Neurocomputing 552 (2023) 126558

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Survey paper

Multi-camera multi-object tracking: A review of current trends and future advances

Temitope Ibrahim Amosa [a,*], Patrick Sebastian [a,b], Lila Iznita Izhar [a,c], Oladimeji Ibrahim [e], Lukman Shehu Ayinla [a,e], Abdulrahman Abdullah Bahashwan [a], Abubakar Bala [a], Yau Alhaji Samaila [a,d]

[a] Department of Electrical and Electronics Engineering, Universiti Teknologi PETRONAS, 32610 Bandar Seri Iskandar, Tronoh, Perak, Malaysia
[b] Institute of Autonomous Systems, Universiti Teknologi PETRONAS, Malaysia
[c] Institute of Health and Analytics, Universiti Teknologi PETRONAS, Malaysia
[d] Department of Electrical and Electronics Engineering, University of Ilorin, 240103 Ilorin, Nigeria
[e] Department of Computer Engineering, University of Ilorin, 240103 Ilorin, Nigeria

ARTICLE INFO

Communicated by Zidong Wang

Keywords:
Multi-object tracking
Computer vision
Visual tracking
Intelligent video surveillance
Video surveillance
Multi-camera system

ABSTRACT

The nascent applicability of multi-camera tracking (MCT) in numerous real-world applications makes it a significant computer vision problem. While visual tracking of objects, especially in video obtained from single camera setup, has drawn huge research attention, the constant identification and tracking of targets as they transit across multiple cameras remains an open research problem. In addition to the linking of target appearance and trajectory information across frames, effective association of such data across multiple cameras is also very critical in MCT. Occlusion, appearance variability, camera motion, as well as nonrigid object structure and motion are widely recognized constraints and major sources of concerns in MCT. In recent years, several literatures have been contributed suggesting a variety of approaches to addressing various problems in MCT. However, studies that critically review and report the advances and trends of research in MCT are still limited. This current study presents a comprehensive and up-to-date review of visual object tracking in multi-camera settings. In this paper, we analyze and categorize existing works based on six crucial facets: problem formulation, adopted problem solving approach, data association requirements, mutual exclusion constraints, benchmark datasets, and performance metrics. Furthermore, the study summarizes the outcomes of 30 state-of-the-art MCT algorithms on common datasets to allow quantitative comparison and analysis of their experimental results. Finally, we examine recent advances in MCT and suggest some promising future research directions.

1. Introduction

Visual object tracking has constantly attracted considerable research attention due to the significance of the challenges posed to the research community and its commercial potential [1]. The underlying concern of visual tracking is to analyze video sequences from single or multi-camera systems, while aiming to determine the location of targets over a series of frames [2]. It plays an essential role in computer vision due to its capacity in several real-world applications, such as video surveillance, autonomous driving, human computer interaction, anomaly action detection, crowd behavior analysis, traffic scene understanding, etc. However, the presence of various photometric and geometric constraints has made visual tracking remain a challenging computer vision task [3]. Moreover, with the availability of low-cost high-quality, high-frame-rate security cameras and rising computing power, the deployment of multi-camera networks for various applications has grown significantly [4,5]. As a result, there is growing demand for a specific type of tracking algorithms known as *"multi-camera trackers"*. These trackers are designed to handle the complexities of tracking targets across multiple camera views and are necessary for many real-world multi-camera tracking applications. In Fig. 1, we provide a distinctive illustration of a multi-camera network layout and sample frames from each camera.

In the visual tracking community, the dominant research focus used to be around tracking either a single object (SOT) or multiple objects (MOT) within a single camera. However, with the increasing adoption of

* Corresponding author.
E-mail address: temitope_21002075@utp.edu.my (T.I. Amosa).

https://doi.org/10.1016/j.neucom.2023.126558
Received 31 January 2023; Received in revised form 6 July 2023; Accepted 8 July 2023
Available online 13 July 2023
0925-2312/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

multi-camera networks across various applications, the demand for multi-camera tracking systems has surged, surpassing the usage of single cameras [7]. This trend is primarily driven by the remarkable advantages offered by multi-camera networks, such as the ability to achieve comprehensive coverage of the monitored area and obtain a density of spatio-temporal sampling that outstrips what a single camera can offer [5,7]. As such, multi-camera tracking systems have become an indispensable tool for various applications where accuracy, efficiency, scalability, and reliability are paramount. Also, camera networks permit a mix of cameras of various kinds (e.g., omnidirectional, pan-tilt-zoom, fixed-lens thermals, etc.) to operate jointly, and are reliable for detecting and tracking targets. While multi-camera networks offer all these alluring advantages, the tracking of targets in such systems presents a more complex challenge than single-camera tracking. This is because multi-camera tracking requires the tracking of targets across multiple camera views, which poses several unique challenges [8–10]. These challenges include dealing with camera viewpoint variations, accurate calibration and synchronization of multiple cameras, handling occlusions and object appearance changes, and performing accurate data fusion and association, among others. These challenges demand advanced computer vision techniques and sophisticated algorithms to ensure accurate and reliable object tracking across multiple cameras.

In challenging scenarios, tracks may not be complete and instead appear as fragmented tracklets. The responsibility of trackers is then to correctly associate these tracklets in order to construct complete tracks. Track fragmentation typically occurs due to occlusions, which is very common in most multicamera arrangements, where objects are partially or fully obscured from the camera's view. Modern algorithms have attempted to tackle this issue by leveraging detection responses and pairwise interaction models [8,11,12]. Multi-target tracking algorithms exploring the detections responses, take the detection outputs from an object detection algorithm as their input, with the primary aim to match them together to obtain complete trajectories. However, many tracking methods that exploit this scheme often assume that the motion of each pedestrian is independent, disregarding the intricate and significant interactions that occur between individuals [13]. On the other hand, methods for predicting target movement based on target interaction besides detection responses are followed by other multi-camera tracking methods. These algorithms integrate the learned behavioral models from video and such models are trained by observing target movements in natural environment [14–16]. Some of the tracking approaches adopting this scheme are known for formulating the target trajectory patterns as a series of continuous optimizations [11,13,17]. However, the interaction models employed in most literature are restricted to a few hand-designed interactions, such as attraction and repulsion [18].

