HiEve ACM MM Grand Challenge 2020: Pose Tracking in Crowded Scenes

Lumin Xu
The Chinese University of Hong Kong
Hong Kong SAR, China
luminxu@link.cuhk.edu.hk

Ruihan Xu
Peking University
Beijing, China

Sheng Jin
Hong Kong University
Hong Kong SAR, China

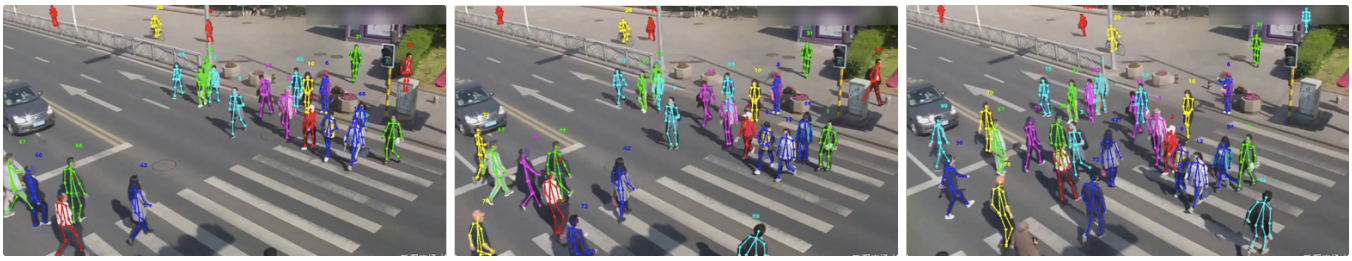

Figure 1: Crowd pose tracking results of our team *SimpleTrack* on HiEve test set.

ABSTRACT

This paper tackles the challenging problem of multi-person articulated tracking in crowded scenes. We propose a simple yet effective top-down crowd pose tracking algorithm. The proposed method applies Cascade-RCNN for human detection and HRNet for pose estimation. Then IOU tracking and pose distance tracking are applied successively for pose tracking. We conduct extensive ablation studies on the recently released HiEve crowd pose tracking benchmark. Our final model achieves 56.98 Multi-Object Tracking Accuracy (MOTA) without model ensembling on the HiEve test set. Our team SimpleTrack won the 3rd place in the ACM MM'2020 HiEve Challenge.

CCS CONCEPTS

• **Computing methodologies** → **Tracking**; **Object detection**.

KEYWORDS

Pose estimation; pose tracking; crowd

ACM Reference Format:
Lumin Xu, Ruihan Xu, and Sheng Jin. 2020. HiEve ACM MM Grand Challenge 2020: Pose Tracking in Crowded Scenes. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3394171.3416295

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3416295

1 INTRODUCTION

Human Pose Estimation (HPE) [5, 8, 11, 15, 17, 19, 22, 23, 25] is a very important and fundamental task in the field of computer vision, which localizes the human keypoints in the images or videos. Human pose tracking (or human articulated tracking) is more challenging, as it requires not only to estimate the human poses on each frame, but also to track human instances across frames (see Figure. 1). The development of human pose tracking can significantly promote higher-level applications, such as action recognition, scene analysis, and safety monitoring.

Recent multi-person tracking methods can be categorized into the top-down and bottom-up approaches. Top-down pose tracking approaches [9, 29, 31] first detect human bounding boxes, then perform pose estimation and human-level tracking. Bottom-up pose tracking approaches [7, 12–14, 16] first detect body keypoints and then group them into individuals across frames.

In this paper, we propose a simple top-down human tracking framework to estimate poses on each frame and associate persons of the same identity across frames. We first adopt Cascade R-CNN [4] with HRNet [25] backbone as our human detector to predict the bounding boxes of human candidates in each frame. Then, we estimate the localization of keypoints in each bounding box using HRNet [25]. Tracking is performed at the person-level. A simple but powerful method, IOU (Intersection over Union) tracking [3], is applied to associate persons across frames. IOU tracker computes the intersection over union between each pair of person bounding boxes on the adjacent frames. Then we greedily match the human instances according to the IOU of bounding boxes. If a person is not matched by the IOU tracker, we further apply a pose distance tracker to find out the matched pairs.

Evaluated on the HiEve test dataset, our team SimpleTrack won the 3rd place in the ACM MM'2020 HiEve Challenge [21], without using model ensembling.

