JRDB-Pose: A Large-scale Dataset for Multi-Person Pose Estimation and Tracking

Edward Vendrow[1]*, Duy Tho Le[2]*, Jianfei Cai[2], Hamid Rezatofighi[2]
[1]Stanford University, [2]Monash University
evendrow@stanford.edu, {tho.le, jianfei.cai, hamid.rezatofighi}@monash.edu

Abstract

Autonomous robotic systems operating in human environments must understand their surroundings to make accurate and safe decisions. In crowded human scenes with close-up human-robot interaction and robot navigation, a deep understanding of surrounding people requires reasoning about human motion and body dynamics over time with human body pose estimation and tracking. However, existing datasets captured from robot platforms either do not provide pose annotations or do not reflect the scene distribution of social robots. In this paper, we introduce JRDB-Pose, a large-scale dataset and benchmark for multi-person pose estimation and tracking. JRDB-Pose extends the existing JRDB which includes videos captured from a social navigation robot in a university campus environment, containing challenging scenes with crowded indoor and outdoor locations and a diverse range of scales and occlusion types. JRDB-Pose provides human pose annotations with per-keypoint occlusion labels and track IDs consistent across the scene and with existing annotations in JRDB. We conduct a thorough experimental study of state-of-the-art multi-person pose estimation and tracking methods on JRDB-Pose, showing that our dataset imposes new challenges for the existing methods. JRDB-Pose is available at https://jrdb.erc.monash.edu/.

1. Introduction

Visual scene understanding of human environments is a difficult and crucial task for autonomous driving, human-robot interaction, safe robotic navigation, and human action recognition. Although rough predictions of human location are sufficient for some applications, a deep understanding of crowded human scenes and close-up human-robot interaction requires reasoning about human motion and body dynamics with human body pose estimation and tracking. Developing an AI model to predict human body pose is made more difficult by the varied and highly imbalanced range of human motion found in daily living environments, including a variety of scales, occlusions, and overlapping humans, representing a long-tailed distribution of human poses which is difficult for existing methods.

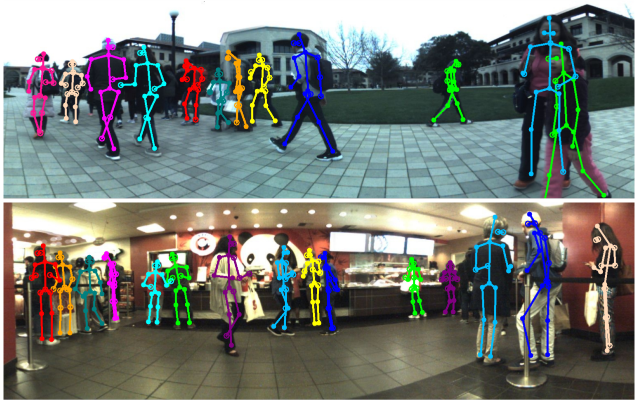


Figure 1. JRDB-Pose provides high frequency annotations of tracks and body joints in long scenes of crowded indoor and outdoor locations featuring dynamic motion and occlusion.

Human pose estimation and tracking is an active research area with many new large-scale datasets [13, 16, 26, 46] contributing to significant recent progress; however, these datasets do not primarily target robotic perception tasks in social navigation environments, and thus rarely reflect specific challenges found in human-robot interaction and robot navigation in crowded human environments, *e.g*. shopping malls, university campus, *etc*.

JRDB [28] previously introduced a large-scale dataset and a benchmark for research in perception tasks related to robotics in human environments. The dataset was captured using a social manipulator robot with a multi-modal sensor suite including a stereo RGB 360° cylindrical video stream, 3D point clouds from two LiDAR sensors, audio and GPS positions. JRDB [28] additionally introduced annotations for 2D bounding boxes and 3D oriented cuboids. Recently, JRDB-Act [15] further introduced new annotations on the JRDB videos for individual actions, human social group

*Equal contribution

Dataset	# Poses	# Boxes	Tracks	Crowd	ppF	Occlusion	Action	Indoor + Outdoor	Robot Navigation	Multi-Modal	Multi-Task
MPII [3]	40K				1-17			✓			
Penn Action [54]	160k	160k			1		✓	✓			
COCO [26]	250k	500k			1-20	✓		✓			✓
KITTI [18]		80k	✓	✓		✓				✓	✓
H3D [35]		460k	✓	✓						✓	
MOT20 [12]		1.65M	✓	✓		✓		✓			
THÖR [41]		2.5M	✓						✓	✓	
PoseTrack21 [13]	177k	429k	✓	✓	1-13	✓		✓			
Waymo [46]	173K	**9.9M**	✓	✓	*unk*	✓				✓	✓
JRDB-Pose	**636k**	2.8M	✓	✓	**1-36**	✓	✓	✓	✓	✓	✓
JTA† [16]	10M			✓	0-60	✓		✓			
MotSynth† [16]	40M		✓	✓	0-125	✓		✓			

Table 1. Comparison of existing public datasets related to 2D pose estimation and tracking. For each dataset we report the numbers of poses, boxes, as well as the availability of tracking information, crowd data, people per frame (ppF), occlusion labels, action labels, scene type, and if the data comes from robot navigation in human environments. We mark if a dataset has data modalities besides RGB frames, and if it contains annotations for multi-task types. Note that JRDB-Pose is a multi-modal dataset captured from a social navigation robot, addressing different research challenges than many existing works. †*Synthetic dataset. unk: Unknown.*

formation, and social activity of each social group.

JRDB was collected from a robotic navigation platform in crowded human environments, diversely capturing both indoor and outdoor scenes. Additionally since the robot's camera is located at person-level, and moves around, the data is not just collected from a far-off view but captures close-up scenes.

For robotic systems to safely navigate dynamic human environments and perform collision risk prediction, they must be able to accurately track and forecast motion of people in their surroundings. Human motion is often fast and requires high frame rate data for accurate prediction and tracking, making high-frequency annotated human pose data crucial for the development and evaluation of robotic perception systems in human environments. Complex social interactions add difficulty and similarly benefit from high-frequency data. In crowded scenes with high levels of occlusions or overlap with other humans, tracking may be also difficult.