In the context of multi-object tracking, specifically in a single-camera configuration, the goal is to create accurate appearance and/or motion models, count objects in a video sequence, and maintain the distinctiveness of their identities as they move through a sequence of video frames [19]. However, for MOT in a multiple-camera setup, there are additional challenges that must be considered, such as understanding the topology of the camera network and accounting for viewpoint variations [5,20,21]. Object re-identification (re-ID) is also a crucial research area in multi-camera tracking [22], where algorithms analyze detections of the same91 target from different cameras to identify and extract representative properties of the target. Similar to single-camera tracking, the effectiveness of multi-camera visual tracking heavily relies on accurate and robust object detection algorithms for input observation, such as detections. Additionally, superior feature extraction algorithms play a crucial role in obtaining representative characteristics from detections and computing accurate similarities for improved multi-camera tracking performance [23].

1.1. Difficulties and challenges in MCT

Over the years, several well-defined detectors and trackers with competitive performance have been proposed. However, besides the fact that most of these approaches focus on a single camera scenario, the presence of several conflicting challenges also makes multi-target tracking a difficult task. Additionally, because most real-world applications employ multiple cameras, it is essential that such frameworks must be scalable to multi-camera scenarios. Since visual tracking is still an evolving research frontier for multi-camera setting, when formulating Multi-Camera Multiple Object Tracking (MCMOT) problem, several considerable practical difficulties and challenges need to be considered, including:

- *Dramatic appearance variabilities*: To track several objects, it is necessary to link together segments of moving objects (*tracklet*) in each frame of a video. The most common method for making this connection is by comparing the features of the foreground objects that have been segmented. From the motion object blob, features such as the edges, centroid, stereo disparity, texture, motion field, and gradients are often retrieved for association [24]. However,

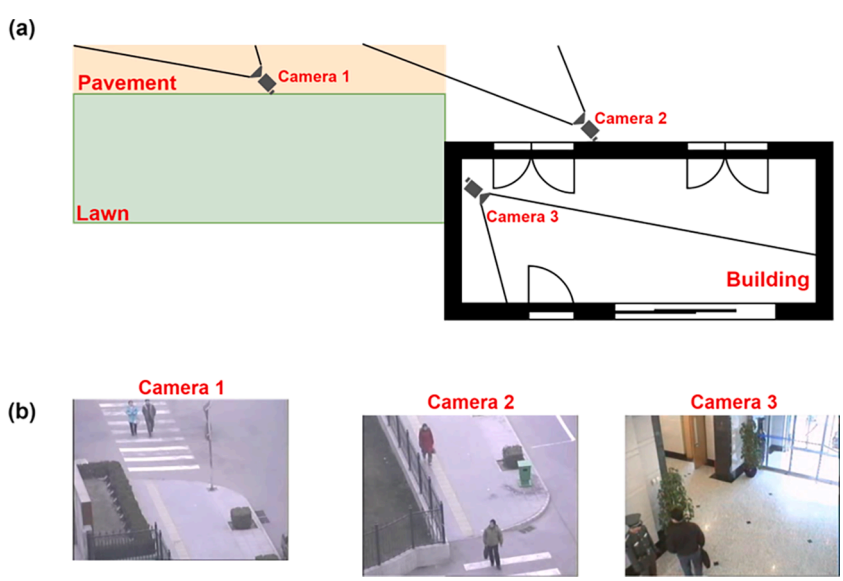


Fig. 1. A diagrammatic illustration of multi-camera network (a) camera layout depicting a non-overlapping camera FOV arrangement (b) Sample frames from each camera in the network .

appearance variabilities like cluttered backgrounds, non-rigid deformation, illumination changes, prolonged occlusions, shadows, and so on all can make the same object appear different, thereby making object tracking a difficult task [4,25].

- *Camera overlapping ratio*: When compared to traditional single-camera tracking, the problem of multi-camera system tracking is still a very recent one. New datasets and evaluation metrics that are focused on multi-camera scenarios are thus evolving. Additionally, multi-camera datasets mainly have the ability to record two kinds of scenarios. One is that different cameras are focused on the same scene. In other words, the cameras' field-of-views (FOVs) completely overlap with one another. The second category of dataset is made up of images from cameras, which barely or never overlap. The level of FOVs overlap however, affects the formulation of tracking problems in multi-camera systems, and thus influences the overall performance of the resulting tracking model [26].
- *Varying object motion*: Combining target motion and appearance similarities is a frequent practice in multi-object tracking in multi-camera systems. However, the precision of the calculated similarity is significantly impacted by any kind of fluctuation in target motion.
- *Camera motion*: It is difficult to track and associate various views of the same target across moving cameras since its appearance, orientation, and scale can vary dramatically. The importance of camera motion in visual tracking cannot be understated: approximation of image motion and target motion relative to the camera is important to determine scene properties and motion parameters [27].
- *Unknown number of targets and cameras:* The cameras capturing the presence of each target and the total number of targets across the multi-camera network are indeterminate. Consequently, this further complicates the task of determining the overall trajectory of each target [28]. This issue becomes even more complex when a tracker erroneously produces multiple local tracklets for a single target within a single camera [29].
- *Data association*: Considerably while data association in single-camera MOT applications is tough, it is made even more complex in MCMOT settings by the possibility of uneven lighting, prolonged occlusion [30], lack of spatio-temporal constraints among cameras [31], among others. The feature vectors generally associated with an object can alter significantly as the target moves from an indoor to an outdoor location. This is because most object detectors and Re-ID models may not take this change into consideration [32]. This causes a disruption in the relationship between an object's trajectory representation and its prior feature vectors, which results in incorrect identification (ID) assignments. For an online tracking system, failures in data linkage at a certain frame could have long-term negative implications [30,33,34]. Data association remains a critical part in evaluating the efficacy of an MCMOT and has received a lot of attention in recent times.

1.2. Prior literature reviews

In recent years, several review articles have been contributed to visual tracking domain, primarily with emphasis on the wide range of approaches for tracking single or multiple targets. So far, most of these review papers only taken into account visual tracking in single camera setup. For instance, an earlier survey by Yilmaz et al. [25] classified tracking techniques based on the object and motion representations employed. It gives detailed descriptions of the leading techniques in each category and weighs the benefits and drawbacks of each. The authors also cover important tracking-related topics such as choosing the right motion models, utilizing the appropriate image features, and object detection. Similarly, In [5], the author analyses from the perspective of computer vision and pattern recognition, the development of pertinent multi-camera system technologies. In this paper, the author discusses a variety of subjects, including cooperative video surveillance using both active and static cameras, computing the topology of camera networks, multi-camera calibration, multi-camera activity analysis, multi-camera tracking, and object re-identification. In [35], the authors present an overview of multi-camera tracking of people based solely on data fusion between different sensors.

