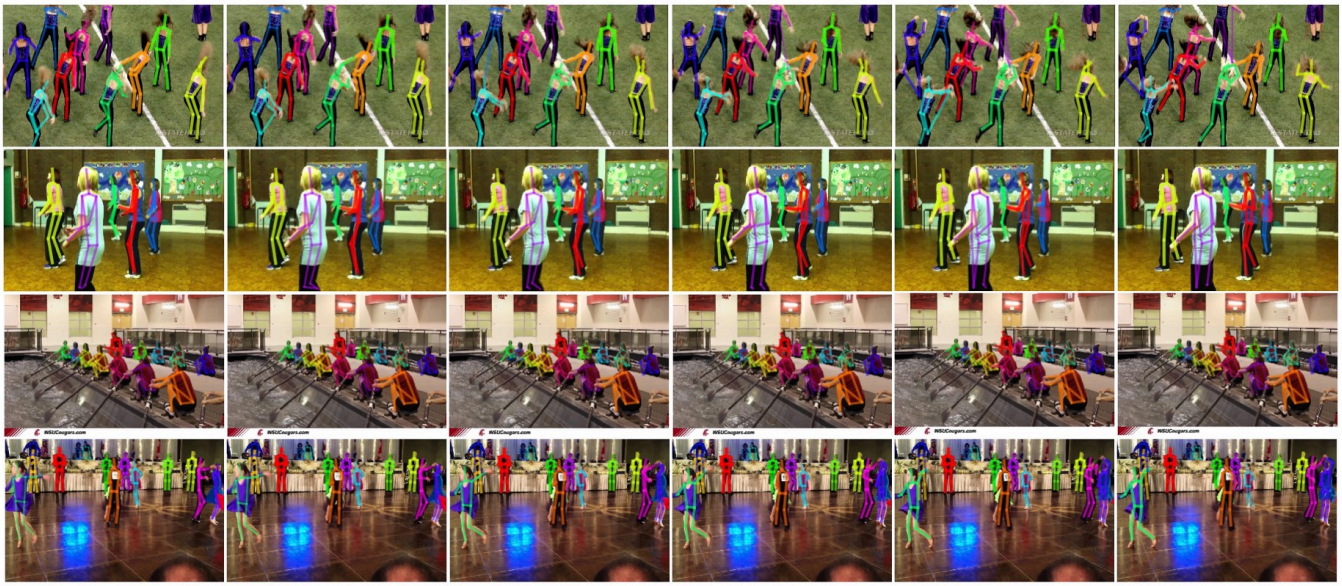

Figure 4: Some final posetracking results in videos

4.3 Ablation Studies

Pose Flow Building

Pose Flow Building module is responsible for constructing pose flow. We use a naive pose matching crossing frames, without the use of Eq. 6. Table 2 shows that without flow building, naive pose matching can not achieve a decent result. It is because pose matching should be considered in a long-term manner.

Pose Flow NMS

Without pose flow NMS, the final pose tracking results will include many disjoint and redundant pose flow, which will damage the final tracking performance significantly. To evaluate the effectiveness of pose flow NMS, we compare it with conventional frame-by-frame NMS that applies in pose flow building results. As in Table 2, results show that Pose Flow NMS can guarantee accurate pose flows, which means high mAP value and high MOTA value in comparison with conventional frame-by-frame NMS.

5 Conclusion

We have presented a scalable and efficient top-down pose tracking framework, which mainly leverages spatio-temporal information to build pose flow to significantly boost pose tracking task. Two important techniques, Pose Flow building and Pose Flow NMS were proposed. In ablation studies, we prove that the combination of Pose Flow Building and Pose Flow NMS can guarantee a remarkable improvement in pose tracking tasks. Moreover, our proposed pose tracker that can process frames in video at 100 FPS (excluding pose estimation in frames) has great potential in realistic applications. In the future, we hope our pose tracker can help to improve long-term action recognition and offer a powerful tool for understanding complex scenes.

References

[Andriluka and Iqbal, 2017] Mykhaylo Andriluka and Iqbal. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. 2017.

[Bing Wang, 2015] Li Wang Bing Wang, Gang Wang. Tracklet association by online target-specific metric learning and coherent dynamics estimation. *CoRR*, abs/1511.06654, 2015.

[Cao *et al.*, 2016] Zhe Cao, Shih-En Wei, Tomas Simon, Yaser Sheikh, Shih-En Wei, and Yaser Sheikh. Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields. 2016.

[Chen *et al.*, 2017] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded Pyramid Network for Multi-Person Pose Estimation. 2017.

[Chéron and Laptev, 2015] Guilhem Chéron and Ivan Laptev. P-CNN: Pose-based CNN Features for Action Recognition. 2015.

[Choi, 2015] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. *CoRR*, abs/1504.02340, 2015.

[Chu *et al.*, 2017] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-Context Attention for Human Pose Estimation. 2017.

[Fang *et al.*, 2016] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional Multi-person Pose Estimation. 2016.

[Girdhar *et al.*, 2017] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-Track: Efficient Pose Estimation in Videos. 2017.

[Gkioxari *et al.*, 2017] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and Recognizing Human-Object Interactions. 2017.

[Insafutdinov *et al.*, 2016a] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. ArtTrack: Articulated Multi-person Tracking in the Wild. 2016.

[Insafutdinov *et al.*, 2016b] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. 2016.

[Iqbal *et al.*, 2017] Umar Iqbal, Anton Milan, and Juergen Gall. PoseTrack: Joint Multi-Person Pose Estimation and Tracking. 2017.

[Kim *et al.*, 2015] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M. Rehg. Multiple hypothesis tracking revisited. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4696–4704, Washington, DC, USA, 2015. IEEE Computer Society.

[Mykhaylo Andriluka, 2014] Leonid Pishchulin Mykhaylo Andriluka. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[Newell *et al.*, 2016] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 483–499. 2016.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. 2015.

[Rublee *et al.*, 2011] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, Nov 2011.

[Song *et al.*, 2017] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-Slicing Network: A Deep Structured Model for Pose Estimation in Videos. 2017.

[Su *et al.*, 2017] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven Deep Convolutional Model for Person Re-identification. 2017.

[Tang *et al.*, 2016] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Multi-Person Tracking by Multicut and Deep Matching. 2016.

[Wang and Schmid, 2013] Heng Wang and Cordelia Schmid. Action recogntiion with improved trajectories. *Proc. Int. Conf. Computer Vision*, pages 3551–3558, 2013.

[Wang *et al.*, 2015] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 4305–4314, 2015.

[Yang *et al.*, 2017] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning Feature Pyramids for Human Pose Estimation. 2017.

[Zhang and Shah, 2015] Dong Zhang and Mubarak Shah. Human Pose Estimation in Videos. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2012–2020, 2015.

[Zheng *et al.*, 2017] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose Invariant Embedding for Deep Person Re-identification. 2017.

[Zolfaghari *et al.*, 2017] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. Chained Multi-stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection. 2017.