Supplementary Material for PoseTrack21: A Dataset for Person Search, Multi-Object Tracking and Multi-Person Pose Tracking

Andreas Doering* [1] Di Chen* [2,3] Shanshan Zhang [2] Bernt Schiele [3] Juergen Gall [1]

[1] University of Bonn [2] Nanjing University of Science and Technology [3] MPI for Informatics

A. Dataset Statistics

We visualize additional statistics for the validation set of PoseTrack21 in Fig. 1. In particular, we illustrate the distribution of multi-object tracking (MOT) and multi-person pose tracking (Pose Tracking) instances within each frame in Fig. 1 a). Some sequences in PoseTrack21 contain more annotated bounding boxes than poses. For that reason, there are some cases in which the number of frames containing n persons (*i.e.* $n = 72$) is higher for Pose Tracking than for MOT. For illustrative reasons, Fig. 1 a) is scaled logarithmically. Additionally, we visualize the distribution of bounding box sizes in Fig. 1 b). In Fig. 1 c), we measure the number of annotated frames for MOT and Pose Tracking, respectively. PoseTrack21 follows PoseTrack 2018 [1], which provides densely annotated keypoints in the center of every sequences. The remaining frames are sparsely annotated with a step size of four frames. In contrast, PoseTrack21 provides annotated bounding boxes for every frame. Finally, we illustrate the total amount of persons for MOT and Pose Tracking, which have a maximum bounding box intersection over union in an interval of $[0, 1]$ in Fig. 1 d). This shows that the majority of persons occlude each other with an intersection-over-union of at least 20%.

B. Ablation Studies

We provide additional ablative analysis for multi-person pose tracking, multi-object tracking and person search.

Multi-Person Re-ID Pose Tracking We additionally analyzed the performance of our proposed baselines in terms of false-positives and MOTA score with respect to the size of the person bounding boxes. Fig. 3 a) shows the false positives rate for our baselines. As expected, the number of false positives decreases for larger bounding-boxes. *TrackTorWCorr* has significantly more false positives than the remaining baselines. This additionally confirms that keypoint correspondences are not a suitable motion model and that keypoint-based non-maximum suppression fails to remove duplicate detections reliably.

As highlighted in Fig. 3 b), all baselines achieve the highest MOTA scores for bounding box sizes between 800 and 1000 pixels, afterwards the performance drops significantly. The drop in performance might be caused by a weakness of our underlying person detector, which we used in all our Pose Tracking baselines. In particular, we used FasterRCNN [7] with FPN [5] as our person detector.

Multi-Object Tracking In a similar fashion, we analyzed the performance of our multi-object tracking baselines [3, 8, 9] and *CorrTrackWReid*. As shown in Fig. 4, MOT approaches show very similar performance patterns, compared to multi-person pose tracking baselines (Fig. 3).

Person Search Fig. 5 shows our ablative experiments for person search. In Fig. 5 a-c), we plot the number of matches, mAP and top-1 scores with respect to bounding box size. In Fig. 5 d), we plot the number of matches with respect to bounding box visibility (1 - IoU). Except for the top-1 score, the performance of all approaches degenerates for a bounding box sizes larger than 1100 pixels, which is similar to multi-person pose tracking and multi-object tracking. We conclude that the respective person detector models are not sufficiently trained on large bounding boxes, which can be caused by limited availability of large-scale persons during training. To our surprise, *SeqNet* [4] performs worse than related approaches for box sizes between 1000 and 1200 pixels, even though *SeqNet* utilizes a cascaded detection architecture.

C. Keypoint HOTA

HOTA [6] is an evaluation metric that has been originally proposed for the task of MOT. It tries to balance detec-

