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

Abstract

Current research evaluates person search, multi-object tracking and multi-person pose estimation as separate tasks and on different datasets although these tasks are very akin to each other and comprise similar sub-tasks, e.g. *person detection or appearance-based association of detected persons. Consequently, approaches on these respective tasks are eligible to complement each other. Therefore, we introduce PoseTrack21, a large-scale dataset for person search, multi-object tracking and multi-person pose tracking in real-world scenarios with a high diversity of poses. The dataset provides rich annotations like human pose annotations including annotations of joint occlusions, bounding box annotations even for small persons, and person-ids within and across video sequences. The dataset allows to evaluate multi-object tracking and multi-person pose tracking jointly with person re-identification or exploit structural knowledge of human poses to improve person search and tracking, particularly in the context of severe occlusions. With PoseTrack21, we want to encourage researchers to work on joint approaches that perform reasonably well on all three tasks.*

1. Introduction

Multi-person pose tracking [30, 34, 36, 39, 43, 44], multi-object tracking [4, 5, 10, 37, 45, 50], and person search [7–9, 23, 41, 42] are very active research areas in computer vision and relevant for many applications like sports, autonomous driving, and security. Although these research areas share very common sub-tasks, they are currently studied independently and the progress in one area does not necessary yield a progress in the other areas. For instance, person search aims at re-identifying a query person within a set of target images or video frames. In contrast to person re-identification, person search also requires the detection of all persons in an image and it is more difficult due to inaccurate detections, missed detections, or false positives. Person search is thus a related sub-task of multi-object tracking where persons need to be detected in all frames and associated across frames for tracking. If a person is occluded for several frames, the association problem is similar to person search. Approaches for multi-object tracking thus often use a person re-identification module [17, 18, 27, 48, 49]. Multi-object tracking, however, considers only the identities of a person within a video sequence, but not across videos. In contrast, person search aims at identifying persons even across videos. Finally, multi-person pose tracking is related to multi-object tracking, but instead of estimating bounding boxes for each person the full pose needs to be estimated, including which joints are visible or not.

The reason why multi-person pose tracking, multi-object tracking, and person search are studied independently is the lack of a dataset that allows to evaluate all three tasks jointly on real video sequences. Datasets for multi-object tracking [11, 22, 29, 32] contain track-ids and bounding boxes, but no ground-truth for human poses and person-ids across videos. Datasets for multi-person pose tracking [1, 19] do not contain person-ids across video sequences either, but they also do not contain bounding boxes. While it is possible to compute bounding boxes from human poses, they are not accurate and reliable due to occlusions and truncations that occur highly frequently in these datasets. Synthetic datasets [13, 14] that are generated using the computer game engine GTA are an exception, but they are intended for training and cannot be used for evaluating the performance of approaches on real-world data. Furthermore, they lack the pose diversity of real videos, which also include diverse sport sequences. In this work, we close this gap and propose PoseTrack21, a large-scale dataset for multi-person pose tracking, multi-object tracking, and person search. It is based on the extended set of videos from the PoseTrack 2018 training and validation set [1], but the videos are completely re-annotated. Besides the refined human poses, the dataset contains accurate anno-

*equal contribution

PoseTrack 2018 PoseTrack21

Figure 1. Comparison between PoseTrack 2018 [1] and PoseTrack21. We densely annotated crowded scenes to increase the difficulty for multi-person pose tracking, multi-object tracking and person search. Ignore regions are drawn in red color. Best viewed in color with a PDF reader.

tations of joint occlusions, accurate bounding box annotations even for small persons, and person-ids within and across video sequences. The dataset can thus be used for evaluating approaches for multi-person pose tracking, multi-object tracking, and person search. Furthermore, the dataset allows to compare methods for multi-person pose tracking and multi-object tracking, which has been so far not possible due to missing bounding box or human pose annotations. We thus propose a few baselines where we combine techniques for multi-person pose tracking, multi-object tracking, and person search and address the question whether extending approaches for multi-object tracking to multi-person pose tracking or vice versa is more promising. We finally provide a detailed analysis regarding strengths and limitations of existing approaches and baselines. The dataset and source code of baselines are available at https://github.com/andoer/PoseTrack21.

2. Related Datasets

We discuss the most commonly used datasets for the tasks of multi-person pose tracking [1, 14, 19], multi-object tracking [11, 21, 22, 29, 32] and person search [21, 40, 47] and summarize the major differences in Tab. 1. Compared to previous multi-person pose tracking datasets, PoseTrack21 contains around 22% more human pose annotations per sequence. Further, we provide over $420,000$ additional bounding box annotations. Therefore, our dataset contains much more person instances compared to previous datasets for multi-person pose tracking. In addition, PoseTrack21 is the only dataset for the task of multi-person pose tracking, multi-object tracking, and person search since it provides continuous person identities throughout the entire dataset. In the following, we describe the other datasets more in detail.