We introduce JRDB-Pose, a large-scale dataset captured from a mobile robot platform containing human pose and head box annotations. JRDB-Pose provides 600k pose annotations and 600k head box annotations, each with an associated tracking ID. JRDB-Pose includes a wide distribution of pose scales and occlusion levels, each with per-keypoint occlusion labels and consistent tracking IDs across periods of occlusion. The combination of JRDB-Pose with JRDB and JRDB-Act forms a valuable multi-modal dataset providing a comprehensive suite of annotations suited for robotic interaction and navigation tasks.

Our contributions are:

- We introduce JRDB-Pose, a large-scale pose estimation and tracking dataset providing pose annotations and head boxes with tracking IDs and per-keypoint occlusion labels.
- In addition to adopting the popular metrics, we introduce new metrics, OSPA-Pose and OSPA$^{(2)}$-Pose for pose estimation and tracking, respectively.
- We conduct a comprehensive evaluation of state-of-the-art methods on JRDB-Pose and discuss the strengths and weaknesses of existing methods.

2. Related Datasets

We summarize the commonly used public datasets for pose estimation [3,26,26,47,54], pose tracking [13,46] and multi-object tracking [12, 18, 35, 41] in Tab. 1.

Single and Multi-person Pose Estimation Datasets: The task of 2D pose estimation involves predicting the pixel locations of human skeleton keypoints on an image. The MPII Human Pose Dataset [3] is a popular multi-person pose estimation dataset and benchmark including videos from everyday human activities, and the larger Penn Action dataset [54] provides both pose and action annotations in sports settings with a single pose per frame. MS COCO Keypoints Challenge [26] proposed a large-scale dataset including a diverse set of scenes and occlusion labels. All of these datasets label individual frames, limiting them to single-frame pose estimation. Human3.6M [20] annotates videos and single-person 3D poses from a controlled indoor scene, while 3DPW [47] provides estimated 3D human meshes in-the-wild with up to 2 people per frames. In comparison, JRDB-Pose is captured from a robotic platform in-the-wild with crowded and manually-annotated indoor and outdoor scenes, addressing a different set of challenges. JRDB-Pose also includes diverse data modalities including cylindrical video, LiDAR point clouds, and rosbags.

Multi-Person and Multi-Object Tracking Datasets: Multi-Person Pose Tracking (MPPT) and Multi-Object Tracking (MOT) are crucial tasks in robotic perception and navigation, where the challenge is to track, across a video, the body keypoints of individuals or the locations of objects,

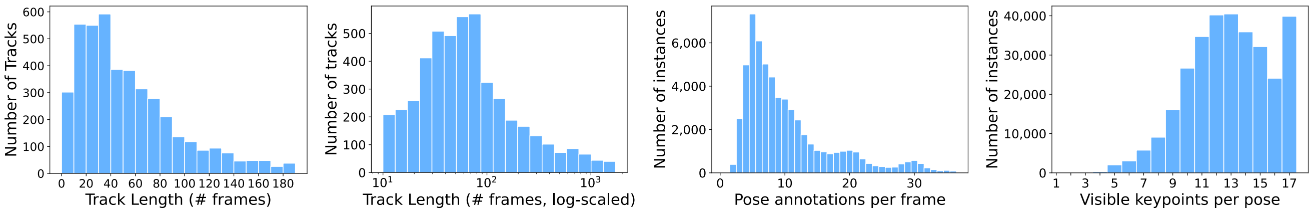
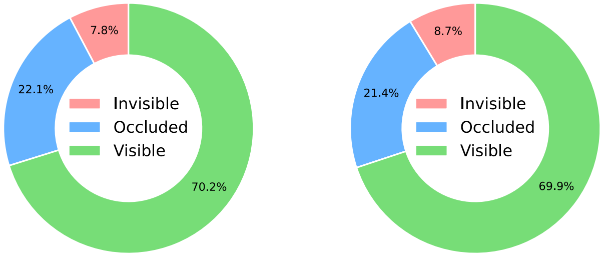


Figure 2. Various statistic for JRDB-Pose, which provides visibility labels and track ids consistent across long periods of occlusion. From left to right: *1)* The distribution of track lengths, with the long tail truncated. *2)* A log-scaled distribution showing all track lengths. JRDB-Pose track lengths are varied and as high as 1700 frames long. *3)* Number of pose annotations in each panoramic frame. *4)* Number of keypoints in each pose annotated as visible.



(a) Train set

(b) Test set

Figure 3. Distribution of keypoint visibility annotations in the JRDB-Pose train/validation and test splits. While most joints are visible, JRDB-Pose contains a large number of occluded and invisible joints.

respectively. MOT has attracted significant attention from the community with large-scale challenges [12, 25, 30], and MPPT is increasingly becoming recognized for its significance in human activity understanding and human-object interaction. Despite high performance on easy scenes, current MPPT methods struggle in crowded environments with occlusions and scale fluctuations.

Posetrack 2018 [2] introduced a benchmark for pose estimation and tracking using annotated video sequences featuring a variety of in-the-wild scenarios such as sports and dancing. Since PoseTrack was not densely labeled in crowded scenes, it required large "ignore regions" contains regions of images and videos lacking annotations. More recently, Posetrack21 [13] used the videos from PoseTrack 2018 to provide dense annotations of crowded scenes. Both datasets target diverse pose tracking scenarios, containing videos such as surveillance-like footage that are different from typical robotic navigation and human-robot interaction scenarios. Additionally, each video sequence is limited to five seconds, while sequences in JRDB-Pose are up to 2 minutes long. JRDB-Pose is a challenging dataset and benchmark for real-world tracking tasks, containing videos captured in crowded environments with humans fading into and emerging from the crowd.

