

Appendices for “Beyond the Survey: A Systematic Empirical Study of Detection and Association in Visual MOT”

Linh Van Ma, Juhua Hu, Wei Cheng, Unse Fatima, Moongu Jeon

1 Multi-camera Multi-object Tracking

In this section, we introduce multi-camera multi-object tracking and its relationship to the main tracking diagram shown in Fig. 1.

1.1 3D Multi-camera Multi-object Tracking with Camera Calibration

Compared to single-camera multi-object tracking, multi-camera multi-object tracking (MCMOT) is significantly more complex, requiring two levels of association: (1) across cameras and (2) within cameras. Nevertheless, MCMOT enables stronger 3D perception by triangulating object positions from multiple viewpoints. Traditional MCMOT methods Herzog et al. (2023); Hofmann et al. (2013) begin with 2D bounding box detection, generate per-camera tracklets, and then associate them into global tracks. Other approaches predict 3D locations from 2D detections and perform association in 3D space Cheng et al. (2023). More recently, end-to-end methods bypass 2D detection entirely by projecting image features into Bird’s-Eye View (BEV) and conducting both detection and tracking in 3D Teepe et al. (2024a,b); Nguyen et al. (2022).

As illustrated in Fig. 18, 3D people tracking from multiple cameras can be achieved via three primary paradigms. Approach (a) Hou et al. (2020); Chavdarova et al. (2018); Teepe et al. (2024b)

trains 3D detectors using ground-truth 3D annotations to predict object locations on the ground plane, then applies single-camera multi-object tracking algorithms Ma et al. (2023). While effective, this method lacks full 3D object extent when using 2D bounding boxes as input. However, when pose keypoints are provided, it enables recovery of detailed 3D keypoints Isakov et al. (2019); Liao et al. (2024); Ye et al. (2022); Tu et al. (2020), though it remains limited to pose-aware objects (e.g., humans).

Approach (b) Ma et al. (2024c); Ong et al. (2022) detects objects in 2D, extracts re-identification features, and fuses information using Bayesian multi-sensor multi-object tracking frameworks to estimate the full 3D state, including center coordinates (X, Y, Z) and dimensions (width, length, height). Despite producing rich 3D outputs, these methods incur high computational cost due to multiple-hypothesis data association, achieving less than 30 FPS even with only three targets on an Intel Core i7 CPU, which falls short of real-time requirements.

Finally, approach (c) tracks objects independently in each camera and subsequently fuses the per-camera trajectories using the method in Nguyen et al. (2025b) to produce 3D outputs. This paradigm leverages batch processing to significantly improve the accuracy of the tracking. It turns out that these three approaches follow exactly the diagram shown in Fig. 1: the complex architecture

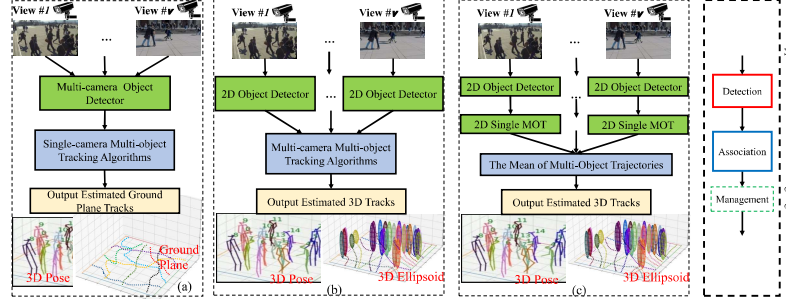


Fig. 18 Multi-view architectures: (a) Multi-view detection + single-sensor multi-object tracking [Chavdarova et al. \(2018\)](#); (b) Monocular detection + multi-sensor multi-object tracking [Ong et al. \(2022\)](#); [Ma et al. \(2024c\)](#); (c) 2D single multi-object tracking with the mean of multi-object trajectories [Nguyen et al. \(2025a\)](#). This paradigm of multi-camera multi-object tracking follows the same approach as shown in Fig. 1.