Dataset	# Boxes	# Poses	track-ids	person-ids	# queries	real
MP PoseTrack [19]		16,219	✓			✓
PoseTrack 2017† [1]		80,144	✓			✓
PoseTrack 2018† [1]		144,688	✓			✓
MOT15 [22]	101,345		✓			✓
MOT17 [29]	292,733		✓			✓
MOT20 [11]	1,652,040		✓			✓
DukeMTMC◇ [32]	46,261		✓	✓		✓
PathTrack [28]	16,287 ‡	✓			✓	
CUHK-SYSU [40]	96,143			✓	2,900	✓
PRW [47]	34,304			✓	932	✓
P-DESTRE [21]	∼ 14.8 M			✓	253	✓
PoseTrack-ReID◇ [15]	84,443			∗		✓
PoseTrack21	428,949	177,164	✓	✓	1,313	✓
JTA [14]		∼10 M	✓			
MotSynth [13]		$\sim 40M$		✓		

Table 1. Comparison with different datasets for multi-person pose tracking, MOT and person search. Datasets marked with †: only training and validation set. ∗: no manual annotations. ◇: dataset not available. ‡: Only reports total number of tracks.

Multi-Person Pose Tracking Multi-person pose tracking datasets [1, 19] are large-scale datasets for multi-person pose estimation and tracking in videos. PoseTrack consists of challenging scenarios from multiple in-the-wild scenarios, such as sport and dancing, with a high degree of occlusion in many crowded scenes. The videos also strongly differ in terms of camera view and camera motion. The PoseTrack 2017 dataset [1, 19] follows the split of the MPII Human Pose dataset [2], which splits the dataset into 292, 50 and 208 videos for training, validation and testing. In total, PoseTrack 2017 provides around 23,000 labeled frames with 153,615 annotated poses, where each pose is annotated with 15 keypoints. PoseTrack 2018 extends the previous dataset and contains 593, 170 and 375 videos for training, validation and testing. The majority of the sequences range from 41 to 151 frames and contain 30 densely annotated frames around the middle of each sequence. In addition, validation and test sequences are densely annotated with a step size of four frames. In total, PoseTrack 2018 consists

of 46,933 labeled frames[1].

In Tab. 1, we provide the statistics for the extended version. As the test set is not publicly available, we can only report the number of total poses for the training and validation set. In addition to the human poses, PoseTrack provides ignore regions to exclude crowds and small people that have not been annotated, head bounding boxes to estimate the scale of a person, which is required for evaluation, and track identities. Unfortunately, track identities are not unique throughout the dataset, not even within a single sequence. If a person leaves a scene and re-enters, for instance, it is assigned a new track id. This can result in ambiguities for appearance-based similarity approaches.

Multi-Object Tracking The most important benchmark for multi-object tracking is the MOTChallenge[2], which consists of three separate tracking benchmarks: 2D MOT 15 [22], MOT16/17 [29] and MOT20 [11]. Each of the benchmarks consists of challenging person tracking sequences, mostly in surveillance scenarios or street-scenes with several degrees of occlusions and crowded scenarios. For instance, the MOT20 dataset is split into a training and a testing subset, each containing four sequences. Additionally, the dataset contains annotations of different object classes, such as cars, reflections or crowds, which are ignored during the evaluation. MOT20 sequences are mostly limited to surveillance scenarios in which the persons are in an upright position. The diversity of poses is thus much lower compared to datasets for multi-person pose tracking. Similar to PoseTrack, track identities are not always continued, *e.g.*, if a person leaves and re-enters the scene, a new track id is associated. Especially in very crowded scenarios, this results in a lot of identity switches, which potentially harms the training of appearance-based association methods, such as person re-identification.

DukeMTMC [32] is another dataset for multi-object tracking with a surveillance-based setup. Differently, DukeMTMC provides recordings of the same scene from different cameras and provides a benchmark for multi-target multi-camera tracking. For that reason, this dataset provides unique person identities throughout the entire dataset. Unfortunately, this dataset is no longer available due to ethics issues. Manen *et al*. [28] propose another large-scale MOT dataset called PathTrack, containing 720 sequences with a total length of 172 minutes and a total of 16287 person trajectories. Similar to PoseTrack, the dataset contains sequences of different categories, such as sports, dancing or street.

Person Search and Person Re-Identification Both, person search datasets [21, 40, 47] and video-based person re-identification datasets [21,38,46] are divided into query and gallery subsets. The query subset contains all persons of interest that have to be matched with the gallery subset. Query images are usually provided as tight person crops and mostly contain the person of interest only. Though, in many real world scenarios such as surveillance or sports, persons are often occluded by obstacles. Further, these scenarios contain a lot of crowded scenes in which persons frequently occlude each other. This leads to a lot of ambiguities, especially if a person is partially visible or multiple persons are present within a single crop. Contrary, PoseTrack21 provides queries of variable difficulty ranging from single person crops to highly occluded scenarios in which multiple persons are visible within a single crop. In this way, the dataset allows to study person search in a realistic setup and scenarios that differ from surveillance-like footage with limited pose variations. Fig. 2 shows a few examples of the queries of PoseTrack21. Further, we provide pose annotations for every query person, which can be used as additional guidance for person search.

