Pose Flow: Efficient Online Pose Tracking

Yuliang Xiu
yuliangxiu@sjtu.edu.cn
Jiefeng Li
ljf_likit@sjtu.edu.cn
Haoyu Wang
why2011btv@sjtu.edu.cn
Yinghong Fang
yhfang@sjtu.edu.cn
Cewu Lu
lucw@cs.sjtu.edu.cn

Machine Vision and Intelligence Group
Shanghai Jiao Tong University
Shanghai, China

Abstract

Multi-person articulated pose tracking in unconstrained videos is an important while 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 the association of cross-frame poses and form pose flows (PF-Builder). Second, a novel pose flow non-maximum suppression (PF-NMS) is designed to robustly reduce redundant pose flows and re-link temporal disjoint ones. Extensive experiments show that our method significantly outperforms best reported results on two standard Pose Tracking datasets ([12] and [8]) 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 pose tracker is very minor, guaranteeing online **10FPS** tracking. Our source codes are made **publicly available**[1].

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 [6, 14] to multi-person pose estimation in complex and unconstrained scenes [3, 7]. Beyond static human keypoints in individual images, pose estimation in videos has also emerged as a prominent topic [17, 22]. Furthermore, human pose trajectory extracted from the entire video is a high-level human behavior representation [19, 20], naturally providing us with a powerful tool to handle a series of visual understanding tasks, such as Action Recognition [5], Person Re-identification [18, 24], Human-Object Interaction [9] and numerous downstream practical applications, e.g., video surveillance and sports video analysis.

To this end, multi-person pose tracking methods are developed, whose dominant approaches can be categorized into top-down [8] and bottom-up [11, 12]. Top-down methods,

[1] https://github.com/YuliangXiu/PoseFlow

also known as two steps scheme, first detect human proposals in every frame, estimate 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 [8]. By contrast, bottom-up methods, also known as jointing scheme, 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.

Currently top-down methods have largely outperformed bottom-up methods both in accuracy (mAP and MOTA) and tracking speed, since bottom-up approaches lose a global pose view due to the mere utilization of second-order body parts dependence, which directly cause ambiguous assignments of keypoints, like Figure 1 a). Furthermore, joint schemes are computationally heavy and not scalable to long videos, making it unable to do online tracking. Therefore, top-down methods may be a more promising direction. Following this direction, however, there remains many challenges. As shown in Figure 1 b) c) d), due to frame degeneration (e.g. blurring due to fast motion), truncation or occlusion, pose estimation in an individual frame can be unreliable. To tackle this problem, we need to associate cross-frame detected instances to share temporal information and thus reduce uncertainty.

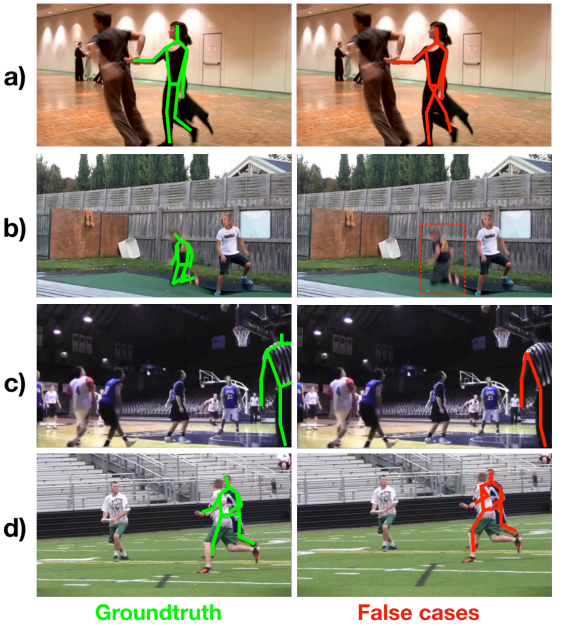

Figure 1: Failure cases of previous pose estimation methods, ground-truth in green and false cases in red. a) Ambiguous assignment. b) Missing detection. c) Human truncation. d) Human occlusion.

In this paper, we propose an efficient and decent method to achieve online pose tracking. Apart from applying an improved RMPE[7] as pose estimator, our proposed method includes two novel techniques, namely Pose Flow Building (PF-Builder) and Pose Flow NMS(PF-NMS). First, we associate the cross-frame poses that indicate the same person. To achieve that, we iteratively construct 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 effective 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, PF-NMS takes 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 pose estimators and only takes minor extra computation for tracking. Given detected poses in individual frames, our method can track poses at 10 FPS.