2 METHOD

Our simple top-down human tracking framework consists of three main components, *i.e.*, the human detector, the pose estimator, and the human pose tracker.

2.1 Human Detector

Following the pipeline of top-down human pose tracking, we first detect all the person candidates for each frame. We apply Cascade-RCNN [4] with HRNet [25] backbone as our human detector and predict the human bounding boxes. Since HiEve [21] is a video dataset with 19 video sequences for training, the diversity of scenes is limited. Due to lack of training set diversity, the human detector will overfit in the HiEve training set. To avoid overfitting, we resort to combine multiple publicly available datasets. We first train the model on the MS-COCO dataset [20]. Then, the pre-trained model is finetuned on HiEve together with CrowdHuman [24] and CityPersons [32]. Multi-scale testing is used to further improve the performance.

2.2 Pose Estimator

We follow HRNet [25] to perform pose estimation in each human bounding box. HRNet maintains high-resolution representations through the whole process and is proved to achieve state-of-the-art performance for pose estimation. In order to increase the generalization ability of our pose estimator, we train the network on the HiEve dataset together with four additional well-known public multi-person pose estimation datasets, *i.e.*, MSCOCO [20], AI Challenger [28], MPII [2] and PoseTrack'18 [1]. We follow [10] to design a multi-head model for dataset-specific pose estimation. The backbone shares the common representation from multiple datasets, while the separate heads are specific for different domains. The model is first jointly trained with all datasets to learn generic features and then the heads of the model are finetuned on specific domains to further improve the pose estimation accuracy.

2.3 Pose Tracker

We conduct pose tracking at the person level. Each person on the first frame is assigned one unique person ID. Starting at the second frame, IOU tracking and person distance tracking are applied successively. The matched pairs are assigned the same person ID and a person without association is assigned a new person ID.

IOU Tracker [3]. IOU (Intersection over Union) is a classical measure of the relationship between two bounding boxes, which represents the proportion of intersection area to union area. As the assumption that the position change of the same person is small in adjacent frames, a person is assigned the same person ID as the one with the maximum IOU (if above threshold 0.3) in the last frame. In order to ensure the uniqueness of person ID on the current frame, the already matched person on the last frame will never serve as candidates any more.

Pose Distance Tracker. As the HiEve dataset is extremely challenging and includes frequent occlusion, IOU tracking may cause ID switch due to imperfect detection results. The distance of human poses is another convincing cue to decide whether two persons own the same identity across frames. We define the person distance as the average Euclidean distance of valid keypoint pairs. For persons not matched during IOU tracking, we greedily assign the person ID as the one with the minimum pose distance in the last frame if the minimum distance is smaller than 60 pixels. As the localization of keypoints may be inaccurate in challenging scenes, we ignore keypoints with confidence scores lower than 0.2.

3 EXPERIMENTS

3.1 Dataset and Evaluation Metrics

HiEve Dataset [21] consists of 32 video sequences, with 19 for training (HiEve-full-train) and 13 for testing (HiEve-test). Since there is no official validation set, we selected 4 representative video sequences (id 1, 12, 15 and 16) from HiEve-full-train as our validation set, named HiEve-4val. The rest 15 video sequences are used as the real training set (HiEve-15train).

Official evaluation code [1] is used for evaluating pose estimation and pose tracking performance. Total mean Average Precision (mAP) is used for multi-person pose estimation, and the standard Multi-Object Tracking Accuracy (MOTA) is used for pose tracking. For ablative experiments, the models are evaluated on the held out HiEve-4val dataset, if not specified.

3.2 Human Bounding Box Detection

We compare different design choices for human bounding box detection, in terms of backbones, datasets and test-time augmentation.

Backbones. We compare models with different backbones, namely ResneXt-101 [30] and HRNet [26]. Comparing #1 and #2 in Table 1, we find that HRNet backbone significantly outperforms the ResNext101 backbone (0.292 mAP vs 0.250 mAP). Both models are first trained on MS-COCO datasets [20] and then finetuned on HiEve-15train dataset, and evaluated on HiEve-4val dataset. Since the HiEve dataset has only 19 training videos and images in a video sequence are similar, the scene diversity of the dataset is somewhat limited. We find that fine-tuning on HiEve dataset for only 2 epochs will result in the best performance. And more training epochs will lead to overfitting and performance drop.