*equal contribution

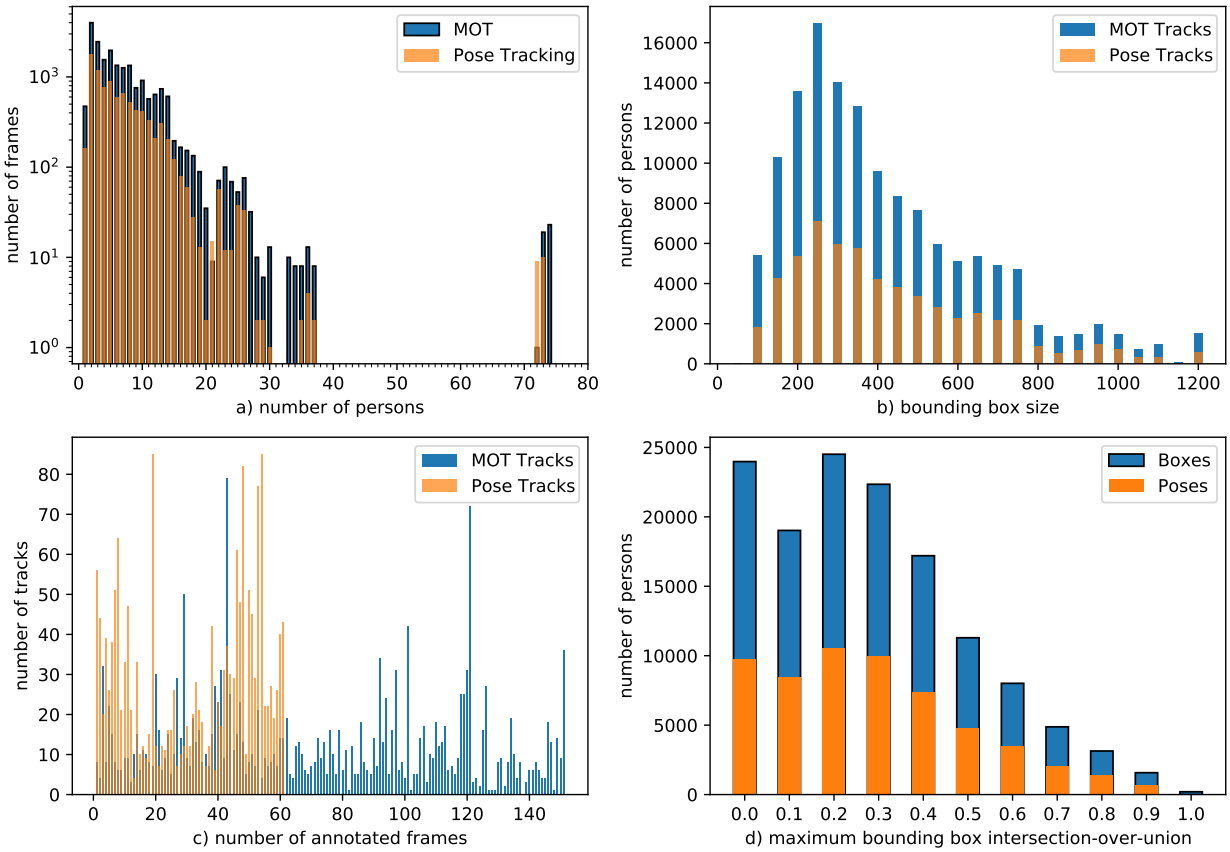

Figure 1. Statistics for the validation set of PoseTrack21. Best viewed with a PDF reader with zoom function.

tion accuracy and association accuracy of underlying tracks. For the association of detected tracks to ground truth tracks, HOTA utilizes a localization similarity, which relies on the IoU of bounding boxes. For the task of multi-person re-id pose tracking, we extend HOTA to keypoint HOTA. In particular, we rely on the head-normalized percentage of correct keypoints (PCKh) [2] as localization similarity. This allows us to measure the tracking performance on a keypoint level and we define the person-level tracking performance as the average tracking performance among all keypoint classes.

Furthermore, HOTA aims to maximize true positive assignments and prefers an assignment with minimal identity switches (IDSW) to an assignment with high localization accuracy. This is achieved by an additional global alignment score (cf. (15) in [6]). PoseTrack21 contains very challenging scenarios with high degrees of occlusions. For that reason, we aim to penalize re-identification errors strictly and do not use the global alignment score in keypoint HOTA, since it tolerates identification errors if two persons are close as illustrated in Fig. 2.

The source code of the evaluation metrics is available at https://github.com/andoer/PoseTrack21.