In [36], the authors examine the available literature regarding the physical architecture and configurations, calibrations, algorithms, as well as benefits and drawbacks of multi-camera systems. In [37,38], the authors present a detailed review and experimental survey of deep learning (DL)-based visual trackers. In [35], the authors divide existing deep learning-based trackers into three categories based on network structure, network function, and network training. Furthermore, they performed an experimental survey on sixteen selected deep trackers to compare the representative methods on three popular benchmark datasets. Similarly, Marvasti-Zadeh et al. in [36] summarizes the underlying characteristics, key impulses, and major contribution of deep learning-based trackers from nine crucial aspects, namely, aerial-view tracking, long-term tracking, online tracking, network architecture, network exploitation, network training for visual tracking, network objective, and network output. They also conducted an experimental evaluation on the state-of-the-art DL-based trackers to compare their performance on seven benchmarks.

1.3. Motivation

Although several review articles [37–42] have been contributed on visual object tracking in recent years, a large number of these articles focused on specific aspects of architecture, technologies, and applications. Furthermore, they are also mostly restricted to visual tracking under single-camera settings. While numerous approaches for tracking under multiple camera setup have been proposed, studies that thoroughly analysis, comprehensively evaluate, and insightfully review and report recent trends and advances in MCT are still limited. To fill this apparent knowledge gap, this current study emphasizes the formulation, classification, and key concepts of multi-camera tracking (MCT). Although we acknowledge that multi-object tracking can be applied to both 2D and 3D data, the objective of this paper is to provide a comprehensive review of the existing literature on visual tracking in scenarios involving multiple cameras, with a specific emphasis on approaches that utilize 2D data. Also, efforts were made in this study to address important questions of research interests, and by evaluating 30 MCT trackers and top-rated baselines, we make several useful conclusions and point out future research directions. Our review provides a valuable resource for researchers and practitioners in the field of multi-camera multi-object tracking and serves as a roadmap for advancing this important area of research.

1.4. Contribution

We present the visual tracking community with a comprehensive review of the MCT problem, which we believe will enhance better understanding of the problem, its major challenges, drawbacks, and the state of the art. To sum up, the main contributions of this study are as follow:

1) We derive a consolidated formulation of the MCT problem that incorporates the majority of existing MCT methods (Section 2.1), as well as two distinct ways to classify MCT methods (Section 2.2).
2) We examine various essential parts of a MCT system, break down each into its component parts, and thoroughly discuss each part's guiding principles, benefits, and drawbacks (Section 3).
3) The popular multiple camera visual tracking benchmark datasets are compared based on their core characteristics (e.g., the number of videos, frames, number of cameras, sequence attributes and overlap) (Section 4).
4) Experimental findings on well-known datasets for various approaches are presented, facilitating future experimental comparison.

Some intriguing observations and discoveries are made by examining the provided results (Section 5).

5) We highlight current MCT research problems by summarizing the MCT review. Additionally, open issues are discussed to pinpoint potential future research trajectories (Section 6)

1.5. Organization

The goal of this study is to provide a broad overview of the essential elements of the MCT task. These components cover the current state of MCT research, various issues that must be considered when developing MCT system, and how to evaluate their performance. Section 2 provides a description of the MCT problem, including its general formulation (Section 2.1) and common classification schemes (Section 2.2). Specifically, the classification of existing MCT methods is discussed under the following dedicated headings: Multiple camera arrangement (Section 2.2.1), problem solving technique (Section 2.2.2), and tracklet association criteria (2.2.3). The most prevalent parts of modeling MCT are discussed in Section 2.2.3, specifically the appearance model (sub-Section (a)), dynamic motion model (sub-Section (b)), target dynamic interaction model (sub-Section (c)), optimization methods (Section 2.2.3.4), and mutual exclusion constraints (Section 2.2.3.5). Section 3 presents a detailed explanation of the MCT core components, that includes the single camera tracking (Section 3.1) and inter-camera tracking (Section 3.2). In addition, evaluation-related issues are covered in Section 4, including metrics (Section 4.1), public datasets (Section 4.2), publicly available trackers (Section 4.3), and benchmark results (Section 4.4). After this, Section 5 summarizes the pertinent problems and intriguing issues for the MCT research community's foreseeable future. Section 6 concludes the paper.

2. Multi-Camera tracking problem

It is challenging to comprehend the MCT problem from a high-level point of view because it has been presented in previous works from different perspectives. In this section, we provide a general formulation and argue that it can be used to unify existing works. To the best of our knowledge, other than the work of [43], limited prior work has been done in support of this attempt.

2.1. Problem formulation

In most surveillance and monitoring applications, the area of interest is assumed to be constantly watched by N cameras, denoted by the set $C = \{c_1, ..., c_N\}$. Each camera's detector and/or single camera tracker generates K observations of each target of interest, resulting in a set of observation $O = \{o_1, ..., o_k\}$, where K denotes the total number of observations observed up to the most recent time T. The appearance feature A_k, the track X_k, and the camera, $\mu_k \in C$ in which it appears make up the *kth* observation, where $o_k = \{A_k, X_k, \mu_k\}$. The track X_k is a collection of all track histories of observation o_k seen from camera μ_k i.e., $X_k = \{x_k^1, \cdots, x_k^l\}$ where l is the length of track of o_k recorded until time T and $x_k^l = (u_k^l, v_k^l, w_k^l, h_k^l, t_k^l, y_k^l)$ is the *lth* track history of o_k comprising the time stamp, t_k^l, position information (u_k^l, v_k^l) and the size information (w_k^l, h_k^l) with respect to the camera's (i.e., μ_k) image plane, and $y_k^l = (x_k^l, y_k^l)$ is a position on the ground plane. It should be noted that, depending on the setting, x_k^l may or may not contain y_k^l.

The MCT system generates a set of multi-camera tracks ,$M = \{M_1, ..., M_p\}$, using the observation set described above, where p is the size of the multi-camera tracks set, M. By definition, $M_p = \{oI_p\}_{I_p \subseteq \{1:K\}}$ is the *pth* multi-camera track consisting of observations, i.e., a collection of observations (or a single observation) with the same identity. I_p denotes an index set whose elements represent elements of a set O, implying that oI_p is a subset of O. A simplified notational representation suggested by [43] introduced $\{I_p^d\}$ such that d is the length of set I_p, (i.e., $|I_p|$) and I_p^l (the *lth* element of I_p) indicates an observation that is also *lth* observation of the multi-camera track M_p. Therefore, one can write $M_p^d = oI_p^d = o_k$, suggesting that the observation o_k is the same as the observation of multi-camera track, M_p. Also implying that the number of elements in I_p, denoted as d is the same as the number of observations contained in M_p, which is same as that contained in o_k denoted as l. Thus, $M_p^l = oI_p^l = o_k$. Ultimately, there is this constraint that each observation must belong to a unique multi-camera track for all observations and multi-camera tracks. The mathematical formulation of this constraint is as shown in Equation 2.1.