Autonomous Driving Datasets: Recently, autonomous vehicle datasets have been released featuring large-scale and detailed annotations of the surrounding environment, posing difficult MOT and MPPT challenges. H3D [35] annotates human detections and tracks. More recently, the Waymo Open Dataset [46] filled the gap by providing 172.6K human pose annotations from their autonomous vehicle navigating outdoor road and highway environments. While the data is valuable for driving applications, the dataset contains exclusively outdoor road and highway environments making it unsuitable for non-driving robotic tasks, *e.g*. navigating a crowded shopping mall.

Robotic Navigation: Robotics operating in a social environment must learn to navigate in a socially-compliant manner by detecting and reacting to other agents in the environment. Existing datasets for robotic navigation [23, 28, 41] provide relatively few annotations and are often created in limited environment. The THÖR dataset provides human trajectories on a ground plane, and is filmed in a controlled indoor environment. SCAND [23] captures a variety of crowded indoor and outdoor environments without annotations for human detection or tracking. In human environments with complex interactions, it is crucial to understand human motion more deeply than a bounding box. To this end, JRDB-Pose is one of the first datasets providing large-scale pose annotations in a robotic navigation environment.

Synthetic Datasets: Several recent works have proposed large synthetic datasets for human pose estimation and tracking by generating data using a video game engine. The Joint Track Auto (JTA) [17] dataset generates 10M annotated human body poses, which MotSynth [16] extends to 40M human poses with associated track IDs. While synthetic datasets have been shown to be helpful to the existing frameworks' performance boost when combined with the real-world datasets [16], they may not reflect underlying biases and distributions in real-world data, preventing them from being the only training data resources for evaluating approaches in real-world applications.

State-of-the-art Frameworks: Multi-person pose estimation involves predicting body keypoint locations for all people in an image and identifying which keypoints belong to which individuals. There are two common types of approaches to this problem: top-down methods first detect all people and then execute pose estimation for each individual [24, 34, 49, 53], while bottom-up methods first identify

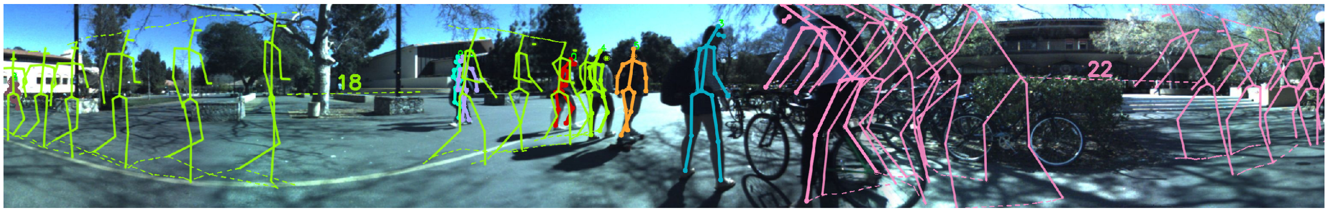


Figure 4. A portion of the panoramic frame with annotated pose instances. We show some of our tracking annotations by visualizing pose histories for two people. The gaps in the pose history correspond to periods of occlusion denoted by the numbers marking the length of the occlusion in frames. JRDB-Pose provides track IDs which are consistent even across long periods of occlusion.

keypoints and then group them into individual people [7], including disentangled representations [19], associative embedding [32], HGG [22], YOLO-Pose [27], and CID [48].

Multi-person pose tracking approaches [7, 34, 44, 53] often identify keypoints or poses using a pose estimation method, and then predict tracks using the estimated keypoints. Top-down techniques like openSVAI [34] and PGG [21] decompose the task into three discrete stages: human detection, pose estimation, and then pose tracking. Recently, UniTrack [50] proposed a unified framework for multiple object and keypoint tracking utilizing shared general-purpose appearance models.

The multi-object tracking task can be categorized into tracking by detection (TBD) [8, 14, 42, 51, 55] and joint detection and tracking (JDT) [5, 11, 29, 45, 52, 56, 57]. SORT-based [6] approaches [8, 14, 51] have gained popularity with methods including DeepSORT [51] and StrongSORT [14] In addition, OC-Sort [8] improved these methods by proposing an enhanced filter and recovery strategy suitable for tracking non-linearly moving and frequently occluded objects. Recently, ByteTrack [55] achieved efficient and precise tracking results using a general association technique that utilizes nearly every detection box. Popular JDT methods include Tracktor++ [5] which uses integrated detection-tracking modules and CenterTrack [56] which employs two-frames tracking algorithm for real-time and accurate tracking. Overall, since significant research in multi-person pose and multi-object tracking focuses on refining track consistency, occlusion recovery, and missing and false detection handling, JRDB-Pose is a useful and challenging benchmark for existing and future works.

3. The JRDB-Pose Dataset

We create JRDB-Pose using all videos from the JRDB dataset [28], including 64 minutes of sensory data consisting of 54 sequences captured from indoor and outdoor locations in a university campus environment. JRDB-Pose includes 57,687 annotated panoramic (3760 x 480) frames, containing total of 636k annotated pose instances with 11 million labeled keypoints with occlusion labels for each keypoint. We annotate 17 keypoints for each body pose including head, eyes, neck, shoulders, center-shoulder, elbows, wrists, hips, center-hips, knees and ankles. Each pose includes a tracking ID which is consistent for each person across the entire scene, including across long periods of full occlusion (see Fig. 4), and is also consistent with person IDs for existing 2D and 3D bounding box annotations from JRDB. Fig. 2 shows the distribution of track lengths present in JRDB-Pose. In total, JRDB-Pose provides 5022 pose tracks with an average length of 124 frames and some tracks as long as 1700 frames.

Annotations	Quantity	Track IDs
2D Body Bounding Box [28]	2,400,000	✓
3D Body Bounding Box [28]	1,800,000	✓
Atomic Action [15]	2,800,000	N/A
Social Group [15]	600,000	N/A
2D Head Bounding Box	600,000	✓
2D Body Pose	600,000	✓

Table 2. A summary of annotations in JRDB [28], JRDB-Act [15], and JRDB-Pose, which together provide a unique multi-modal dataset suitable for multi-person detection, pose estimation tracking, and activity recognition. Bolded annotations are introduced in this paper.