To verify the effectiveness of proposed framework, we conduct extensive experiments on two standard pose tracking datasets, **PoseTrack Dataset** [12] and **PoseTrack Challenge Dataset** [2]. Our proposed approach significantly outperforms the state-of-the-art method [8], achieving 58.3% MOTA and 66.5% mAP in PoseTrack Challenge validation set, 51.0% MOTA and 63.0% 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 [10], 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 the absence of global context and structural information. OpenPose [3] introduces Part Affinity Fields (PAFs) to associate body parts with individuals in an image, but ambiguous assignments still occur in crowds.

To address this limitation, multi-stage pipeline [4, 7] 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 the 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 greatly 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[13] and MPII[1].

2.2 Multi-Person Articulated Tracking in Video

Based on the multi-person pose estimators described above, it is natural to extend them from still image to video. PoseTrack [12] and ArtTrack [11] in CVPR'17 primarily introduce multi-person pose tracking challenge and propose a new graph partitioning formulation, building upon 2D DeeperCut [10] by extending spatial joint graph to spatio-temporal graph. Although plausible results can be guaranteed by solving minimum cost multicut problem, hand-crafted graphical models are not scalable for long clips of unseen types of scenes. It is worth noting that optimize this sophisticated IP requires tens of minutes per video, even implemented with state of the art solvers.

Hence, another line of research tends to explore more efficient and scalable top-down method by first operating multi-person pose estimation on each frame, and then link them in terms of appearance similarity and temporal relationship between pairs of boxes. Yet some issues should be dealt with properly: 1) how to filter redundant boxes correctly with

信息的融合

the fusion of information from adjacent frames, 2) how to produce robust pose trajectories by leveraging temporal information, 3) how to connect human boxes with the same identity meanwhile keeping away from disturbance of scale variance.

Although one latest work, 3D Mask R-CNN[8], which is designed for correcting the location of keypoints by leveraging temporal information in 3D human tubes, tries to give their solution to these problems, it do not employ pose flow as a unit. Besides, the tracker just simplify 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 motion and pose information into account, which is essential in tracking the occasional truncated human. To address this limitation, meanwhile maintaining its efficiency, we put forward a new pose flow generator, which combines Pose Flow Builder and Pose Flow NMS.

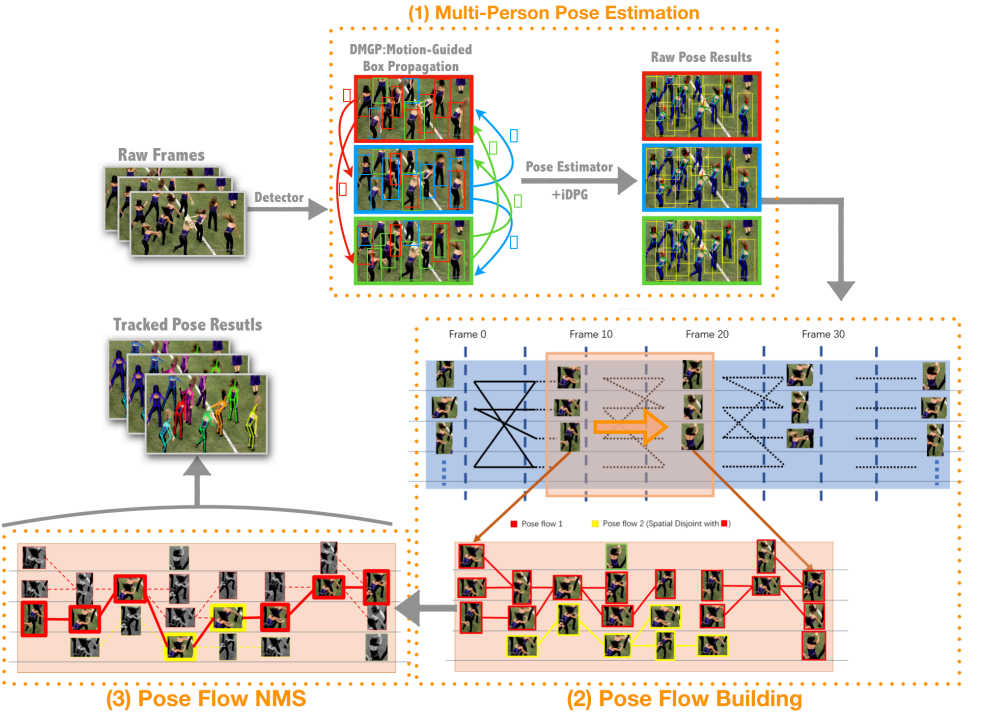

Figure 2: Overall Pipeline: 1) Pose Estimator. 2) Pose Flow Builder. 3) Pose Flow NMS. First, we estimate multi-person poses. Second, we build pose flows by maximizing overall confidence and purify them by Pose Flow NMS. Finally, reasonable multi-pose trajectories can be obtained.