$$M_k \cap M_p = \varnothing, \quad \forall M_k, M_p \in M, \quad k \neq p \tag{2.1}$$

To correctly assert for or against any two observations being co-identified, that is, assume the same identity, numerous strategies have been proposed over the years [44–49]. Most of these approaches, however, agreed that there exists strong indication for the co-identity of two observations if their appearance descriptors are similar and their times and locations are consistent with typical walking speeds. Evidence is inconclusive if two observations differ in appearance or take place at nearby time instants in distant locations [50]. Another crucial challenge in MCT is correctly connecting a group of observations from the same target, with the basic understanding that mutual dependencies affect the mobility, appearance, and visibility of objects and must be taken into account. This necessitates probabilistic inference, typically maximum a-posteriori (MAP), in a posterior distribution across numerous non-independent variables [51,52]. The resulting optimization problem is substantially nonconvex (in the case of a continuous domain) or non-submodular (in the case of a discrete domain), and hence cannot be optimized globally without extensive simplifications.

This issue is often quickly and effectively resolved by using a graph model [30,53,54]. First, a solvable graph model is created using input observations (detections, tracklets, trajectories, or pairs) through the use of graph modeling. It consists of nodes, edges, and weights. The graph is then solved using an optimization approach, which yields local or global solutions. In subsequent subsections, we provide a classification of existing MCT approaches according to three modalities, that include, camera physical configuration, problem solving methods, and data association criteria.

2.2. Classification of MCT

Due to the significance of MCT, numerous approaches for object tracking in multi-camera systems have been put out during the past two decades or more. These tracking approaches have been categorized in many ways in the literature using various criteria. The following subsections give a distinct classification of current works on MCMOT based on four factors: camera arrangement, solution approach, data association criteria, and MCT optimization methods.

2.2.1. Multiple camera arrangement

A multi-camera system is a configuration of groups of cameras used for capturing images or scenes sequences. A multi-camera network can be either homogeneous (the setup includes just the same type of camera) or heterogeneous (the arrangement includes a variety of camera types) [36]. For high-precision applications, the use of two or more cameras is common, to increase the viewing area's range. In such a situation, the camera design can be divided into two types based on the intersection of the network's cameras: overlapping and non-overlapping field of view (FOV) setups.

(a) Overlapping multi-camera

Each camera has a FOV, which is a zone or range where the camera can precisely and easily detect everything. However, this FOV is finite

for every camera, and as a result, it is difficult to observe the entire trajectory of objects of interest in large areas with just one camera [55,56]. While it is preferable for most visual surveillance systems to have a wider view of targets of interest, multiple cameras are often employed. When using multiple cameras, it is possible to arrange the FOVs of the cameras in the network such that their FOVs overlap, creating common sub-areas where targets are visible to all cameras. In Fig. 2 (a), we illustrated a variety of ways that the fields of view of two cameras can overlap. Some earliest literatures [55,57–61] in MCMOT considered visual tracking under this kind of camera arrangement. For instance, Collins et al. in [57] developed a distributed network of active surveillance video cameras with overlapping FOV that enables activity monitoring in clutter environments. The authors also built a video understanding model that allows automatic detection and tracking of humans and vehicles using the network of overlapping FOV active video sensor. Similarly, Cai and Aggarwal [58] also demonstrated the feasibility of end-to-end tracking of targets from video sequences obtained from synchronized (overlapped) multiple cameras using 3D geometry in the various camera coordinates. The tracking algorithm suggested in that paper complements the known methods in motion detection, segmentation, and recognition, with the motion analysis from 3D geometry in different camera coordinates, while using geometric and intensity characteristics of the object to facilitate detection and tracking.

In [55], the authors address the problem of multi-target tracking in a camera system with multiple overlapping uncalibrated cameras. The authors stressed the importance of tracklet correspondence for each target and formulated it as a problem of consistent object labelling as seen from the perspective of multi-camera arrangement. This work reported the usefulness of FOV line limits as a valuable instrument to distinguish between the multiple possibilities for track association, while also suggesting an automated approach that uses the observation of motion in the environment to obtain these lines. Using multiple synchronized cameras, Mittal and Davis [60] proposed an approach for detecting, segmenting, and tracking multiple people in cluttered environment. The proposed system is fully autonomous and takes detection and tracking decisions using information collected from multiple pairs of overlapping video cameras. In this paper, the authors leverage a Bayesian classification-based segmentation, a region-based stereo algorithm provided only with target projections in two views, to compute the 3D points with the target, and occlusion analysis as fusion mechanism to merge information from various camera pairs. While it can be inferred from most of the prior literatures in this category that solving MCMOT problem requires camera calibration and environmental modelling, majority of literatures considering this configuration solely are limited to two or three decades ago. As also noted by the authors in [31], the practicality of assuming that cameras have overlapping fields of view is often limited by the considerable number of cameras needed and the physical limitations imposed by their placement.

(b) Non-overlapping multi-camera

From the perspective of numerous authors regarding economic and computational considerations, employing multiple cameras with overlapping fields of view can be costly. As a result, non-overlapping field-of-view camera networks are preferred and widely used in real-world applications [29–31,63]. In Fig. 2 (b), we illustrate a non-overlapping FOV multi-camera arrangement with four cameras. Tracking targets in video footage captured by multiple non-overlapping cameras is universally recognized as a daunting task. While the cameras are deliberately positioned far apart to minimize costs, this setup inevitably results in prolonged occlusion of targets and significant variations in viewpoint and illumination across different fields of view, further adding complexity to the tracking process [63,64]. Over the years, numerous authors have considered solving various tracking issues for this kind of arrangement. In [10], Song et al. delved into the challenge of tracking multiple targets in a camera network configured in a non-overlapping arrangement. To tackle the inherent difficulties posed by this layout in addition to known issues like illumination and pose variations, sensor noise, scene clutter, and occlusion, the authors proposed an innovative approach. They introduced a multi-objective optimization framework that combines short-term feature correspondences across cameras with long-term feature dependency models. This integrated approach aims to address the additional issues arising from the non-overlapping arrangement, enhancing the accuracy and effectiveness of target tracking in camera networks. The authors in [63] learn highly discriminative features for both multiple target tracking and Re-ID in a network of non-overlapping cameras. In this work, the author proposed an adaptive weighted triplet loss function that can learn good features for both tracking and Re-ID using convolutional neural networks (CNN) trained using hard-data mining technique.