We started our annotation process using the JRDB bounding boxes and tracking IDs for each person in the scene. Annotators manually label human poses, and all the annotations were carefully quality-assessed to ensure high-quality and temporally consistent annotations. Details on our annotation protocol can be found in Appendix A. We annotated all persons from the 5 camera views of the 360° cylindrical video stream that had either a large enough bounding box (area $\geq$6000 pixels) or clear keypoint locations. For researchers interested in panoramic views, we merge the annotations from the 5 camera views into a single panoramic image. Note that this means not all visible people are labeled, especially those far away from the robot. Poses were annotated at 7.5 Hz and finally upsampled to 15 Hz, providing accurate high-frequency annotations without significant jitter.

JRDB-Pose labels poses for a wide range of scales, as shown in Fig. 6. We visualize the pose scales for each training and validation scene in Fig. 5, showing that the distribution of scales also varies significantly by scene as the types human motions and activities vary. Details on testing scenes

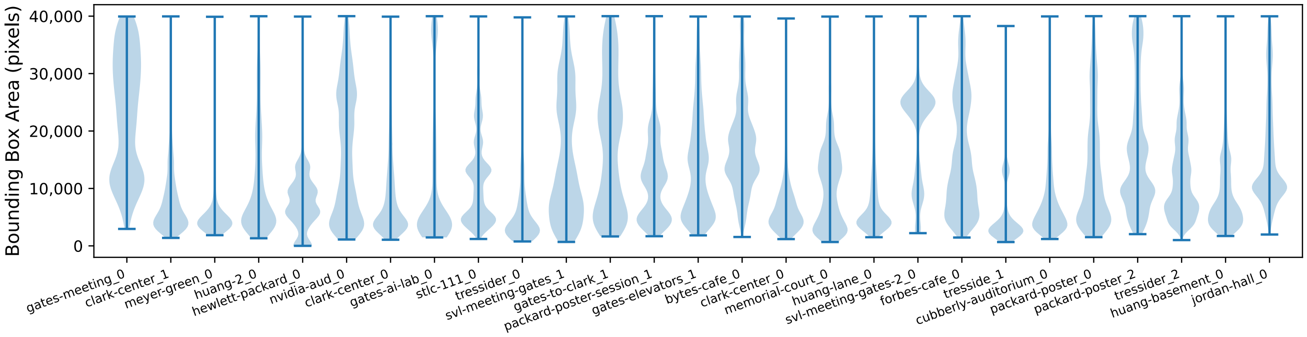


Figure 5. Distribution of bounding box scales in the train and validation scenes. JRDB-Pose contains a wide distribution of pose scale that is different for each scene.

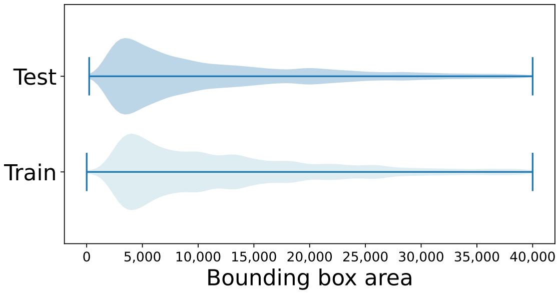


Figure 6. Bounding box distribution of the train/validation and test splits. JRDB-Pose contains a wide distribution of pose scales.

are not shown because test data is held out.

JRDB [28] and JRDB-Act [15] previously introduced annotations including 2D and 3D bounding boxes with tracking IDs, atomic action labels, and social groups. Together with JRDB and JRDB-Act annotations, JRDB-Pose is a multi-task learning dataset for human detection/tracking, pose estimation/tracking, individual action, social group, and social activity detection. Tab. 2 summarizes all annotations now available on videos from JRDB.

Occlusion: Occlusion is a key problem for human pose estimation and tracking because it causes incomplete information for predicting keypoints, leading to major errors in pose estimation tasks [33]. Significant research in the field of pose estimation has focused on occlusion-aware methods for improving pose estimation [1, 9, 10], but may be limited by lack of quantity and diversity in the existing data [9]. JRDB-Pose advances this research by providing large-scale data including occluded scenarios, as well as detailed occlusion labels for every keypoint which we hope will be useful for quantifying and improving performance under occlusion. In particular, we assign each keypoint a label in $\{0, 1, 2\}$ defining its occlusion type, where for every pose we annotate the posistion and occlusion of each keypoint:

ID	Meaning	Description
0	Invisible	The joint is out of frame or especially difficult to annotate, with a location inferred from context by annotators.
1	Occluded	The joint is obscured by an object or another body part, but its location is apparent from the image context.
2	Visible	The joint is fully visible and in view of the camera.

Figure 3 shows the occlusion label distribution in the train and test splits. JRDB-Pose contains a large number of labeled occluded and invisible joints, making it suitable for pose estimation under occlusion. Occlusion varies by scene: Fig. 7 shows that the training scenes of JRDB-Pose contain varied distributions of keypoint visibility, from densely populated indoor to dynamic outdoor scenes in which the robot navigates and interacts with oncoming pedestrians.

3.1. Benchmark and Metrics

JRDB-Pose Splits: We follow the splits of JRDB [28] to create training, validation, and testing splits from the 54 captured sequences, with each split containing an equal proportion of indoor and outdoor scenes as well as scenes captured using a stationary or moving robot. All frames from a scene appear strictly in one split.

3.1.1 Pose estimation

OKS. We define the ground-truth and a corresponding predicted human body keypoints as $x_i \in \mathbb{R}^{2\times J}$ and $y_j \in \mathbb{R}^{2\times J}$, respectively, where J represents the number of the keypoints. We measure a similarity score between x_i and y_j via Object Keypoint Similarity (OKS) [26] as:

$$OKS(x_i, y_j) = \exp\left(-\frac{d_E^2(x_i, y_j)}{2s^2k^2}\right) \quad (1)$$

where $d_E(x_i, y_j)$ is the mean Euclidean distance between two set of keypoints normalized by the product of ground-truth box area s and a sigma constants k. While this metric is commonly used for single-person pose estimation, in multi-person settings the assignment between ground-truth and predicted poses is not known, so it is unclear which pose pairs to use without a matching mechanism.