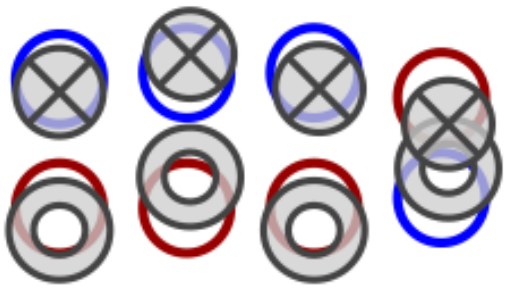

Figure 2. While the blue and red person (ground-truth) cross their path at the last frame, both tracks (cross, ring) keep straight on. In the last frame, the identities of both persons are therefore wrongly estimated, which is penalized by keypoint HOTA. HOTA [6] does not penalize this case if the cross overlaps with the blue circle and the ring with the red circle higher than a threshold α.

References

[1] M. Andriluka, U. Iqbal, E. Ensafutdinov, L. Pishchulin, A. Milan, J. Gall, and Schiele B. PoseTrack: A benchmark for

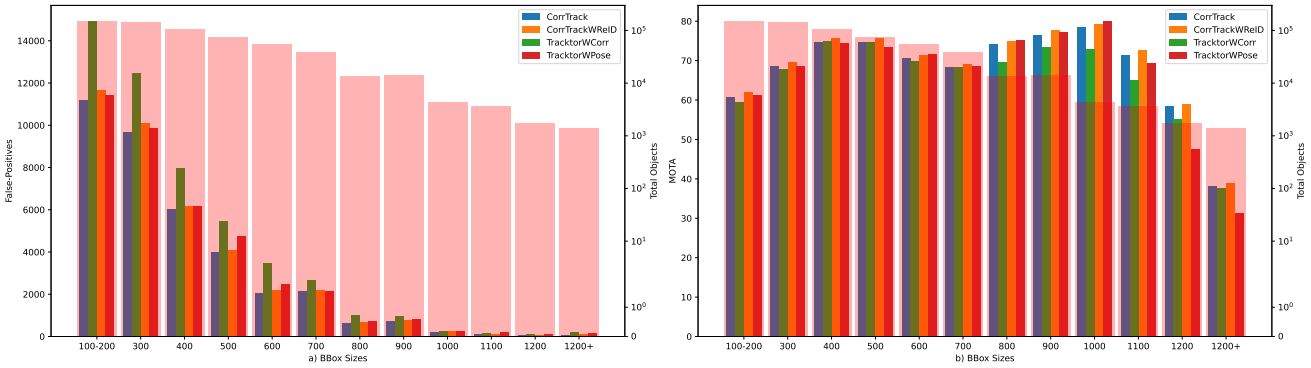

Figure 3. **Multi-person pose tracking**: Ablative evaluation of false positives and MOTA with respect to bounding box size. Best viewed with a PDF reader with zoom function.

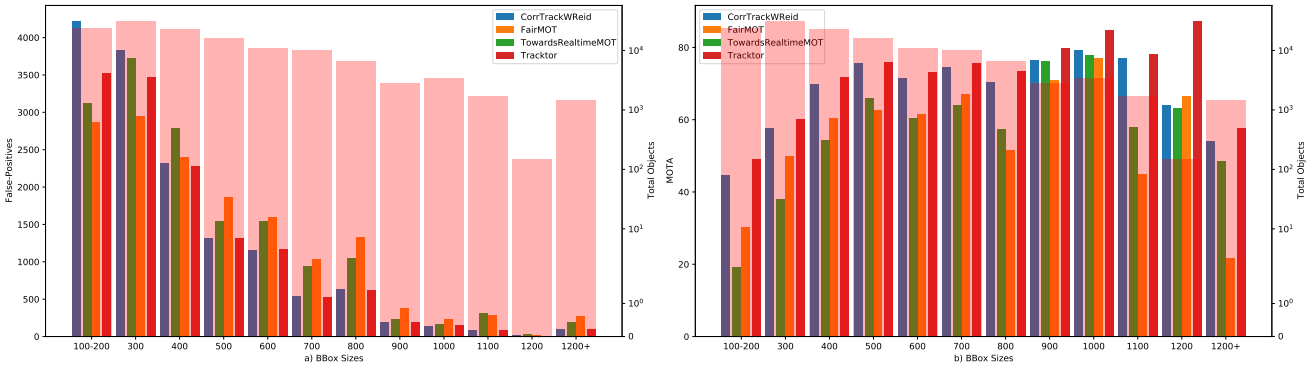

Figure 4. **Multi-object tracking**: Ablative evaluation of false positives and MOTA with respect to bounding box size. Best viewed with a PDF reader with zoom function.