In [64], Chen et al. propose a batch learning scheme for learning spatio-temporal features for tracking target in a non-overlapping camera arrangement. The authors also proposed an approach to remove weak links, which are the main causes of false correspondence and increased computation complexity for tracking in camera networks. An approach for track association builds on the reference set-based appearance was suggested by authors in [65] for a network of nonoverlapping cameras. The researchers in reference [53] utilized Equalized Graph Models

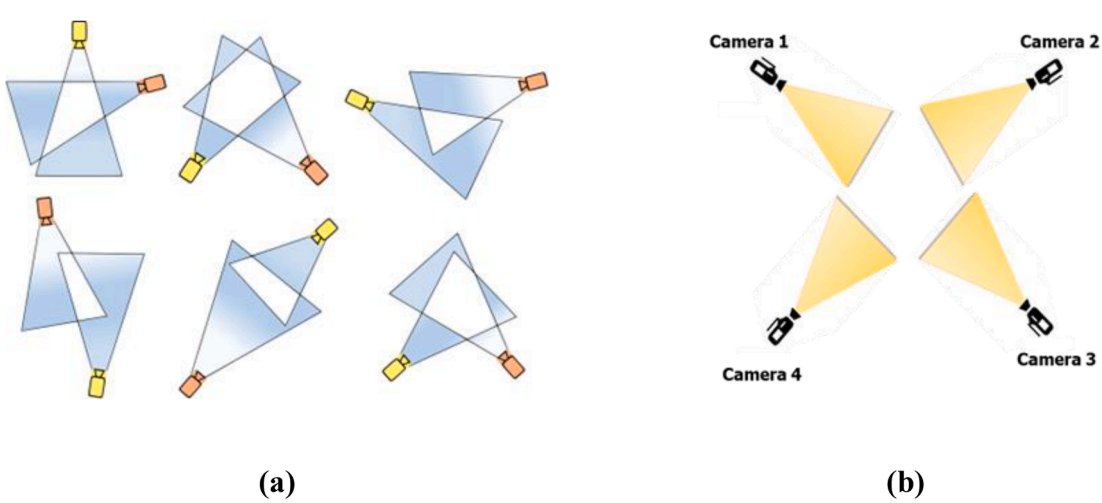


Fig. 2. Camera setup in a multi-camera system. (a) a variety of ways that fields of view can overlap [62] (b) example of non-overlapping arrangement in multi-camera systems.

(EGM) to correspond various within-camera tracklets to obtain the complete trajectories for non-overlapping multi-target tracking. In this approach, each tracklet is represented as a node in a graph, and the connections between tracklets are represented as edges. By finding a min-cost flow from the source node to the sink node, complete trajectories for each target are obtained by linking the tracklets together. Similar to [63], Zhang et al. in their study [66] also develop a resilient feature representation for individual targets by leveraging Convolutional Neural Networks (CNNs) for MCMOT under these camera layouts. Additionally, they propose a novel Feature Re-Ranking (FRR) mechanism to identify correspondences among tracklets with improved accuracy. Furthermore, in recent research [30], social grouping has been incorporated into the framework of Multi-Target Multi-Camera Tracking (MTMCT). This approach utilizes a Conditional Random Field (CRF) model to link local tracklets and determine tracklet associations by minimizing the unary and pairwise energy costs within the CRF model.

In order to enhance the precision of data association results and streamline the search process, the most of the MTMCT algorithms, particularly those considering non-overlapping setting as in [31,53,64,67], employ techniques such as pruning infeasible matching candidates based on camera topology and motion continuity. Additionally, a set of constraints, including coupling and non-overlap constraints [68] as well as spatial–temporal constraints [65], are introduced to ensure adherence to the matching consistency principle. It is also noteworthy that there are research studies that considered the problem of tracking for both overlapping and non-overlapping scenarios jointly. For instance, Quach et al. formulate the problem of data association in both camera arrangements as a link prediction on a graph [30]. In this work, the authors effectively combine while considering temporal information, the dynamic graph formulation with link prediction. Similarly, He et al. solved the cross-camera tracklet association problem in both overlapping and non-overlapping camera setting using the Restricted Non-negative Matrix Factorization algorithm. More recently, there has been a growing focus on addressing the Multi-Target Multi-Camera Tracking (MTMCT) problem, with two solution approaches being explored. Authors in [8] considered tracking of multiple interacting targets in a wide-area camera network that is composed of both overlapping and non-overlapping cameras. To decide the state is merged or split, the authors applied structural SVM to group state, and associate each local tracklets with group state based on the outcome of such decision.

2.2.1.1. Problem solving techniques. To address visual tracking problems in multi-camera settings, two major approaches are adopted in existing literature. The first one is by executing single camera tracking together with inter-camera tracking across all cameras within the network (two-step hierarchical approach). The second one is to globally consider all input detections (global approach) [6]. In this subsection, a detailed overview of these methods is contributed.

(a) Two-step hierarchical approach

As depicted in Fig. 3 (a), earlier visual object tracking methods in multi-camera systems tackle the problem in two distinct steps. The initial step involves estimating the consecutive track sequences, also known as tracklets, for all targets in each camera within the network. This task is commonly referred to as single-camera multi-target tracking and has been extensively researched within the computer vision and pattern recognition communities, with a plethora of studies conducted [49,69–74]. With the remarkable achievements of deep learning in computer vision tasks such as object detection [75–82], the tracking-by-detection approach [49,69,71–73,83] has emerged as the dominant method for multi-target tracking in recent times. In the second step, matching of each target trajectories as it moves across different cameras in the network is carried out utilizing an object re-identification algorithm [28]. This step is also regarded as the "handover" stage [84,85]. Formulating MTMCT problem as two hierarchical steps allows for data association within each camera as well as across cameras, commonly referred to as intra-camera and inter-camera tracking. In particular, the researchers in [30] presented a novel approach for solving the inter-camera tracklet assignment problem by formulating it as dynamic link prediction on a graph. In their study, they assumed that tracking within each camera is performed using a widely used single-camera MOT algorithm, specifically deepSORT [69]. To design a tracker robust to large and crowded scene, Nguyen et al. [86] formulate data association problem in MCT as spatio-temporal lifted multi-cut. They also leverage centralized representation concept to develop pre-cluttering algorithm used for detection grouping to avoid the ID-switch error common to most single camera trackers. Similar to [30], the researchers also assumed a pre-solved single camera tracking problem, while employing a readily available SCT algorithm, precisely centerTrack [87], to generate tracklets within each camera view.