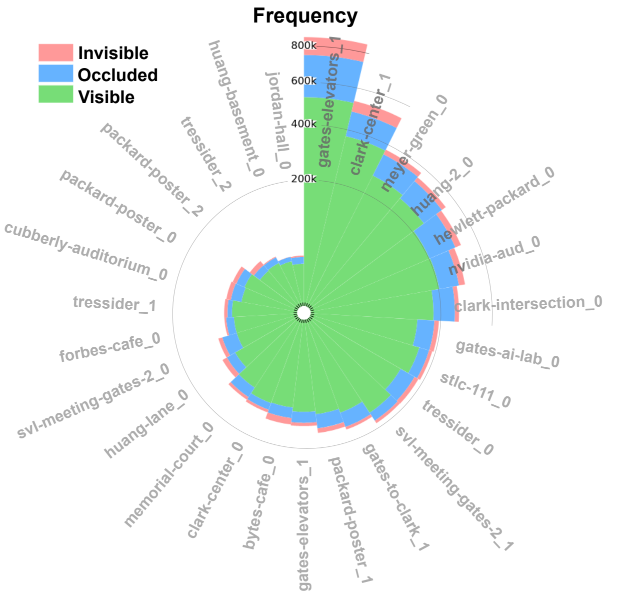


Figure 7. Distribution of keypoint visibility annotations for each JRDB-Pose train and validation scene. Best viewed in color.

Average Precision. Average Precision (AP) and mean AP (mAP) [36] are among the most common multi-person pose estimation metrics. Similarity between predicted and ground-truth poses are calculated via OKS [26], and poses are matched via a non-optimal greedy assignment strategy that matches poses with higher confidence scores first. True and false positive predictions are determined by a threshold on the OKS score. Since the resulting AP score corresponds to the performance at a specific OKS threshold rather than a continuous value based on overall keypoint distances, the COCO benchmark [26] averages AP across multiple OKS thresholds to reward methods with higher localization accuracy. Nevertheless, this strategy does not fully capture keypoint distance nor is the matching strategy optimal. We evaluate AP using an OKS threshold of 0.5.

OSPA-Pose. Optimal Sub-Pattern Matching (OSPA) [43] is a multi-object performance evaluation metric which includes the concept of miss-distance. Recently, OSPA has been further adapted to detection and tracking tasks [38] while preserving its metric properties. We propose OSPA-Pose ($\mathcal{O}_{pose}$), a modification of the OSPA metric suited to multi-person pose estimation.

Let $X = \{x_1, x_2, \dots x_m\}$ and $Y = \{y_1, y_2, \dots y_n\}$ be two sets of body keypoints for all ground-truth and predicted body poses, respectively, with $|Y| \geq |X|$ (otherwise flip X, Y). Considering $d_K(x_i, y_i) = 1 - OKS(x_i, y_i) \in [0, 1]$ as a normalized distance metric between human poses, OSPA-Pose ($\mathcal{O}_{pose}$) error is calculated by

$$\mathcal{O}_{pose}(X,Y)=\frac{1}{n}\left(\min_{\pi\in\Pi_n}\sum_{i=1}^{m}(d_K(x_i,y_{\pi_i}))+(n-m)\right), \quad (2)$$

where $n \geq m \geq 0$, Π_n is the set of all permutations of $\{1, 2, \dots, n\}$, and $\mathcal{O}_{pose}(X, Y)$ = $\mathcal{O}_{pose}(Y, X)$ if $m > n$. We further define $\mathcal{O}_{pose}(X, Y) = 1$ if either X or Y is empty, and $\mathcal{O}_{pose}(\emptyset, \emptyset) = 0$.

While both AP and $\mathcal{O}_{pose}$ use OKS to calculate pose similarity, $\mathcal{O}_{pose}$ measures an optimal distance from the set of continuous keypoint distances consisting of the localization error (first term) and cardinality mismatch (second term), eliminating the need for thresholding.

3.1.2 Pose tracking metrics

Commonly used metrics: We evaluate pose tracking performance with three commonly used tracking metrics, **MOTA** [2], **IDF1** [40], and **IDSW** [31], with modifications to make them suitable for the pose tracking task. Rather than using IoU or GIoU [39] to calculate similarity scores which are more appropriate for bounding boxes, we apply OKS (as defined in Eq. (1)) to obtain a similarity score between keypoint sets, with any keypoint pair of OKS above a threshold of 0.5 considered a positive prediction.

OSPA$^{(2)}$-Pose Inspired by [4, 38] extending the OSPA metric for evaluating two sets of trajectories, we propose a new metric for evaluating two sets of human body pose tracks, namely OSPA$^{(2)}$-Pose ($\mathcal{O}^2_{pose}$).

Let $\mathbf{X} = \{X_1^{\mathcal{D}_1}, X_2^{\mathcal{D}_2}, \dots X_m^{\mathcal{D}_m}\}$ and $\mathbf{Y} = \{Y_1^{\mathcal{D}_1}, Y_2^{\mathcal{D}_2}, \dots Y_n^{\mathcal{D}_n}\}$ to be two sets of body keypoint trajectories for ground-truth and predicted body poses, respectively. Note $\mathcal{D}_i$ represents the time indices which track i exists (having a state-value). Then, we calculate the time average distance of every pair of tracks $X_i^{\mathcal{D}_i}$ and $Y_j^{\mathcal{D}_j}$:

$$\underline{\widetilde{d}}(X_i^{\mathcal{D}_i}, Y_j^{\mathcal{D}_j}) = \sum_{t\in\mathcal{D}_i\cup\mathcal{D}_j} \frac{d_O\left(\{X_i^t\}, \{Y_j^t\}\right)}{|\mathcal{D}_i \cup \mathcal{D}_j|}, \quad (3)$$