Datasets. We explore different combinations of training datasets for human bounding box detection. We experiment with two well-known benchmarks for pedestrian detection, *i.e.*CityPersons [32] (CP) and CrowdHuman [24] (CH). Both of them have crowded scenes as the HiEve dataset does. However, the dataset size of HiEve is much larger than that of CP and CH. To balance the data distribution, in our implementation, we repeat CP and CH for 8 times and then mingle with HiEve. Comparing #4, #5 and #6 in Table 1, we see that adding CrowdHuman improves the performance dramatically, while adding CityPersons has little effect.

Test-time augmentation. Comparing model#5 and model#6 in Table 1, we find that using multi-scale testing (MS) will further improve the human detection results from 0.329 to 0.345 mAP. We use three default scales for test-time augmentation.

3.3 Pose Estimation

We find that directly training the pose estimator on HiEve dataset can not achieve satisfying performance. We train our pose estimator, HRNet [25], with 4 additional publicly available pose estimation

[1] https://github.com/leonid-pishchulin/poseval

Table 1: Ablation study for human detection on HiEve-4val. 'CP' means using Citypersons dataset, 'CH' means using CrowdHuman dataset, 'MS' means multi-scale testing.

#	Method	mAP	AP^{50}	AP^{75}	AP^{m}	AP^{l}	AR^{100}
1	ResNeXt101	0.250	0.510	0.218	0.157	0.320	0.292
2	HRNet	0.292	0.584	0.256	0.185	0.373	0.347
3	#2 + CP	0.290	0.580	0.258	0.195	0.363	0.340
4	#2 + CH	0.330	0.617	0.311	0.233	0.402	0.379
5	#2 + CP + CH	0.329	0.608	0.305	0.233	0.399	0.374
6	#5 + MS	0.345	0.633	0.328	0.248	0.416	0.390

Table 2: Ablation study for human pose estimation on HiEve-4val. Performance reported with human detector#1.

#	Training	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total
1	w/o	35.4	54.6	52.8	54.9	55.6	54.5	53.9	50.6
2	Finetune	35.4	54.6	53.3	55.5	56.5	55.9	56.8	51.4
3	Joint.	35.6	54.9	53.5	55.8	56.2	55.8	55.9	51.4
4	Multi-domain	35.7	55.0	53.7	55.8	56.4	55.9	56.7	51.6

datasets, *i.e.*MSCOCO [20], AI Challenger [28], MPII [2] and PoseTrack [1].

We explore the usage of multiple training settings on pose estimation. As shown in Table. 2, model#1 is trained on 4 extra datasets only, without training on HiEve dataset. Model#2 is obtained by finetuning the pre-trained model#1 on HiEve-15train dataset and the total mAP improves 0.8 mAP (51.4 vs 50.6) on HiEve-4val. Model#3 is jointly trained on all 5 datasets. Model#3 outputs 19 keypoints, which is the intersection of joint types labeled on all these datasets. Model#4 follows [10] to use domain-specific heads. It outputs 33 keypoints, including 14 keypoints from HiEve Dataset, and 19 keypoints from other datasets. The results show that, Model#4 achieves the best performance and is applied to our final submission.

3.4 Pose Tracking

We track the poses by applying IOU tracking and pose distance tracking successively at person level after human bounding box detection and pose estimation. IOU tracking is simple but powerful, and it finds out most matched pairs. The localization of person keypoints is a convincing cue to decide the same identity across frames when the human bounding box detection results are not accurate, especially on the crowded scene.

The accuracy of human bounding boxes is important in our tracking method. Most person IDs are assigned by IOU tracker and the performance of IOU tracker is highly dependent on the accuracy of bounding boxes. We experiment on HiEve training set using the ground truth bounding boxes. 99.6% persons are assigned the correct person IDs applying IOU tracking. Most failure cases happen when persons brush past each other. Therefore, we choose Cascade RCNN with HRNet-w40 backbone as our detector and employ multi-scale testing, which achieves the best detection performance in our ablation study.