3 Our Proposed Approach

In this section, we present our pose tracking pipeline. As mentioned before, pose flow means a set of pose indicating the same person instance in different frames. As Figure 2 shows, 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.

3.1 Preliminary

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

帧内

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 pose distance defined in [7]. We denote p_1^n and p_2^n as the n^{th} keypoints of pose P_1 and P_2 respectively, $n \in \{1,2,...,N\}$, N is keypoint number of one person, $B(p_1^n)$ is box that centers at p_1^n, c_1^n is score of p_1^n. The *tanh* function is to suppress the low score keypoints.

The soft matching function is defined as

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

The spatial similarity among keypoints written as

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

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

$$\begin{aligned} &d_f(P_1,P_2|\Lambda) \\ &= K_{Sim}(P_1,P_2|\sigma_1)^{-1} + \lambda H_{Sim}(P_1,P_2|\sigma_2)^{-1} \end{aligned} \tag{3}$$

where $\Lambda = \{\sigma_1, \sigma_2, \lambda\}$. 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. Bounding boxes surrounding p_1^n and p_2^n are extracted and denoted as B_1^n and B_2^n. The box size is 10% person bounding box size according to the standard PCK [1]. We evaluate the similarity of B_1^n and B_2^n. Given f_1^n DeepMatching feature [16] points extracted from B_1^n, we can find f_2^n matching points in B_2^n. The matching percentage $\frac{f_2^n}{f_1^n}$ can indicate the similarity of B_1^n and B_2^n. Therefore the inter-frame pose distance between P_1 and P_2 can be expressed as:

$$d_c(P_1,P_2) = \sum_n \frac{f_2^n}{f_1^n} \tag{4}$$

3.2 Improved Multi-Person Pose Estimation

We adopt RMPE [7] as our multi-person pose estimator, which uses Faster R-CNN[15] as the human detector and Hourglass Network with PRMs [21] as single person pose estimator. Our pipeline is ready to adopt to different human detectors and pose estimators.

Data Augmentation In video scenoria, human always come and leave video capturing region, resulting in truncation problem. To handle truncation of humans, we propose an improved deep proposal generator (iDPG) as a data augmentation scheme. iDPG aims to produce truncated human proposals using random-crop strategy during training. Specifically, we randomly crop human instance region into quarter or half person. Thus, those random-crop proposals will be used as augmented training data. We observe an improvement of RMPE when it applies to the video frames

Motion-Guided Box Propagation Due to motion blur and occlusion, missing detection happens frequently during human detection phrase. This will increase person id switches (IDs↓), like in Table 4.1, dramatically degrading final tracking MOTA performance. Our idea is to propagate box proposals to previous and next frames by crossing frame matching technique. That is, the box proposals triple. In this way, some missing detected proposals have high channce to be recovered and largely improve the recall (redundant boxes will be filter out by following step). The cross-frame matching technique we used is deepmatching[16].

3.3 Pose Flow Building

We firstly perform pose estimation for each frame. Pose flows are built by associating poses that indicate the same person across frames. The straight-forward method is to connect them by selecting 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 the 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

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

where Ω_{j+1} is the set of pose at $(j+1)^{th}$ frame. In paper, we set $\varepsilon = \frac{1}{25}$ by cross-validation. $\mathcal{T}(P_i^j)$ means possible corresponding pose set in next frame for P_i^j. Without loss 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 $s(Q_i)$ is a function that outputs confidence score of Q_i, which is defined as

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

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 instead of a 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 paths 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.

3.3.1 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, all keypoint confidence are updated by average confidence scores. We believe pose flow should be the basic block and should use single confidence value to represent it. This process is referred to as confidence unification.

3.4 Pose Flow NMS

We hope our NMS can be performed in the spatio-temporal domain instead of individual frame processing. That is, we take poses in a pose flow as a unit in NMS processing, reducing errors by both spatial and temporal information. The key step is to determine the distance between 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 $\{P_a^1,\dots,P_a^N\}$ and $\{P_b^1,\dots,P_b^N\}$, 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,

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

where $d_f(\cdot)$ is the intra-frame pose distance defined in Eq. 3. The median metric can be more robust towards outliers, such as miss-detection due to occlusion and motion blur.

Pose Flow Merging Given $d_{PF}(\cdot)$, we can perform NMS scheme as convection pipeline. First, the pose flow with the 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 a more robust pose flow representing 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. Details of cross-frame pose merging (keypoint-level) can be refered to Figure 3.