where $t \in \mathcal{D}_i \cup \mathcal{D}_j$ is the time-step when either or both track presents. Note that $\{X_i^t\}$ and $\{Y_j^t\}$ are singleton sets, *i.e.* $\{X_i^t\} = \emptyset$ or $\{X_i^t\} = x_i^t \in \mathbb{R}^{2\times J}$ and $\{Y_j^t\} = \emptyset$ or $\{Y_j^t\} = y_j^t \in \mathbb{R}^{2\times J}$. Therefore, inspired by the OSPA set distance, $d_O\left(\{X_i^t\}, \{Y_j^t\}\right)$ can be simplified into the following distance function, :

$$d_O(\{X_i^t\},\{Y_j^t\}) = \begin{cases} d_K(x_i^t, y_j^t) & \text{if } |\{X_i^t\}| \wedge |\{Y_j^t\}| = 1, \\ 1 & \text{if } |\{X_i^t\}| \oplus |\{Y_j^t\}| = 1, \\ 0 & \text{Otherwise,} \end{cases} \quad (4)$$

where $\wedge$ and $\oplus$ are boolean operators AND and OR, $d_K(x_i, y_i) = 1 - OKS(x_i, y_i) \in [0, 1]$.

Finally, we obtain the distance, $\mathcal{O}^2_{pose}$, between two sets of pose tracks, *i.e.* $\mathbf{X}$ and $\mathbf{Y}$ by applying another OSPA distance over Eq. (3), *i.e.*

$$\mathcal{O}^2_{pose}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n}\left(\min_{\pi\in\Pi_n}\sum_{i=1}^{m}\widetilde{\underline{d}}(X_i^{\mathcal{D}_i}, Y_{\pi_i}^{\mathcal{D}_{\pi_i}}) + (n-m)\right), \qquad (5)$$

where $n \geq m \geq 0$, Π_n is the set of all permutations of $\{1, 2, \ldots, n\}$, and $\mathcal{O}^2_{pose}(X, Y) = \mathcal{O}^2_{pose}(Y, X)$ if $m < n$. We further define $\mathcal{O}^2_{pose}(X, Y) = 1$ if either X or Y is empty, and $\mathcal{O}^2_{pose}(\emptyset, \emptyset) = 0$.

Note that the first term $\widetilde{\underline{d}}$ reflects ID switches and localization errors, whereas the cardinality error $(n - m)$ contains false and missed track errors. In Tab. 5, we also present the $\mathcal{O}^2_{pose}$ per occlusion level, where each occlusion level is considered individually. We do not apply localisation error to joints that do not belong to the occlusion of interest by setting their Euclidean distance to zero, indicating that $d_E(x_i, y_j)$ in Eq. (1) is now the average Euclidean distance between two sets of keypoints of a certain occlusion level.

4. Multi-Person Pose Estimation and Tracking Baselines

In this section we evaluate the performance of various state-of-the-art methods for pose estimation and tracking, verifying that JRDB-Pose is a challenging dataset for existing frameworks seeking an opportunity for a dedicated developments in this domain. All methods are evaluated on our individual images and annotations.

4.1. Multi-Person Pose Estimation

We evaluate several popular or recent state-of-the-art methods for multi-person pose estimation models [19, 27, 48, 49]. We evaluate one top-down method using HRNet backbone [49]. The top-down method uses a Faster R-CNN [37] detector to predict all humans in the scene, from which the high-confidence (score > 0.95) predictions are kept. We use a heatmap size of 192x256, and the HRNet backbone with a high-resolution convolution width of 48. We further evaluate three recent bottom-up models, which regress joint locations directly without the need for human detections: DEKR [19], CID [48], an YoloPose [27]. All methods are trained from their respective initializations without COCO pre-training. To help address training difficulties associated with wide panoramic images, we train on individual camera images, and then combine them together to form stitched view predictions. Duplicate poses in the stitched annotation set are eliminated using NMS on the predicted boxes.

Tab. 3 summarises the pose estimation results for each method. To highlight the large number of occluded and invisible labels in our dataset, we include the total contribution of joints of each visibility type to the overall OSPA-Pose as well as average contribution for joints of each visibility. We also include cardinality and localization errors of OSPA-Pose. We find that Yolo-Pose is the best performing baseline, outperforming the other methods in $\mathcal{O}_{pose}$ and AP. DEKR achieves the lowest $\mathcal{O}_{pose}$ localization error but a higher cardinality error as a result of a high number of missed detections. In Tab. 4 we see the contributions to the localization error based on keypoint visibility. DEKR achieves the lower metrics due to its low localization error, besides which Yolo-Pose again performs best. Although visible joints contribute the most to localization error due to their higher overall frequency, on a per-keypoint average the predictions on occluded and invisible joints showed significantly higher errors for all models, confirming that occlusion poses a difficult challenge for existing methods. JRDB-Pose contains a significant number of occluded joints which we hope will be useful for researchers to improve the robustness of pose estimation methods to occlusions. We also find that all methods achieve lower AP as compared to their results on common large-scale benchmarks [26], showing that JRDB-Pose presents a reasonable and difficult challenge for multi-person pose estimation.

Method	$\mathcal{O}_{pose}$	Loc	Card	AP	AR
HRNet [49]	0.480	0.210	0.270	24.6	47.5
DEKR [19]	0.410	**0.113**	0.2968	31.7	46.9
CID [48]	0.377	0.174	0.204	38.6	48.0
YOLO-Pose [27]	**0.368**	0.172	**0.196**	**47.9**	**72.5**

Table 3. Multi-person pose estimation baselines evaluated on JRDB-Pose stitched annotations. Loc and Card are the $\mathcal{O}_{pose}$ localization and cardinality error, respectively.

Method	Total Contribution			Per-Keypoint Avg		
	V↓	O↓	I↓	V↓	O↓	I↓
HRNet [49]	0.131	0.054	0.025	0.051	0.075	0.102
DEKR [19]	**0.070**	**0.031**	**0.012**	**0.030**	**0.052**	**0.089**
CID [48]	0.101	0.047	0.026	0.038	0.063	0.108
YOLO-Pose [27]	0.098	0.048	0.026	0.033	0.055	0.091

Table 4. We break down $\mathcal{O}_{pose}$ localization error from Tab. 3 for visible (V), occluded (O), and invisible (I) joints. The left and right columns represent the total and average per-keypoint contribution (scaled by 1M) to the $\mathcal{O}_{pose}$ localization error, respectively.