Figure 2. From left to right: person search queries in PoseTrack21 with increasing difficulty.

[15] is another dataset based on the PoseTrack 2017 dataset, which was annotated for the purpose of video-based person re-identification. Based on the keypoint annotations, the authors calculated bounding boxes and removed persons with less than 6 keypoints. In contrast to PoseTrack21, the identities have not been annotated. Instead, person identities were obtained by an unsupervised approach without additional verification. The annotations are furthermore not available. Similarly, [6] also evaluates video-based person re-identification on PoseTrack and extracts tracklets from the PoseTrack 2018 videos. Both protocols [6, 15] are for person re-identification but not for person search.

Synthetic Datasets The Joint Track Auto (JTA) [14] dataset is generated from a video game. As other synthetic datasets, the dataset is intended for training, but it is not suitable for evaluating the performance of approaches on real-world data. Nonetheless, the dataset contains more then 10M annotated human body poses within over 460,888 densely annotated frames. The dataset contains 256 videos for the training and validation sets, respectively. Unlike

[1] On training and validation. The testing set can not be measured, as the annotations are not publicly available.

[2] https://motchallenge.net

PoseTrack, track ids are uniquely assigned within each sequence.

The MOTSynth [13] dataset is generated similarly from the same video game and combines 128 sequences from the JTA dataset with 256 newly scenes. All scenes were rendered with different weather conditions and during day and night, totaling 576 and 192 generated scenes for training and validation with more than 40M bounding boxes, over 1.3M annotated frames and 9519 unique person ids. On top, MOTSynth provides 3D poses, segmentation masks and depth information. In both datasets, the diversity of human poses, however, is very low compared to PoseTrack.

3. The PoseTrack21 Dataset

For creating PoseTrack21, we use the videos of the training and validation set of the extended version of the PoseTrack 2018 [1] dataset. It contains 593 videos for training and 170 videos for validation. The annotation has been performed in several steps. First, we annotated the bounding boxes. Since PoseTrack provides pose annotations only for some keyframes, we first annotated the bounding boxes for all keyframes. To this end, we visualized the annotated poses in PoseTrack for a frame and asked the annotators to annotate all persons where the head is visible since the head size is needed for the evaluation [3]. This also included persons that have not been annotated in PoseTrack 2018, which are in particular small persons and persons in crowds. The annotators were asked to draw a tight bounding box that covers the entire person, including occluded body parts. In a second step, we interpolated and manually revised the bounding boxes between the keyframes. In the third step, all bounding box annotations were verified by another person. In the fourth step, we marked ignore regions in each frame, which contain persons that have not been annotated. The ignore regions have been verified by a second person and ensure that a method, which makes a prediction for a person that has not been annotated, is not penalized. In the fifth step, head bounding boxes are annotated for the validation set. The head bounding boxes are required for the evaluation metric [3] and not required for training. In parallel, unique person identities throughout the videos of the training and validation set have been annotated. In a final step, we adapt and annotate person keypoints for all keyframes on the training and validation sets. The original annotations of the PoseTrack 2018 dataset consist of 15 keypoints, where each keypoint contains a flag, whether it is annotated or not. Unfortunately, there exist several cases in which these flags are not set reliably. In our dataset, we re-define the purpose of the keypoint flags and include occluded keypoints. In that way, pose estimation, re-ID and tracking approaches can utilize occlusion information within their training pipelines. We define a joint $j = (x, y, v)$ as occluded, if $x > 0$, $y > 0$ and $v = 0$. A joint is truncated if $x = 0, y = 0$ and $v = 0$. Otherwise, a joint is defined as visible if $v = 1$. After refining the original keypoints, we ran an off-the-shelf pose estimator [30] to estimate the poses for all newly added bounding boxes on keyframes only, which were then refined manually afterwards. All annotated poses have been verified and if necessary corrected by a second person. In total, 23 annotators worked on the dataset with over 16,000 person hours.

Person Search In the context of person search, we sampled 1313 person queries with varying sizes, camera motion and degree of occlusion. Especially in cases of occlusion, queries can contain multiple persons. As this results in ambiguities, we additionally provide keypoint information for the person of interest as shown in Fig. 2. In this way, we aim to encourage researchers to focus on more challenging scenarios for person search.

Data Format We provide different data formats for the respective tasks, which are very similar to the formats used in related datasets. For multi-person pose tracking and person search, we keep the format used in [1]. For multi-object tracking, we adopt the format proposed in [22]. This has the advantage that researchers from the different communities do not need to change their approaches and can easily read the annotations and save the results for evaluation.

4. Multi-Person Re-ID Pose Tracking

In the following, we describe the baselines that we propose for multi-person re-id pose tracking. The first baseline, which will be described in Sec. 4.1, builds on the approach [30] that we extend by including a person re-identification module to re-identify persons after occlusions or after they re-enter the scene. The second type of baselines, which will be described in Sec. 4.2, extends the multi-object tracking approach [4] to multi-person pose tracking.