Besides, the accuracy of pose estimation can significantly influence MOTA. First, pose distance tracking tracks persons according to the localization of keypoints. Pose distance tracking assigns person IDs to person instances in several situations where IOU tracking can not tackle, such as several bounding boxes are very close or the bounding boxes are not accurate. Second, the MOTA metric applies Multiple Object Tracking (MOT) at keypoint level. Inaccurate localization of keypoints cause false negative, false positive and ID switch, which affects the performance of pose tracking significantly. As a result, we select the best pose estimator in our experiments.

Table 3: Effect of thresholding keypoints. Performance reported with human detector#6 and pose estimator#4.

Kpt Thre	Head mAP	Sho. mAP	Elb. mAP	Wri. mAP	Hip mAP	Knee mAP	Ank. mAP	Total mAP	Total MOTA	Total MOTP
0.0	40.2	60.0	58.9	61.3	59.2	58.1	58.8	**55.6**	52.5	81.6
0.1	39.0	58.6	57.5	60.1	58.0	57.2	58.0	54.4	61.1	82.2
0.2	37.5	56.5	54.5	58.1	55.7	54.9	56.4	52.3	**63.1**	82.8
0.3	35.7	54.0	50.8	55.6	53.1	51.6	54.5	49.7	61.9	83.5
0.4	33.2	51.0	46.3	52.6	49.9	48.0	52.4	46.7	58.9	84.1

However, owing to the imperfect human bounding box detection and pose estimation results, false positives damage the tracking performance. We deal with this problem by filtering out possibly inaccurate keypoints, poses or tracklets.

Thresholding keypoint scores. As shown in Table 3, the MOTA improves evidently when keypoints with extremely low confidence are deleted, though mAP drops a little. If the threshold is set too high, the MOTA drops as some accurate keypoints are filtered out and false negatives happen. We select the threshold as 0.2 to reach a compromise between false positives and false negatives.

Thresholding person scores. Besides, the pose with low confidence may come from a false positive of human bounding box detection. We filter persons with low pose confidence to reduce these cases. The pose confidence is defined as the average confidence of keypoints. The threshold is selected as 0.3 in our final submission as it achieves the best MOTA on HiEve-4val dataset as shown in Table 4.

Thresholding short tracklets. Due to several occlusion in crowds, some people can not be detected continuously in a video, which will lead to ID switches and decrease the performance of pose tracking. As shown in Table 5, deleting these short tracklets with no more than 3 frames will improve the MOTA of all the video sequences on the HiEve-4val dataset.

3.5 Implementation Details

Our human detector is trained on MMDetection [6][2]. We simply use the default training settings, using SGD with initial learning rate 0.02, momentum 0.9 and weight decay rate 0.0001. Our pose estimator is trained on MMPose[3] [27], using Adam [18] with initial learning rate 0.001.

3.6 Failure Case Analysis

We have analyzed some main failure cases in HiEve test set, see Figure. 2. Errors are marked with colored circles. We find that rare poses and view points, low image resolution, and occlusion may

[2]https://github.com/open-mmlab/mmdetection
[3]https://github.com/open-mmlab/mmpose