4.2. Multi-Person Pose Tracking

We evaluate three recent state-of-the-art methods, such as ByteTrack [55], Unitrack [50], and OC-SORT [8], in our tracking benchmark. All these methods are from the tracking-by-detection category and predict tracks based on predictions from a pose estimation method. We use the estimates from Yolo-Pose, as our highest-performing baseline, to initialize the predictions for all pose tracking methods.

In Tab. 5 we provide our pose tracking results for all trackers and training strategies. OC-Sort achieves the best results by a wide margin. Since OC-SORT targets improving tracking robustness for objects in non-linear mo-

Pose Estimation Method (Training)	**Tracking Method**	MOTA ↑	IDF1↑	IDSW↓	$\mathcal{O}^2_{pose}$↓	Components		$\mathcal{O}^2_{pose}$↓ by Visibility		
						Card↓	Loc↓	V↓	O↓	I↓
Yolo-Pose [27] (COCO only)	ByteTrack [55]	61.32	55.80	4236	0.715	0.519	**0.196**	0.698	0.703	0.731
	UniTrack [50]	60.80	55.01	2854	0.727	0.473	0.253	0.713	0.716	0.733
	OC-SORT [8]	**66.09**	**59.93**	**2588**	**0.630**	**0.380**	0.249	**0.610**	**0.618**	**0.646**
Yolo-Pose [27] (JRDB-Pose only)	ByteTrack [55]	61.17	52.11	3203	0.693	0.429	**0.264**	0.678	0.692	0.708
	UniTrack [50]	**69.84**	56.06	2565	0.725	0.454	0.271	0.710	0.722	0.734
	OC-SORT [8]	69.22	**60.62**	**1977**	**0.577**	**0.295**	0.283	**0.556**	**0.577**	**0.595**
Yolo-Pose [27] (COCO→ JRDB-Pose)	ByteTrack [55]	67.16	55.38	3325	0.690	0.456	**0.234**	0.674	0.688	0.708
	UniTrack [50]	**72.82**	57.69	2413	0.708	0.435	0.272	0.693	0.707	0.719
	OC-SORT [8]	71.74	**61.15**	**2260**	**0.594**	**0.331**	0.263	**0.573**	**0.592**	**0.613**

Table 5. Multi-person pose tracking baselines evaluated on JRDB-Pose individual camera images.

tion with the improved Kalman-filter and recovery strategy, we believe it is better suited for JRDB-Pose which includes occluded periods during which people are often not moving linearly with respect to the robot's perspective (*e.g.* sequences where both the robot and nearby people are turning or moving simultaneously), thus better recovering from occlusions. This method's lower $\mathcal{O}^2_{pose}$cardinality error further confirms that it can better find accurate tracks. Compared to their performance in other large-scale tracking benchmarks [12, 13], these methods achieve relatively lower overall performance in the same reported metrics, reflecting the unique challenges in our dataset, which we hope will motivate further work within the research community.

4.3. Study on Pre-training Methods

Using our highest performing pose estimation method, we further study how pre-training strategies affect our final model performance. We try inference-only with a COCO-pretrained model, finetuning a COCO-pretrained model, and training from scratch. Tab. 6 and Tab. 5 show pose estimation and tracking performance of our highest performing model across the different training strategies. For pose estimation we find that finetuning from COCO generally performs better than training from scratch in both pose estimation and tracking tasks with an improvement of 1.7% in $\mathcal{O}^2_{pose}$, while running inference from a COCO model without finetune shows much weaker results especially in $\mathcal{O}_{pose}$. The relatively small improvement also suggests that JRDB-Pose contains a varied distribution of scenes and poses suitable for model training.

We also include results for pose tracking across the same three training strategies of Yolo-Pose used to initialize the tracking. In Tab. 5 we find that the JRDB-Pose model outperforms the COCO model finetuned on JRDB-Pose in MOTA and IDF1 but is itself outperformed in IDSW and $\mathcal{O}^2_{pose}$. The COCO model without finetuning performs worse than the other models. Since the COCO model is not trained on the crowded human scenes and overall image distribution present in JRDB-Pose, it struggles to find keypoints and thus has a high miss-rate as indicated by a much higher $\mathcal{O}^2_{pose}$cardinality component compared to the other training schemes. JRDB-Pose contains a varied distribution of scenes in human environments different from COCO. Based on the per-sequence results, the models typically demonstrate superior performance in indoor areas with no obstacles, such as *hewlett-class* and *nvidia-aud*. On the other hand, the models tend to perform more poorly in complex scenes containing numerous background objects, including robot hands, photocopies, coffee machines, trolleys, and so on. Such scenarios can be observed in *gate-ai-lab* and *indoor/outdoor-coupa-caffe*.

Training strategy	**Stitched**		**Per Camera**	
	AP ↑	$\mathcal{O}_{pose}$ ↓	AP ↑	$\mathcal{O}_{pose}$ ↓
COCO [26]	48.0	0.460	60.2	0.320
JRDB-Pose	48.2	**0.357**	61.9	0.280
COCO → JRDB-Pose	**49.6**	0.385	**63.9**	**0.273**

Table 6. We study how training protocols affect performance of YoloPose. Finetuning from COCO achieves best results.

5. Conclusion

In this paper we have introduced JRDB-Pose, a large-scale dataset of human poses and track IDs suitable for multi-person pose estimation and tracking from videos. JRDB-Pose features high-frequency annotations for crowded indoor and outdoor scenes with heavy occlusion, diverse actions, and longer videos than existing large-scale pose estimation and tracking datasets. Finally, we have introduced $OSPA_{pose}$ and $OSPA^2_{pose}$, new metrics for multi-person pose estimation and tracking. We believe that JRDB-Pose will help address limitations in human action understanding for human-robot interaction and navigation in human environments and advance research by proving large-scale and high-quality annotations.