By employing the tracklet-to-tracklet assignment method, He et al. [29] addressed the inherent problem of cross-camera matching. In their paper, the authors suggested the restricted non-negative matrix factorization to solve the resulting assignment problem of inter-camera

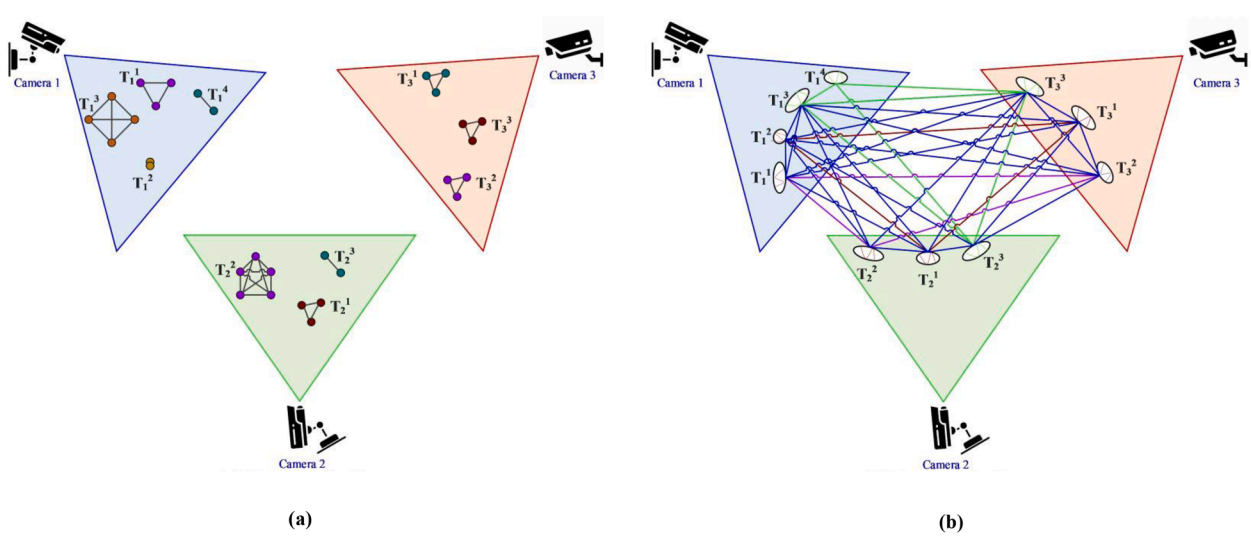


Fig. 3. Illustration of non-overlapping camera arrangement **(a)** Target trajectories is given in each camera. (b) across-camera trajectories of each target are shown as they transit from one camera to another. The arrangement is composed of three cameras and for each camera, there are three targets, except in camera 1 with four targets. Each target as a set of consecutive sequence of track labeled as T_i^n, where i denote the camera id and n is the target number.

tracking, while meeting all the set constraints. They also proposed an approach for estimating the number of targets within the camera network, and the proposed tracker performed better in the face of occlusion and missing detections. Tesfaye et al. [88] introduce a comprehensive three-layer two-step hierarchical framework that leverages the fast constrained dominant sets clustering (FCDSC) technique to address multi-camera tracking challenges. The proposed framework utilizes the first two layers to resolve tracking within individual cameras, while the third layer is designed to simultaneously match tracklets of the same object across all cameras. This unified approach provides a robust solution for tracking in multi-camera setups, allowing for efficient and accurate object tracking across different camera viewpoints. In their vehicle tracking pipeline, Hsu et al. [89] proposed a camera link model (CLM) based on self-supervision to address the local tracklet matching problem in MTMCT. The proposed CLM leverages both topological and appearance features of the targets systematically obtained via graph auto-encoder (GAE) network. The better performance of the proposed model was tested on CityFlow 2019 and 2020 datasets, and competitive performance was obtained. Similarly in [90], the cross-camera tracklet association method proposed [29] was also utilized for vehicle tracking application in a city-scale multi-camera network.

Xu et al. [84] proposed a Bayesian approach in their study to tackle the challenge of local tracklet matching. This approach incorporates a semantic attribute with trajectories through a novel technique called Spatio-Temporal Attributed Parse Graph (ST-APG). By utilizing the semantic attributes of targets along with commonly used cues such as geometry and appearance features, this innovative method effectively prunes matching candidates, resulting in improved accuracy and efficiency in local tracklet matching. In another study by Xu et al. [91], a Hierarchical Composition framework for multi-camera object tracking is introduced. This framework utilizes multiple cues in both 2D and 3D, including object appearances similarity, 3D grounded position, etc. to match local tracklets. By considering various object attributes in a hierarchical manner, the HCT framework offers a robust and comprehensive solution to the matching problem in tracking scenarios. However, the tracker could only obtain a speed of 5 FPS on average during inference, making it unsuitable for real-time application.

While the two-step approach is commonly used in most multi-camera tracking (MCT) literature, the inter-camera tracking step often heavily depends on the output of the intra-camera tracking step. This implies that the performance of the data association component for inter-camera tracking is determined by the output of the intra-camera tracking module, which in turn affects the overall performance of the multi-target multi-camera tracking (MTMCT) system. Furthermore, many existing studies on MTMCT assume that the intra-camera tracking problem is already solved by using an off-the-shelf algorithm, with only a limited number of studies considering the development of a custom intra-camera tracking module from scratch.

(b) Global MCT approach

Available evidence has revealed that performance of the SCT component has a significant impact on the overall tracking performance of MTMCT models based on two steps involving SCT and Re-ID. However, the results of most of the existing SCT approaches frequently include fragments and false positives for challenging scene videos, making them also unstable. The inter-camera tracking (ICT) module is directly harmed by these false positives and fragments resulting in issues ranging from wrong matching, missing tracklets, ID-switch etc. Several authors noted that as long as the multi-camera object tracking (MCT) problem is solved in two steps, these issues will always exist [53,92]. To address this problem, a new solution is sought that integrates the two distinct steps and jointly optimizes them. This approach is referred to as the *Global MCT approach or Global data association*. In Fig. 4, we illustrate the three possible solutions for MCT task.