4.1. Proposed CorrTrack Baselines

CorrTrack [30] is a multi-person pose tracking approach that utilizes a keypoint correspondence network. The approach comprises three steps. Given a new frame, the approach first detects the persons. For a fair comparison among the baselines, we use the same faster R-CNN object detector [31], which is also used in [4], instead of the detector that has been used in [30]. The approach then estimates the human pose for each detected bounding box using a pose estimator, which consists of multiple stages of an adapted GoogleNet [35]. We use the same pose estimator also for the other baselines for a fair comparison. In addition, the approach propagates the poses from the previous frame to the current frame using the keypoint correspondence network. Since occluded joints can become visible in the next frame, the approach re-estimates the poses for the propagated poses using the pose estimator. After applying a

Method	mAP	MOTA	MOTP	FP	IDSW	FN	TP
CorrTrack† [30]	72.0	62.6	**87.7**	**58922**	14896	164823	485634
CorrTrack [30]	72.3	63.0	87.3	59130	15272	161995	488484
CorrTrack [30] w. ReID	72.7	63.8	87.3	62604	9436	158720	491712
Tracktor++ [4] w. poses	71.4	63.3	87.3	59850	**8145**	166886	483558
Tracktor++ [4] w. correspondences	**73.6**	61.6	86.6	75663	20754	**147929**	**502588**
CorrTrack [30] (offline)	72.3	**63.9**	87.3	59132	9577	161997	488482

Table 2. Multi-person pose tracking baselines evaluated with the keypoint MOTA metric on PoseTrack21. Approaches marked with † have a model trained on PoseTrack 2018 [1].

non-maximum suppression on the poses of the new frame, the remaining poses are matched to the poses of the previous frame by bipartite graph matching where the similarity between two poses are measured by the affinity maps that are generated by the keypoint correspondence network. If a pose from the previous frame cannot be matched, the corresponding track ends. A new track starts if a pose of the new frame is not matched to a pose of the previous frame. This frame-wise matching results in a high number of identity switches. [30] also proposed an offline version where tracks can be merged in an additional post-processing step based on the keypoint correspondence similarity. We will report the results for the on-line and off-line variant.

In order to be able to track a person on-line and to reduce the number of identity switches due to occlusions or re-entering the scene, we keep a history of tracks that ended latest $T = 10$ frames ago. If a pose in the new frame is not matched to a track with a pose in a previous frame, we match the pose to the tracks in the history. Note that a track is only added to the history if it is inactive, *i.e.*, it does not contain a pose from the previous frame. For the matching, we use the SeqNet model [23], which we will also evaluate for the task of person search. We compute the re-identification features for the bonding box of the pose that has not been matched and the average re-identification features of the inactive track where we use at most the last T frames of the tracks. The similarity between a pose and an inactive track is then computed by the cosine similarity of the corresponding feature vector. The matching between all inactive tracks and unmatched poses is then performed using the Hungarian algorithm [20]. If the similarity of a match is higher than a threshold $\tau = 0.5$, the matched track is re-activated. The unmatched detections initiate new tracks. We denote this variant by *CorrTrack with ReID*.

4.2. Proposed Tracktor++ Baselines

Tracktor++ [4] is an on-line multi-object tracking approach, which is based on FasterRCNN [31]. During tracking, Tracktor++ aligns frames via image registration by enhanced correlation coefficient maximization [12]. For sequences with low frame rates, Tracktor++ additionally applies a constant velocity assumption for all tracked objects. By applying the respective motion models, bounding boxes of active tracks are then warped into the current frame. Further, warped bounding boxes are refined by the bounding box regression branch of [31]. After bounding box regression, warped boxes with a low confidence are removed and the respective tracks are deactivated. Furthermore, non-maximum suppression based on bounding box intersection over union is applied on all remaining warped and detected bounding boxes. In a second step, unmatched detections are associated with inactive tracks. For the at most last $T = 10$ bounding boxes of an inactive track, the average appearance features extracted from a re-identification model [17] are computed and compared to the appearance feature of each unmatched detection. As distance between two appearance feature vectors, the Euclidean distance is used. The remaining unassociated detections initiate new tracks. We extend the approach [4] towards multi-person re-id pose tracking in two ways. In the first setting, we evaluate Tracktor++ on PoseTrack21 without the constant velocity assumption. Additionally, we remove small tracks with less than three frames as they are likely to be false positives. Afterwards, we estimate the pose for each track with the pose estimation model from [30]. We denote this approach by *Tracktor++ with poses*.

In the second setting, we replace the motion model by a pose warping module which is based on the keypoint correspondence network from [30]. First, we warp the keypoints of the last pose of all active tracks into the next frame. Since occluded joints might become visible, we calculate a bounding box from the warped keypoints and use the bounding box regression module from [4] for bounding box refinement. Second, we re-estimate the poses with the pose estimation model. We perform tracking in a greedy fashion based on non-maximum suppression and pose similarity. The calculation of pose similarity between the warped track poses and estimated poses is performed as in [39]. Tracks that can not be associated become inactive. The association of unmatched detected poses with inactive tracks is done as before. We denote this approach by *Tracktor++ with correspondences*.