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 2.

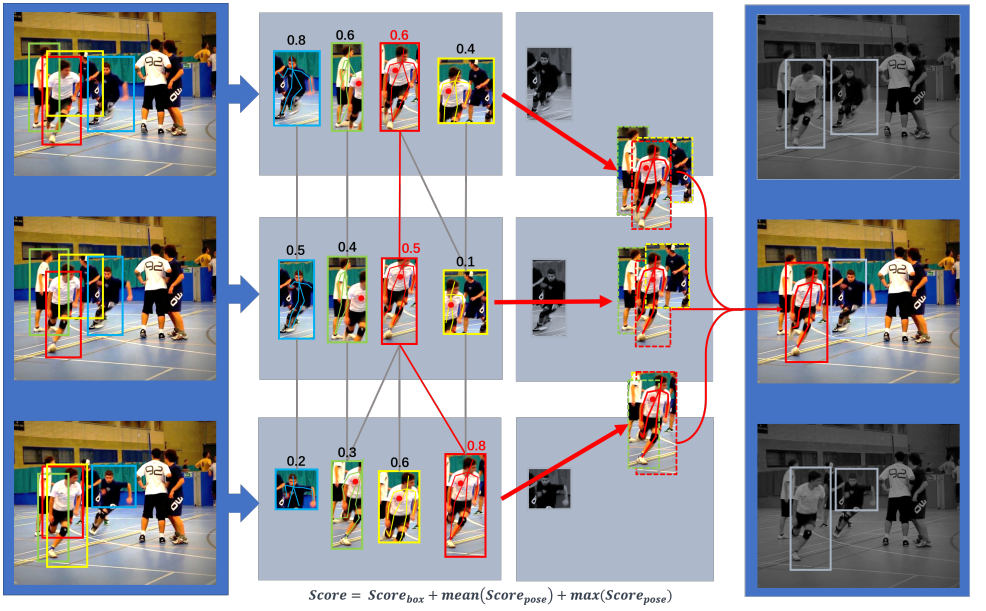

Figure 3: Pose Flow Merging

4 Experiments and Results

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 [12], 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[1]. 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 a fair comparison, we train improved RMPE on 30 training videos and test it on the rest 30 videos like PoseTrack [12] did. Table 4.1 presents tracking results in PoseTrack dataset, and pose estimation results in Table 3. It shows that our method outperforms best reported graph-cut approach by **13.5 mAP** and **25.4 MOTA**.

Method	Rcll↑	Prcn↑	MT↑	ML↓	IDs↓	FM↓	MOTA↑	MOTP↑
Iqbal *et al.* [12]	63.0	64.8	775	**502**	431	5629	28.2	55.7
Ours	**65.9**	**83.2**	**949**	623	**202**	**3358**	**53.6**	**56.4**

Table 1: Multi-person pose tracking results on PoseTrack dataset

PoseTrack Challenge Dataset is released in [2]. Selected and annotated like PoseTrack Dataset, it contains more videos. The testing dataset evaluation includes three tasks, but we only join Task2-Multi-Frame Person Pose Estimation, evaluated by mean average precision (mAP) and Task3-Pose tracking, evaluated by multi-object tracking accuracy (MOTA) metric. Tracking results of validation set and the test set of PoseTrack Challenge Dataset are presented in Table 2. Our method can achieve state-of-the-art results on validation and comparable results on test set. Some qualitative results are shown in Figure 4.

Method	Dataset	MOTA Head	MOTA Shou	MOTA Elb	MOTA Wri	MOTA Hip	MOTA Knee	MOTA Ankl	MOTA Total	MOTP Total	Prcn	Rcll
Girdhar *et al.* [8]	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.* [8]	*(Mini)Test v1.0	55.9	59.0	51.9	43.9	47.2	46.3	40.1	49.6	34.1	81.9	67.4
Ours	testset	52.0	57.4	52.8	46.6	51.0	51.2	45.3	**51.0**	16.9	78.9	71.2

Table 2: Multi-person pose tracking results on PoseTrack Challenge dataset, * Note that this result was computed by online server on a subset of testset, and 51.8 MOTA is Girdhar *et al.* [8] got on full testset.