The two distinct steps comprising the solution approach discussed

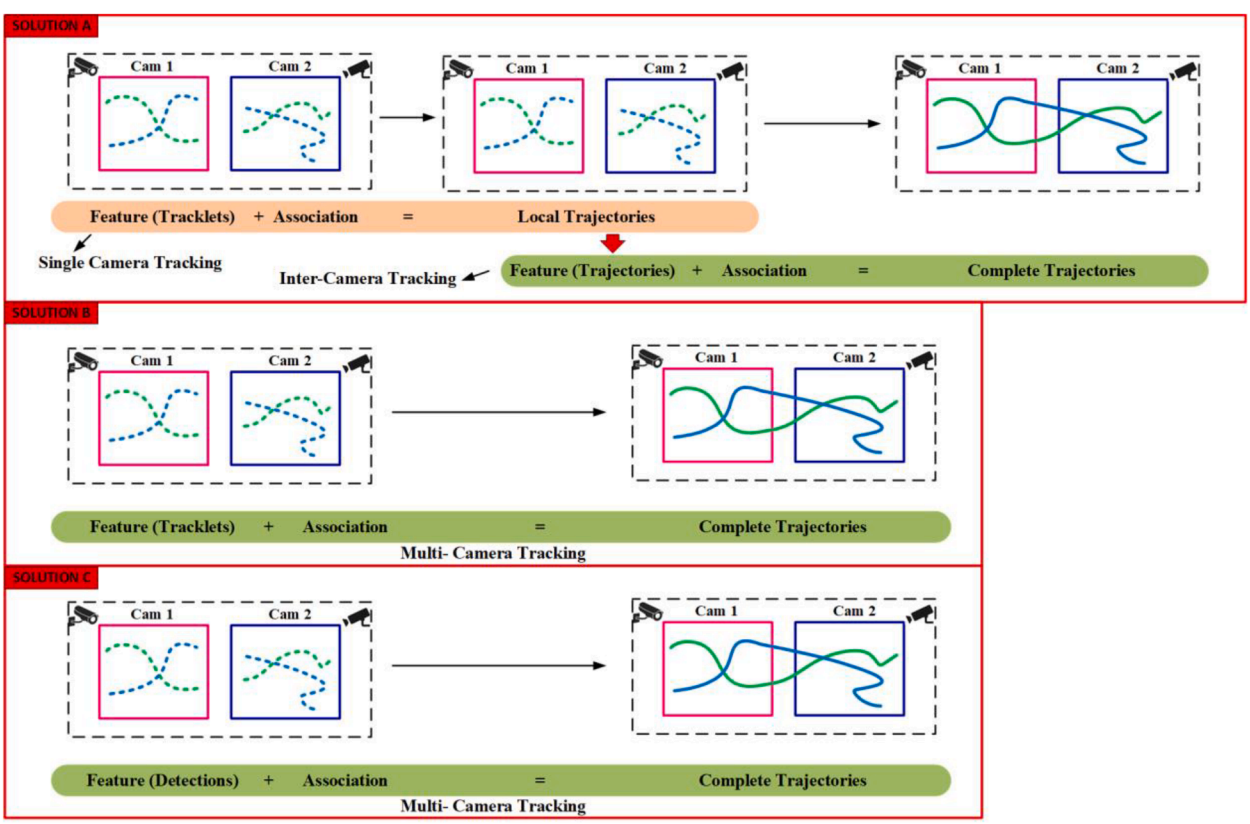


Fig. 4. Illustration of the three viable solutions to multi-camera tracking. Solution A shows the two-step hierarchy approach involving two stages of data association (within and across camera). Solution B and C adopt the global solution approach solving the data association in just one single step.

previously, each have a distinct data association framework, as demonstrated in solution A in Fig. 4. The data correspondence scheme in the first step is performed across frame and is also known as "within-camera data association". In the second step, the across camera data association is performed, and may also be referred to as "inter-camera data association". However, both data association tasks are frequently formulated either as an assignment problem or a network design problem with an optimization solution. Because the formulation of both tasks is the same, merging them into one would offer a lot of benefits and save runtime [23]. To integrate these two data association frameworks, the simplest approach is to create a new data association that takes initial observations (e.g., detections or tracklets) as input and directly outputs final trajectories [53]. Solution B and C in Fig. 4, displayed the two possible solution approaches for global data association. It is also important to mention that global MCT methods have seen a surge in popularity in recent years due to their improved ability to handle the challenges posed by noisy detection input [54]. In these methods, optimization is performed over a batch of frames, rather than just two or a few consecutive frames, resulting in increased temporal locality [20,21,8,27]. This allows for the incorporation of more global properties of targets during the optimization process, thereby enhancing the overall tracking performance [54].

In recent times, several studies [23,53,54,92] have been provided to unify the data association methodologies in single camera tracking (SCT) and inter-camera tracking (ICT) into a single easy methodology. For instance, authors [93] introduced a MCT algorithm that combines motion and appearance features following the global level. A key aspect of their approach is the formulation of data association in a way that addresses the issue of limited temporal-locality, which often result in ID-switching in other methods. To solve the optimization problem arising from their data association approach, the authors utilized Generalized Minimum Clique Graphs. This allowed them to effectively handle the complex data association problem in MCT and overcome the limitations of existing approaches that suffer from ID-switching due to limited temporal-locality. In [53], the authors developed an equalized global graph model for integrating these two steps together, while also contributing an improved similarity measure. To obtain a unifying data association scheme, Liu et al. [23] adopted the generalized maximum multi-clique optimization problem (GMMCP), and formalized MCT as a global graph problem. The study utilized the re-id feature obtained from the Local Maximal Occurrence (LOMO) representation algorithm and the motion information obtained by solving iterative Hankel Total least squares as the inputs to the graph. Like [23], data association in MCT was also formulated as a generalized maximum multi clique problem in [54]. The authors further reveal that Binary-Integer program is the most suitable formulation for the NP hard problem of GMMCP. Without using any heuristics, the issue of miss-detection and occlusion was dealt with in the paper via Aggregated Dummy Nodes.

As global Multi-Camera Tracking (MCT) approaches continue to evolve, a critical question that remains unanswered is how to accurately measure the similarity between two observations in the newly constructed graph. This challenge arises from the fact that some observations are obtained from the same camera, while others are derived from different cameras in a multi-camera setup. Finding a robust and consistent similarity metric that can effectively handle both intra-camera and inter-camera observations remains an ongoing research problem in the field of MCT. If using the same similarity metric, observations from different cameras would typically have a lower average similarity score than observations from the same camera. This is because the appearance and spatiotemporal information of objects in ICT are less reliable than those in SCT due to a variety of factors such as camera settings, viewpoint variations, and illumination changes. Furthermore, the optimization process favors connecting the observations along edges within the same camera over connecting them across cameras, leading to a suboptimal result for the entire MCT. To resolve this issue, as noted in [94], the following two questions must be addressed:

i. *How to distinguish the similarities between observations contributed by the same camera from those obtained from different cameras?*
ii. *How to effectively balance the distribution of these heterogeneous similarities in the new graph?*

2.2.1.2. Tracklets association criteria. Association-based tracking techniques have been suggested as a result of significant advancements in object detection [37,38]. They frequently establish correct linking affinities based on multiple cues between tracklets, also known as track fragments [95,96], and use the Hungarian algorithm [48,96–98], Multiple Hypotheses Tracking (MHT) [99–101], Markov Chain Monte Carlo (MCMC) [102,103], etc. to locate a global solution with the highest likelihood. To establish correspondence between two observations in multi-camera tracking, the similarities between object appearance features and/or the dynamic motion are often considered. In this subsection, we provided a brief overview of the two models as an important data association criterion.