human pose estimation and tracking. In *CVPR*, 2018. 1

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *ICCV*, 2019. 1

[4] Zhengjia Li and Duoqian Miao. Sequential end-to-end network for efficient person search. In *AAAI*, 2021. 1

[5] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[6] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, 2020. 1, 2

[7] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[8] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *ECCV*, 2020. 1

[9] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 2021. 1

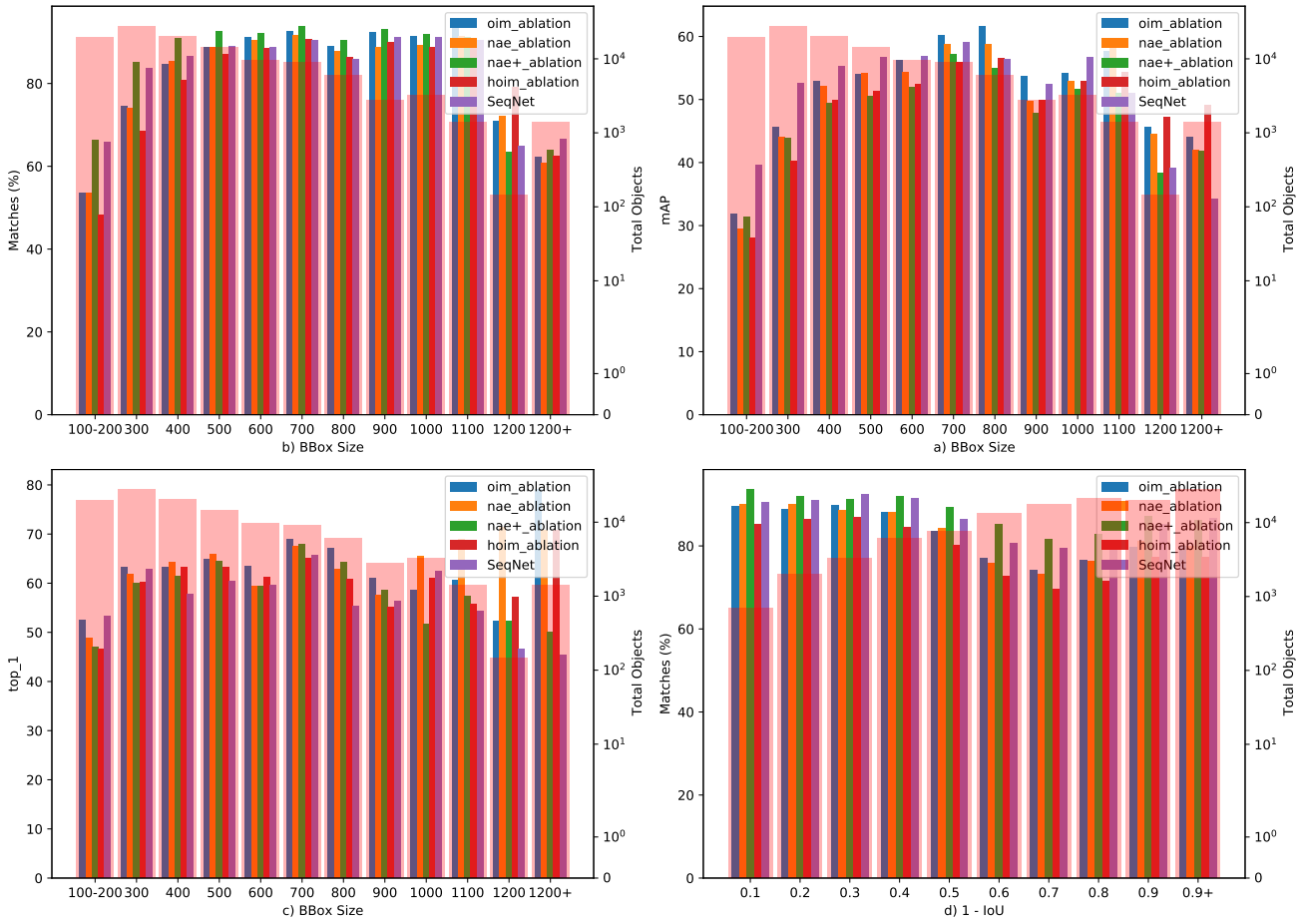

Figure 5. **Person search**: Ablative experiments with respect to bounding box size and visibility. Best viewed with a PDF reader with zoom function.