Method	mAP	HOTA	FA-HOTA	DetA	LocA	AssA	FragA
CorrTrack [30]	72.3	51.13	51.07	45.48	81.94	58.02	57.75
CorrTrack [30] w. ReID	72.7	**52.71**	**52.59**	**46.56**	81.93	60.21	59.66
Tracktor++ [4] w. poses	71.4	52.21	52.03	46.30	**81.95**	59.41	58.61
Tracktor++ [4] w. correspondences	**73.6**	48.90	48.43	44.67	81.26	54.05	52.02
CorrTrack [30] (off-line)	72.3	52.42	52.29	45.48	81.94	**60.93**	**60.37**

Table 3. Multi-person pose tracking baselines evaluated with the keypoint HOTA metric on PoseTrack21.

5. Analysis

We evaluate the performance of related state-of-the-art methods for the tasks of multi-person re-id pose tracking, multi-object tracking and person search on our proposed dataset and analyse strengths and weaknesses.

Evaluation Metrics In the context of multi-person pose tracking, we use the keypoint-based MOTA metric proposed in [1] for evaluation. Contrary to the standard MOTA metric [22] for multi-object tracking, keypoint-based MOTA evaluates the tracking performance for every keypoint class individually. In the context of PoseTrack21, this results in 15 different MOTA scores, which are then averaged into a final MOTA score. In general, the MOTA metric is highly impacted by the localization accuracy of keypoint detections [26]. Hence, better person detectors or pose estimators directly result in a stronger MOTA score.

HOTA [26], in contrast, tries to balance detection accuracy and association accuracy of underlying tracks. For that reason, HOTA consists of sub-metrics measuring the detection accuracy (DetA), the localization accuracy (LocA), the association accuracy (AssA) and the fragmentation accuracy (FragA). FragA penalizes heavily fragmented tracks and extends HOTA to a fragmentation-aware HOTA metric (FA-HOTA).

For the evaluation of multi-person re-id pose tracking, we propose keypoint HOTA and replace HOTA's localization similarity by the head-normalized percentage of correct keypoints (PCKh) [2]. Further, and in contrast to HOTA, we strictly penalize identification errors. For more details, we refer to the supplementary material.

Multi-Person Re-ID Pose Tracking For a fair comparison, we utilized the same person detector and the same pose estimation model for all our baselines. In particular, we use a FasterRCNN [31] with a Resnet50-FPN [24] as our person detector. The model is pre-trained on MSCOCO [25], which we obtained from the TorchVision model-zoo[3]. We further fine-tuned the person detector on PoseTrack21 for 30 epochs, following the training protocol proposed in [4]. Similarly, we train a three-stage pose estimation model from [30] on MSCOCO and PoseTrack21 for 215 and 16 epochs, respectively. The learning rate was reduced from 1e-3 to 1e-4 after 200 epochs and further reduced to 1e-5 after 4 epochs on PoseTrack21.

We report mAP for the pose estimation performance based on the PCKh-metric proposed in [1] and evaluate the tracking performance of our baselines with two different metrics, namely MOTA [1, 22] and HOTA [26]. Tab. 2 and Tab. 3 summarize the results. Note the difference in mAP: It is common practice in multi-person pose tracking approaches [30, 34, 36, 39, 43, 44] to sacrifice mAP for better MOTA scores. Consequently, differences in outlier handling result in different pose estimation results.

Nearly all proposed baselines outperform the on-line version of *CorrTrack* in terms of MOTA and HOTA. *CorrTrack w. ReID* even outperforms the off-line version of *CorrTrack* in terms of HOTA. This shows that a tracklet history in combination with appearance-based feature matching can boost the overall tracking performance. On the other hand, keypoint correspondences do not seem to work reliably as a motion model (*Tracktor++ w. correspondences*) for two reasons: 1) keypoint correspondences can only warp visible keypoints of the previous frames, which results in bounding boxes that do not cover the entire person. In consequence, the appearance features have a limited descriptiveness. 2) the keypoint similarity based non-maxima suppression (NMS) fails to remove duplicate detections which cover different keypoints. A look into Tab. 3 further confirms this behaviour: the association accuracy (AssA) and fragmentation accuracy (FragA) are much lower compared to the remaining baselines. *CorrTrack w. ReID* and *Tracktor++ w. poses* show a similar tracking performance, though *CorrTrack w. ReID* has a higher FragA and generates less fragmented tracks.