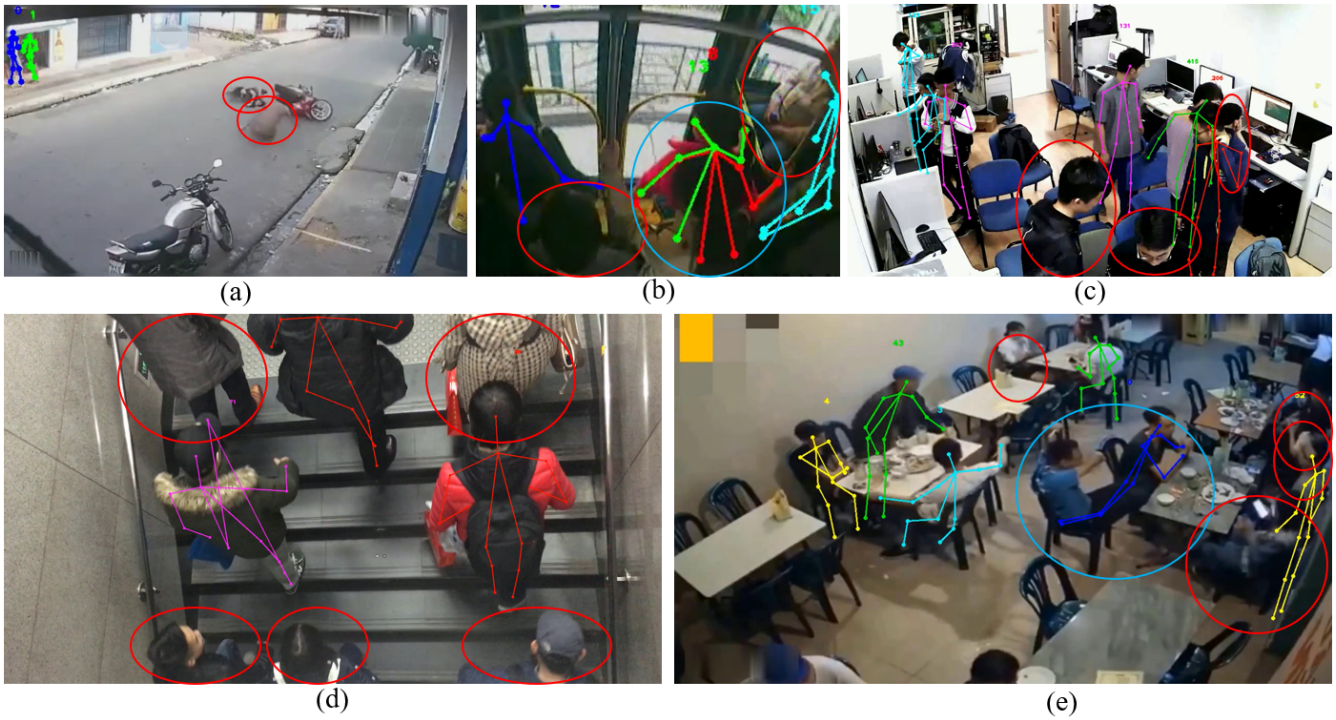

Figure 2: Common failure cases on HiEve test sets: (a) rare poses, (a)(b) low image resolution, (b)(d) rare perspectives (b)(c) occlusion (e) close proximity.

Table 4: Effect of thresholding people. Performance reported with human detector #6 and pose estimator #1.

Person Thre	Head mAP	Sho. mAP	Elb. mAP	Wri. mAP	Hip mAP	Knee mAP	Ank. mAP	Total mAP	Total MOTA	Total MOTP
0.0	39.9	59.4	58.6	60.0	59.1	57.3	58.5	**55.0**	40.8	82.1
0.1	39.8	59.2	58.5	59.9	59.0	57.1	58.3	54.9	43.2	82.2
0.2	39.4	58.4	57.8	59.0	58.0	56.2	57.2	54.1	47.6	82.5
0.3	38.4	56.4	56.5	57.1	55.2	53.7	54.0	52.1	**50.5**	83.2
0.4	37.3	54.2	54.8	54.3	50.9	50.2	49.6	49.3	41.6	83.7
0.5	35.5	50.7	51.9	50.3	46.4	46.1	45.0	45.8	39.8	84.6

Table 5: Effect of thresholding short tracklets. Performance reported with human detector #6 and pose estimator #4.

Length Thre	Video#1		Video#12		Video#15		Video#16	
	mAP	MOTA	mAP	MOTA	mAP	MOTA	mAP	MOTA
1	44.3	47.2	61.2	62.8	68.3	71.0	47.6	14.2
2	44.8	50.5	61.1	64.5	67.9	71.8	48.1	21.0
3	45.0	51.6	61.1	64.9	68.0	72.1	48.3	24.9
4	45.2	52.2	61.0	65.0	67.9	72.0	48.1	27.5

lead to false negatives. Close proximity of two human instances may also confuse the pose model.

Table 6: Crowd Pose Tracking in Complex Events on HiEve'20 Challenge test set.

Team Name	MOTA	MOTP	Total AP
Seedland.Tech	63.9686	52.8769	71.6342
Try	61.7941	54.9723	76.5478
SimpleTrack (Ours)	56.9834	55.9895	66.2943
DeepBlueAI	55.1543	52.3774	65.9403
Commander_test4	53.7671	57.4751	64.0834
PoseFlow [31] (Baseline)	44.1732	48.3287	60.1049

3.7 Final Performance on HiEve'20 Challenge

We evaluate our final performance on the test set, which was 56.98 MOTA. Our team SimpleTrack won the 3rd place in the ACM MM'2020 HiEve Challenge. Table 6 is copied from the leaderboard[4].