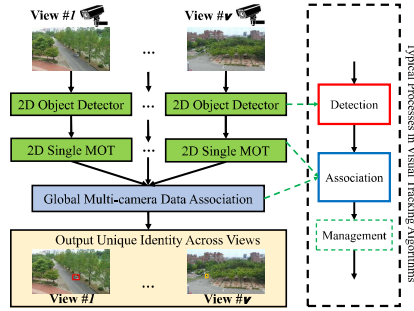


Fig. 19 A diagram of multi-camera multi-object tracking. Its structure is similar to that shown in Fig. 1. Unlike 3D multi-camera multi-object tracking, objects are tracked independently in each camera, and the results are then combined to assign consistent identities across all cameras. This global multi-camera association updates only the object identities, without modifying their spatial locations.

can be broken down into two modules: detection first, followed by association.

1.2 Multi-camera Multi-object Tracking without Camera Calibration

MCMOT with non-overlapping fields of view and no camera calibration presents significant challenges in constructing coherent global trajectories. The core difficulty lies in cross-camera

tracklet matching, where each target’s path is fragmented into disjoint local tracklets within individual camera views. These fragments must be robustly associated across cameras despite variations in appearance, occlusions, and the lack of spatial calibration. An overview of this tracking framework is illustrated in Fig. 19. In such settings, object detection and feature extraction, especially person or vehicle re-identification, play a crucial role in associating objects across views. This method also shares the same spirit as our unified framework shown in Fig. 1: detections are performed in each camera, and association comes later to output trajectories.

A common pipeline for addressing this problem involves single-camera tracking followed by inter-camera tracking. This approach effectively corrects local tracking errors caused by occlusions and missed detections, producing complete global trajectories and even estimating the total number of targets analytically.

2 Datasets and Evaluation

MCMOT Tracking algorithms are typically evaluated using CLEAR [Bernardin and Stiefelhagen \(2008\)](#), HOTA [Luiten et al. \(2021\)](#) scores, and OSPA⁽²⁾ metric. In 3D multi-pose tracking, performance is measured by the PCP3D score [Dong et al. \(2019\)](#) and extended Average Precision [Pishchulin et al. \(2016\)](#), where a pose is correct if its MPJPE [Tu et al. \(2020\)](#) is below K mm. There are several datasets to benchmark the performance of tracking algorithms as follows:

Multi-modalities Multi-object Tracking Datasets: Multi-modal object tracking integrates multiple sensing modalities, such as RGB, depth, thermal, language, and audio, to track objects for applications like autonomous driving and surveillance. The nuScenes dataset [Caesar et al. \(2020\)](#) offers a large-scale autonomous driving benchmark with multi-sensor data, including LiDAR, radar, and RGB cameras in urban environments.

Multi-camera Multi-object Tracking with Camera Calibration Datasets: For indoor 3D pose estimation, CMU PanOptic [Joo et al. \(2015\)](#), and CMC [Ong et al. \(2022\)](#) provides high-resolution multi-camera data, while Wildtrack [Chavdarova et al. \(2018\)](#), MultiviewX [Hou et al. \(2020\)](#) offer outdoor views for pedestrian detection and localization. Several scene-specific benchmarks-including MMPTRACK [Han et al. \(2023\)](#), MCT [Chen et al. \(2016\)](#), EPFL [Fleuret et al. \(2007\)](#), CAMPUS [Belagiannis et al. \(2014\)](#), and PETS09-S2L1 [Ferryman and Shahrokni \(2009\)](#) provide varying configurations of overlapping and non-overlapping camera views. Finally, CityFlow [Tang et al.](#)

[\(2019\)](#) serves as a comprehensive urban benchmark, utilizing 40 synchronized HD cameras across multiple intersections to facilitate large-scale research in MCMOT tracking and traffic analysis.

Multi-camera Multi-object Tracking without Camera Calibration Datasets: Unlike geometrically grounded datasets, many MCMOT benchmarks lack camera calibration, forcing cross-camera association to rely on appearance (ReID) and temporal reasoning. Key examples include DukeMTMC [Ristani et al. \(2016\)](#), SyntheCity [Herzog et al. \(2023\)](#), and MSMT17 [Wei et al. \(2018\)](#), which prioritize feature-based re-identification over spatial geometry.