References

[1] Felix Achilles, Alexandru-Eugen Ichim, Huseyin Coskun, Federico Tombari, Soheyl Noachtar, and Nassir Navab. Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications. In *MICCAI*, pages 491–499. Springer, 2016. 5

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, pages 5167–5176, 2018. 3, 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*, June 2014. 2

[4] Michael Beard, Ba Tuong Vo, and Ba-Ngu Vo. A solution for large-scale multi-object tracking. *IEEE Transactions on Signal Processing*, 68:2754–2769, 2020. 6

[5] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, pages 941–951, 2019. 4

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 ICIP*, pages 3464–3468. IEEE, 2016. 4

[7] Guillem Brasó, Nikita Kister, and Laura Leal-Taixé. The center of attention: Center-keypoint grouping via attention for multi-person pose estimation. In *ICCV*, pages 11853–11863, 2021. 4

[8] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv:2203.14360*, 2022. 4, 7, 8, 12

[9] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *AAAI*, volume 34, pages 10631–10638, 2020. 5

[10] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In *ICCV*, pages 723–732, 2019. 5

[11] Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, and Zicheng Liu. Transmot: Spatial-temporal graph transformer for multiple object tracking. *arXiv:2104.00194*, 2021. 4

[12] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003*, 2020. 2, 3, 8

[13] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *CVPR*, pages 20963–20972, 2022. 1, 2, 3, 8

[14] Yunhao Du, Yang Song, Bo Yang, and Yanyun Zhao. Strongsort: Make deepsort great again. *arXiv:2202.13514*, 2022. 4

[15] Mahsa Ehsanpour, Fatemeh Saleh, Silvio Savarese, Ian Reid, and Hamid Rezatofighi. Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection. In *CVPR*, pages 20983–20992, 2022. 1, 4, 5

[16] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *ICCV*, pages 10849–10859, 2021. 1, 2, 3

[17] 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 (ECCV)*, 2018. 3

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 2

[19] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*, pages 14676–14686, 2021. 4, 7

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, jul 2014. 2

[21] Sheng Jin, Wentao Liu, Wanli Ouyang, and Chen Qian. Multi-person articulated tracking with spatial and temporal embeddings. In *CVPR*, pages 5664–5673, 2019. 4

[22] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *ECCV*, pages 718–734. Springer, 2020. 4

[23] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *arXiv:2203.15041*, 2022. 3

[24] Rawal Khirodkar, Visesh Chari, Amit Agrawal, and Ambrish Tyagi. Multi-instance pose networks: Rethinking top-down pose estimation. In *ICCV*, pages 3122–3131, 2021. 3

[25] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942*, 2015. 3

[26] 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*, pages 740–755. Springer, 2014. 1, 2, 5, 6, 7, 8

[27] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *CVPR*, pages 2637–2646, 2022. 4, 7, 8, 12

[28] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *TPAMI*, 2021. 1, 3, 4, 5

[29] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, pages 8844–8854, 2022. 4

[30] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv:1603.00831*, 2016. 3

[31] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv:1603.00831*, 2016. 6

[32] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017. 4

[33] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 5

[34] Guanghan Ning, Ping Liu, Xiaochuan Fan, and Chi Zhang. A top-down approach to articulated human pose estimation and tracking. In *ECCV Workshops*, pages 0–0, 2018. 3, 4

[35] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *ICRA*, pages 9552–9557. IEEE, 2019. 2, 3

[36] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, pages 4929–4937, 2016. 6

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 7

[38] Hamid Rezatofighi, Tran Thien Dat Nguyen, Ba-Ngu Vo, Ba-Tuong Vo, Silvio Savarese, and Ian Reid. How trustworthy are performance evaluations for basic vision tasks? *arXiv:2008.03533*, 2020. 6

[39] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019. 6

[40] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35. Springer, 2016. 6

[41] Andrey Rudenko, Tomasz P Kucner, Chittaranjan S Swaminathan, Ravi T Chadalavada, Kai O Arras, and Achim J Lilienthal. Thör: Human-robot navigation data collection and accurate motion trajectories dataset. *RA-Letters*, 5(2):676–682, 2020. 2, 3

[42] Fatemeh Saleh, Sadegh Aliakbarian, Hamid Rezatofighi, Mathieu Salzmann, and Stephen Gould. Probabilistic tracklet scoring and inpainting for multiple object tracking. In *CVPR*, pages 14329–14339, 2021. 4

[43] Dominic Schuhmacher, Ba-Tuong Vo, and Ba-Ngu Vo. A consistent metric for performance evaluation of multi-object filters. *IEEE transactions on signal processing*, 56(8):3447–3457, 2008. 6

[44] Michael Snower, Asim Kadav, Farley Lai, and Hans Peter Graf. 15 keypoints is all you need. In *CVPR*, pages 6738–6748, 2020. 4

[45] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv:2012.15460*, 2020. 4

[46] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020. 1, 2, 3

[47] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, pages 601–617, 2018. 2

[48] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *CVPR*, pages 11060–11068, 2022. 4, 7

[49] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 43(10):3349–3364, 2020. 3, 7

[50] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Advances in Neural Information Processing Systems*, 34:726–738, 2021. 4, 7, 8, 12

[51] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 ICIP*, pages 3645–3649. IEEE, 2017. 4

[52] Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda. Transcenter: Transformers with dense queries for multiple-object tracking. *arXiv:2103.15145*, 2021. 4

[53] Dongdong Yu, Kai Su, Jia Sun, and Changhu Wang. Multi-person pose estimation for pose tracking with enhanced cascaded pyramid network. In *ECCV Workshops*, pages 0–0, 2018. 3, 4

[54] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, pages 2248–2255, 2013. 2

[55] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21. Springer, 2022. 4, 7, 8, 12

[56] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, pages 474–490. Springer, 2020. 4

[57] Tianyu Zhu, Markus Hiller, Mahsa Ehsanpour, Rongkai Ma, Tom Drummond, Ian Reid, and Hamid Rezatofighi. Looking beyond two frames: End-to-end multi-object tracking using spatial and temporal transformers. *TPAMI*, 2022. 4