4 CONCLUSION

We have presented a simple yet effective approach to crowd pose tracking. Our approach builds upon the state-of-the-art frame-level human detection model (cascade R-CNN [4]), pose estimation model (HRNet [25]) and a simple pose tracker. The effectiveness of our design choices are validated through extensive ablative experiments. Our team SimpleTrack won the 3rd place in the ACM MM'2020 HiEve Challenge.

[4]http://humaninevents.org/oltp.html?title=4

REFERENCES

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

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Erik Bochinski, Volker Eiselein, and Thomas Sikora. 2017. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–6.

[4] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6154–6162.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008* (2018).

[6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155* (2019).

[7] Andreas Doering, Umar Iqbal, and Juergen Gall. 2018. Joint Flow: Temporal Flow Fields for Multi Person Tracking. *arXiv preprint arXiv:1805.04596* (2018).

[8] Haodong Duan, Kwan-Yee Lin, Sheng Jin, Wentao Liu, Chen Qian, and Wanli Ouyang. 2019. TRB: A Novel Triplet Representation for Understanding 2D Human Body. In *Proceedings of the IEEE International Conference on Computer Vision*. 9479–9488.

[9] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. 2017. Detect-and-Track: Efficient Pose Estimation in Videos. *arXiv preprint arXiv:1712.09184* (2017).

[10] Hengkai Guo, Tang Tang, Guozhong Luo, Riwei Chen, Yongchen Lu, and Linfu Wen. 2018. Multi-Domain Pose Network for Multi-Person Pose Estimation and Tracking. In *ECCV2018 posetrack workshop*.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. *arXiv preprint arXiv:1703.06870* (2017).

[12] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, Bernt Schiele, and Saarland Informatics Campus. 2017. ArtTrack: Articulated multi-person tracking in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Umar Iqbal, Anton Milan, and Juergen Gall. 2016. Pose-Track: Joint Multi-Person Pose Estimation and Tracking. *arXiv preprint arXiv:1611.07727* (2016).

[14] Sheng Jin, Wentao Liu, Wanli Ouyang, and Chen Qian. 2019. Multi-person Articulated Tracking with Spatial and Temporal Embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. 2020. Differentiable Hierarchical Graph Grouping for Multi-Person Pose Estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[16] Sheng Jin, Xujie Ma, Zhipeng Han, Yue Wu, Wei Yang, Wentao Liu, Chen Qian, and Wanli Ouyang. 2017. Towards Multi-Person Pose Tracking: Bottom-up and Top-down Methods. In *IEEE International Conference on Computer Vision Workshop*.

[17] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020. Whole-Body Human Pose Estimation in the Wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. 2019. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10863–10872.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[21] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Guo-Jun Qi, Rui Qian, Tao Wang, Nicu Sebe, Ning Xu, Hongkai Xiong, and Mubarak Shah. 2020. Human in Events: A Large-Scale Benchmark for Human-centric Video Analysis in Complex Events. arXiv:2005.04490 [cs.CV]

[22] Wentao Liu, Jie Chen, Cheng Li, Chen Qian, Xiao Chu, and Xiaolin Hu. 2018. A cascaded inception of inception network with attention modulated feature fusion for human pose estimation. In *The Thirty-Second AAAI Conference on Artificial Intelligence*.

[23] Alejandro Newell, Zhiao Huang, and Jia Deng. 2017. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*.

[24] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. 2018. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123* (2018).

[25] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5693–5703.

[26] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. 2019. High-Resolution Representations for Labeling Pixels and Regions. *arXiv preprint arXiv:1904.04514* (2019).

[27] Can* Wang, Sheng* Jin, Haodong Duan, Xuanyi Li, Zhizhong Li, Wentao Liu, Kai Chen, Chen Qian, and Dahua Lin. 2020. mmpose. https://github.com/open-mmlab/mmpose.

[28] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. 2017. AI challenger: a large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475* (2017).

[29] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[30] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1492–1500.

[31] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. 2018. Pose Flow: Efficient Online Pose Tracking. *arXiv preprint arXiv:1802.00977* (2018).

[32] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. 2017. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3213–3221.