Further, we evaluated the performance of all baseline methods with respect to different attributes, such as size of the bounding box, bounding box visibility and the number of keypoints. The results are shown in Fig. 3. Bounding box size denotes the maximum side length of a given bounding box. As tracking results did not contain bounding box information, bounding boxes were generated from keypoints. Interestingly, all baselines achieve the best performance for bounding box sizes between 400-500 and 800-1000 pixels. In the person visibility study, where we measured the IoU between the person's ground truth bounding

[3] https://pytorch.org/vision/stable/models.html

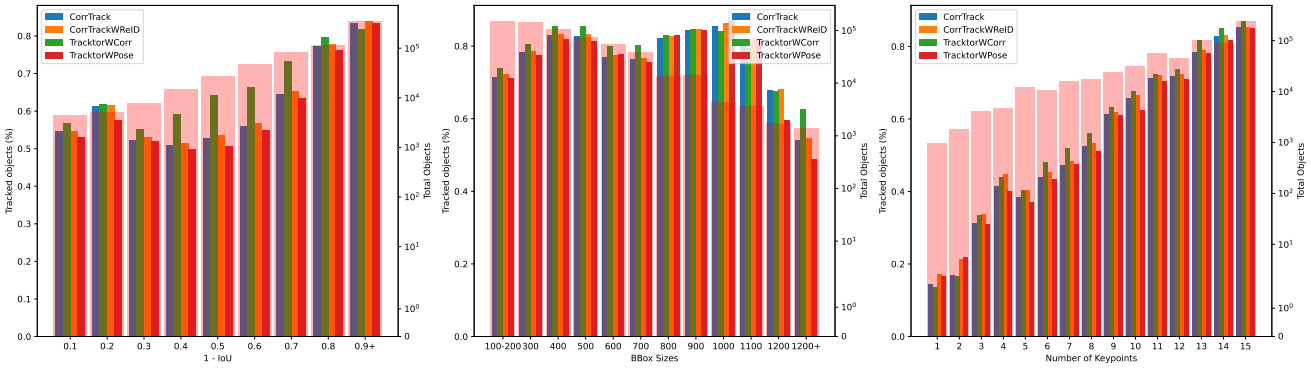

Figure 3. Pose tracking ablation studies. The plots show the recall (tracked objects) and number of objects for different bounding box visibility (1-IoU), size of the bounding box, and number of keypoints. Best viewed with a PDF reader and zoom function.

Method	IDF1	IDP	IDR	Rcll	Prcn	MOTA	MOTP
TRMOT [37]	57.3	70.0	46.6	59.2	85.5	47.2	75.4
FairMOT [45]	63.2	**81.0**	51.8	60.6	**94.9**	56.3	80.5
Tracktor++ [4]	69.3	76.4	**63.5**	**71.6**	86.2	**59.5**	**80.7**
CorrTrack + ReID	**66.5**	72.4	61.4	68.8	81.2	52.0	78.9

Table 4. Multi-object tracking baselines evaluated with the MOTA metric on PoseTrack21MOT.

Method	HOTA	DetA	AssA	LocA	RHOTA
TRMOT [37]	46.85	40.91	54.98	79.92	49.06
FairMOT [45]	53.53	47.43	61.45	83.16	55.37
Tracktor++ [4]	**58.29**	**52.71**	**65.43**	**83.09**	**62.58**
CorrTrack + ReID	56.95	51.33	64.19	82.80	61.86

Table 5. Multi-object tracking baselines evaluated with the HOTA metric on PoseTrack21MOT.

boxes and denoted visibility as 1 - IoU, all baselines perform surprisingly well in highly occluded scenarios (1-IoU $\in [0.1, 0.2]$). However, a closer look at the performance for different numbers of visible keypoints reveals that the recall massively drops when persons are only partially visible. Fig. 3 further reflects our previous conclusion on the performance of *TracktorWCorr* that achieves a higher recall at the cost of more false positives and identity switches.

Multi-Object Tracking We evaluate the performance of state-of-the-art multi-object tracking approaches [4, 37, 45] on our PoseTrack21 dataset. In particular, each baseline provided a pre-trained model on the MSCOCO dataset. We fine-tuned each baseline on PoseTrack21, following the training protocols as proposed in [4, 37, 45].

For *Tracktor++*, we additionally trained TriNet [17], a re-identification model based on ResNet50 [16] as proposed in [4]: For each minibatch, we sampled 18 crops of size 256×128 from the PoseTrack21-MOT subset and trained TriNet for 29270 iterations. Both *TRMOT* [37] and *FairMOT* [45] directly train a dedicated re-identification head within their proposed object detection networks, using the cross-entropy loss.

Similar to Sec. 5, we measure the performance with the MOTA and HOTA metrics. We additionally evaluate IDF1, IDP and IDR [33], which are common metrics to evaluate the MOT performance. We also report the recall-HOTA (RHOTA) [26], which combines detection recall and association accuracy.

In Sec. 5, we have evaluated the performance of a MOT baseline on the task on multi-person pose tracking. Consequently, we want to evaluate, how well multi-person pose tracking approaches can solve the task of MOT. In this regards, we calculated bounding boxes from all poses and removed all pose information from *CorrTrack w. ReID*, which we introduced in Sec. 4.1 and converted the results into the respective MOT format. As shown in Tab. 4 and Tab. 5, *CorrTrack w. ReID* achieves competitive results. This confirms that the MOT and pose tracking tasks complement each other.

In an additional set of ablative experiments, we evaluated the overall tracking performance for different features, such es bounding box size, visibility and number of keypoints similar to Sec. 5. Based on different bounding box sizes, *Tracktor++* significantly outperforms the remaining baselines in the amount of tracked objects, independent of the size. Note that the total number of objects is reported on a semi-logarithmic scale. A similar behaviour applies for the number of available keypoints. In terms of bounding box visibility (1 - IoU), *Tracktor++* struggles to properly track in occluded scenarios, in which *FairMOT* significantly outperforms the other baselines.