Method	Dataset	Head mAP	Shoulder mAP	Elbow mAP	Wrist mAP	Hip mAP	Knee mAP	Ankle mAP	Total mAP
Iqbal *et al.* [1]	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.* [8]	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.* [8]	*(Mini)Test v1.0	65.3	66.7	59.7	51.2	58.6	55.8	48.8	58.5
Ours	PoseTrack Challenge(test)	64.9	67.5	65.0	59.0	62.5	62.8	57.9	**63.0**

Table 3: Multi-person pose estimation results on all PoseTrack dataset,* Note that this result was computed by online server on a subset of test set, 59.6 mAP is Girdhar *et al.* [8] got on full testset.

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 100ms per frame only. Therefore, it will not be the bottleneck of whole system, comparing to the speed of pose estimation.

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 single person pose estimation 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.

4.3 Ablation Studies

We evaluate the effectiveness of four proposed components: Deepmatching based Motion-guided box propagation (DMGP), improved deep proposal generator (iDPG), Pose Flow Builder (PF-Builder) and Pose Flow NMS (PF-NMS). The ablative studies are conducted on the validation of PoseTrack Challenge dataset, by removing these modules from the pipeline or replacing them with naive solvers, i.e., we replace the PF based tracker with box IoU based maximum weight bipartite matching tracker (IoU-tracker) used by [8].

PF-Builder and PF-NMS PF-Builder is responsible for constructing pose flow. Due to its a global optimum solution, like Table 4 shows, it can guarantee better tracking performance than IoU-Tracker even without PF-NMS. PF-NMS can robustly merge redundant pose flows and re-link temporal disjoint ones, thus it can simultaneously polish pose estimation and tracking results by 1.9 mAP and 2.5 MOTA.

Method	mAP	MOTA	MOTP	Prcn	Rcll
PoseFlow, full	**66.5**	**58.3**	67.8	87.0	70.3
w/o PF-NMS	64.6	55.8	66.0	82.2	90.3
IoU-Tracker	64.6	52.1	61.2	82.2	90.3
w/o DMGP	62.2	53.7	63.4	89.2	62.3
w/o iDPG	65.4	57.8	66.9	87.0	70.3

Table 4: Ablation comparison. "IoU-Tracker" means naive box IoU based matching tracker used by [8]. "w/o PF-NMS" means only using PF-Builder without PF-NMS. "w/o DMGP" means removing motion-guided box propagation. "w/o iDPG" means without improved deep proposal generator.

DMGP and iDPG DMGP is used for propagating adjacent boxes bidirectionally to recover missing boxes, so this module can improve tracking performance 4.6 MOTA by decreasing IDs dramatically. Because high recall of detections can fully exploit the power of PoseNMS module in RMPE framework [7], 4.3 mAP is also increased thanks to this high recall. iDPG aims mainly to locate hard keypoints more accurately, because pose information is also leveraged during tracking, iDPG ultimately improve results by 1.1 mAP and 0.5 MOTA.

Figure 4: Some final posetracking results in videos

5 Conclusion

We have presented a scalable and efficient top-down pose tracker, which mainly leverages spatio-temporal information to build pose flow to significantly boost pose tracking task. Two novel techniques, Pose Flow builder and Pose Flow NMS were proposed in this paper. In ablation studies, we prove that the combination of PF-Builder, PF-NMS, iDPG, and DMGP can guarantee a remarkable improvement in pose tracking tasks. Moreover, our proposed pose tracker that can process frames in a video at 10 FPS (excluding pose estimation in frames) has great potential in realistic applications. In the future, we would like to analyze long-term action recognition and scene understanding based the proposed pose tracker.

References

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

[2] Mykhaylo Andriluka, Umar Iqbal, Anton Milan, Eldar Insafutdinov, Leonid Pishchulin, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. *arXiv preprint arXiv:1710.10000*, 2017.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, volume 1, page 7, 2017.

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *arXiv preprint arXiv:1711.07319*, 2017.

[5] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3218–3226, 2015.

[6] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. *arXiv preprint arXiv:1702.07432*, 1(2), 2017.

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[8] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. *arXiv preprint arXiv:1712.09184*, 2017.

[9] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. *arXiv preprint arXiv:1704.07333*, 2017.

[10] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[11] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. Arttrack: Articulated multi-person tracking in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 4327, 2017.

[12] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2017.

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

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

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[16] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Deepmatching: Hierarchical deformable dense matching. *International Journal of Computer Vision*, 120(3):300–323, 2016.

[17] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. *ArXiv170310898 Cs*, 2017.

[18] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3980–3989. IEEE, 2017.

[19] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE, 2013.

[20] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4305–4314, 2015.

[21] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017.

[22] Dong Zhang and Mubarak Shah. Human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2012–2020, 2015.

[23] Hong Zhang and Naiyan Wang. On the stability of video detection and tracking. *arXiv preprint arXiv:1611.06467*, 2016.

[24] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017.