Person Search On our PoseTrack21-PersonSearch subset, we evaluate [7, 8, 23, 41] as our state-of-the-art baselines. All baselines rely on FasterRCNN [31] with a ResNet50 [16] backbone. Additionally, FasterRCNN is extended by an additional re-identification head. For *OIM* [41], we use the re-implementation of [8]. Following [8], we train *OIM* for 22 epochs on an image resolution of 900×1500 with

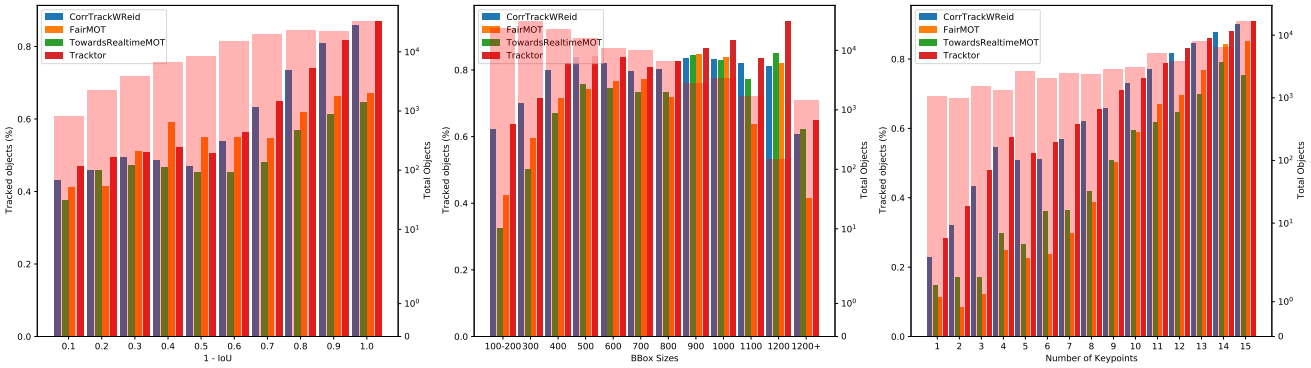

Figure 4. MOT ablation studies. The plots show the recall (tracked objects) and number of objects for different bounding box visibility (1-IoU), size of the bounding box, and number of keypoints. Best viewed with a PDF reader and zoom function.

Method	Recall	AP	mAP	top-1	top-5	top-10
OIM† [41]	68.05	63.84	57.58	**86.82**	88.65	89.49
NAE [8]	67.98	62.29	55.83	84.31	87.59	88.96
NAE+ [8]	74.45	62.56	54.48	83.93	87.89	89.11
HOIM [7]	**79.54**	66.08	52.77	84.69	87.66	88.35
SeqNet [23]	72.49	**69.36**	**65.12**	85.91	**91.01**	**92.23**

Table 6. Person search evaluation on PoseTrack21. Approaches marked with † where re-implemented [8].

a learning rate of 3e-3 which is additionally decayed by a factor of 10 after 16 epochs.

NAE [8] extends [41] with a norm-aware embedding head, which minimizes the embedding norm of background features towards 0 and maximizes the norm of person embeddings towards 1. We train NAE as proposed in [8] with the same training protocol used for *OIM*. *NAE+* [8] is a pixel-wise extension of NAE which is initialized with a pre-trained *NAE* network. *NAE+* is further fine-tuned for 11 epochs with a learning rate of 3e-3, which is decayed by a factor of 10 after 9 epochs. *HOIM* [7] extends [41] with a different re-identification loss, which considers multiple background embeddings as additional negative examples in combination with an InfoNCE loss [41]. We follow [7] and train the network similar to [41]. *SeqNet* [23], on the other hand, proposed a cascaded architecture for FasterRCNN [31]. In particular, *SeqNet* consists of a second bounding box regression head which refines the predicted bounding boxes of the first stage. Further, the second stage comprises a norm-aware re-identification head from [8]. We train *SeqNet* for 20 epochs with a learning rate of 3e-3, which is decayed by a factor of 10 after 16 epochs.

Due to a stronger person detection model, *SeqNet* outperforms all other baselines as shown in Tab. 6. Surprisingly, *OIM* outperforms the remaining baselines. We argue that the corresponding baselines are highly optimized on common person search datasets such as PRW [47] or CUHK-SYSU [40] in terms of hyperparameters, which we adopted from the respective datasets.

6. Discussion

In this work, we propose PoseTrack21, a joint dataset with annotated bounding boxes, human keypoints and person-ids, suitable for the task of multi-person pose tracking, multi-object tracking and person search. With this dataset, we want to encourage researchers to work on joint approaches, which can reliably solve multi-person pose tracking, multi-object tracking and person search. As we have shown in our experiments, MOT approaches in combination with a pose estimation model can be utilized as reliable baselines for multi-person pose tracking. Vice versa, the pose estimation baselines perform fairly well in the context of MOT. Person search models, on the other hand, usually rely on an object detector with an extended re-identification head which subsumes the exact baseline of modern MOT approaches. We belief that PoseTrack21will help to address the discussed limitations and increase the synergy between the three related, but separated research areas multi-person pose tracking, multi-object tracking and person search. Finally, it needs to be noted that the dataset will be only for research purposes and it will be forbidden to use the dataset for training or evaluating commercial surveillance systems or other systems that can potentially harm societies or individuals.

Acknowledgements We want to express our gratitude to Alexandra Splettstößer and to Yasaman Abbasi for their reliable work during the annotation process and quality control of the dataset. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GA 1927/8-1 and Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (61861136011).

References

[1] M. Andriluka, U. Iqbal, E. Ensafutdinov, L. Pishchulin, A. Milan, J. Gall, and Schiele B. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 1, 2, 4, 5, 6

[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, 6

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 4

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *ICCV*, 2019. 1, 4, 5, 6, 7

[5] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. *CVPR*, 2020. 1

[6] Di Chen, Andreas Doering, Shanshan Zhang, Jian Yang, Juergen Gall, and Bernt Schiele. Keypoint message passing for video-based person re-identification. *AAAI*, 2022. 3

[7] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. Hierarchical online instance matching for person search. In *AAAI*, 2020. 1, 7, 8

[8] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *CVPR*, 2020. 1, 7, 8

[9] Lequan Chen, Wei Xie, Zhigang Tu, Jinglei Guo, Yaping Tao, and Xinming Wang. Multi-attribute enhancement network for person search. *arXiv-Preprint*, 2021. 1

[10] Peng Dai, Renliang Weng, Wongun Choi, Changshui Zhang, Zhangping He, and Wei Ding. Learning a proposal classifier for multiple object tracking. *CVPR*, 2021. 1

[11] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian D. Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. MOT20: A benchmark for multi object tracking in crowded scenes. *CVPR*, 2019. 1, 2, 3

[12] Georgios D. Evangelidis and Emmanouil Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *TPAMI*, 2008. 5

[13] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljosa Osep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? *ICCV*, 2021. 1, 2, 4

[14] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *ECCV*, 2018. 1, 2, 3

[15] Xing Fan, Wei Jiang, Hao Luo, Weijie Mao, and Hongyan Yu. Instance hard triplet loss for in-video person re-identification. *Applied Sciences*, 2020. 2, 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv-Preprint*, 2017. 1, 5, 7

[18] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, 2020. 1

[19] Umar Iqbal, Anton Milan, and Juergen Gall. Pose-track: Joint multi-person pose estimation and tracking. *CVPR*, 2017. 1, 2

[20] Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *NRL*, 1955. 5

[21] A. Kumar, E. Yaghoubi, A. Das, B. Harish, and H. Proença. The p-destre: A fully annotated dataset for pedestrian detection, tracking and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security*, 2020. 2, 3

[22] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. 2015. 1, 2, 3, 4, 6

[23] Zhengjia Li and Duoqian Miao. Sequential end-to-end network for efficient person search. In *AAAI*, 2021. 1, 5, 7, 8

[24] 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. 6

[25] 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 *ECCV*, 2014. 6

[26] 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. 6, 7

[27] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. *ICCV*, 2019. 1

[28] S. Manen, M. Gygli, D. Dai, and L. V. Gool. Pathtrack: Fast trajectory annotation with path supervision. In *ICCV*, 2017. 2, 3

[29] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. 2016. 1, 2, 3

[30] Umer Rafi, Andreas Doering, Bastian Leibe, and Juergen Gall. Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In *ECCV*, 2020. 1, 4, 5, 6

[31] 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. 4, 5, 6, 7, 8

[32] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCVW*, 2016. 1, 2, 3

[33] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. *ECCVW*, 2016. 7

[34] Michael Snower, Asim Kadav, Farley Lai, and Hans Peter Graf. 15 keypoints is all you need. In *CVPR*, 2020. 1, 6

[35] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 4

[36] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *CVPR*, 2020. 1, 6

[37] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *ECCV*, 2020. 1, 7

[38] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018. 3

[39] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 5, 6

[40] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *CVPR*, 2017. 2, 3, 8

[41] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. *CVPR*, 2017. 1, 7, 8

[42] Yichao Yan, Jingpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-free person search. *CVPR*, 2021. 1

[43] Yiding Yang, Zhou Ren, Haoxiang Li, Chunluan Zhou, Xinchao Wang, and Gang Hua. Learning dynamics via graph neural networks for human pose estimation and tracking. In *CVPR*, 2021. 1, 6

[44] Rui Zhang, Zheng Zhu, Peng Li, Rui Wu, Chaoxu Guo, Guan Huang, and Hailun Xia. Exploiting offset-guided network for pose estimation and tracking. In *CVPRW*, 2019. 1, 6

[45] 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, 7

[46] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 3

[47] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *CVPR*, 2017. 2, 3, 8

[48] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, 2019. 1

[49] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *TPAMI*, 2021. 1

[50] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ECCV*, 2020. 1