(a) Appearance Model

In MOT, appearance serves as a crucial clue for determining correspondence between observations. An appearance model has two parts: visual feature representation and statistical measurement [104]. The visual characteristics of an object are described using some features, either based on a single cue (such as color, edge etc.) or multiple cues [105,106]. Additionally, the selection of appropriate features holds a critical position in visual tracking and is closely related to object representation. Various visual features have been used to model the appearance of a target over the years. The following are the details of common visual features:

i. *Edge-based features:* Typically, object borders create significant variations in image intensity. An important characteristic of edges is that they are less susceptible to changes in illumination than color characteristics. Edges are typically used as the representative feature by algorithms that monitor object boundaries [53]. In the process of extracting edge features, edge detection such as Canny edge detection [107] and Sobel edge detection [108] is a crucial step. This feature have been leverage in several MTMCT works such as [109].
ii. *Color-based features*: Numerous computer vision tasks have demonstrated that color-based features are the most effective options. This is because color characteristics retain powerful clues that capture human perception in a low-dimensional space and may be generated with less computational overheads than other complex features. The most popular representation of color-based features is color histogram, and this representation is adopted by several authors [8,43,50].
iii. *Texture-based features*: the visual characteristics of homogeneous areas that do not result from the presence of a single color or intensity are captured by this category of feature. Moreover, these areas may have distinct visual patterns or spatial pixel arrangements, and the color information in a region may not be adequate to describe them. Several authors [110–112] utilized this type of feature to build their MCT model.
iv. *Depth-based features*: In comparison to color, edge and texture, depth is another valuable information that can be exploited to provide object appearance representation that is invariant to illumination, background clutter and color change. It is possible to derive a person's shape and physical characteristics from a full-size point cloud and skeleton using only depth information. In addition, the depth value associated with each pixel makes it simpler to distinguish people from the scenery, which greatly reduces the impact of the latter. MCT approaches that employ

depth-based appearance features are still evolving, example include as [113,114].

v. *Region-based features*: A region is any set of points that are surrounded by other points. For region-based features, the pixels of interest are clustered together based on their proximity to one another and their shared characteristics [31]. Several commonly employed features in MOT fall into this category including HOG, pixel template, covariance matrix etc.

vi. *Optical flow:* Optical flow is an approximate notion of image motion that is based on the projection of the velocities of three-dimensional surface points onto the image plane of a visual sensor [115]. It is used as a stand-in for the full representation of motion, correspondence between image features, correlations, or attributes of intensity hierarchies. The assumption for measuring image motion with optical flow is that local time-varying image patches have relatively consistent intensity structures during motion [115,116], this underlying assumption is referred to as "brightness constancy". This feature is commonly used jointly with other features as in [117,118].

vii. *Deep learning features*: Other aforementioned features are created and specified manually by individuals (designer), and they are known as the *"hand-crafted features."* However, due to the exceptional ability of deep learning architectures to automatically extract and learn highly distinctive features from images, many modern MCT methods rely on these features. These features are commonly referred to as *"deep features."* Several researchers exploit these features for multi-target tracking such as in [63,119–121]

viii. *Others features representations*: Other commonly employed appearance features are gait feature [122], which is unique for individual persons, covariance matrix descriptor, Color names descriptor [123], etc.

The use of multiple cues to define the target appearance model increases the descriptiveness in visual tracking. The integration of multiple features might be done at the tracker or statistical measurement level. Statistical measurement, on the other hand, is the computation of similarity between distinct observations based on their visual representation. While tracker level fusion allows for the deployment of a broader range of trackers, measurement level fusion is preferred to avoid numerous runs of the whole tracker, hence lowering the number of similar and redundant tracking hypotheses [106]. The similarity between two observations u and v can be formally expressed as:

$$S_{uv} = G(o_u, o_v) \tag{2.2}$$

where o_u and o_v represent the visual representation of observations u and v respectively, and $G(\cdot,\cdot)$ is a function that determines their similarity. Furthermore, visual representation utilizing a single feature is either translating distance into similarity or calculating similarity directly [29]. Table 1 provides a comparison in terms of metric category and mathematical formulation for commonly employed affinity measures in visual tracking.

(b) Dynamic Motion Model

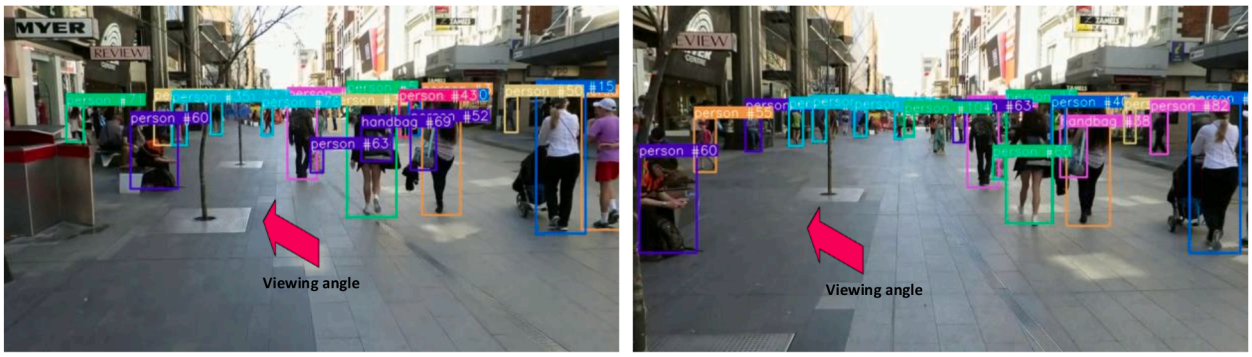
The motion model captures the movement pattern of an object. It estimates the possible positions of objects in future frames, reducing the search space [104]. In most instances, it is assumed that objects move smoothly in the world and, consequently, in the image space (except for abrupt motions). Generally, in visual tracking, motion models can be categorized into linear [48,51,149–155], nonlinear [96,152], and relative [95] based on the target movement pattern. In multi-camera tracking, the linear motion model is typically the most used motion model. This category of motion model assumes a constant velocity between successive frames and can be constructed in three ways on the premise of this assumption [104], namely, position smoothness, velocity smoothness, and acceleration smoothness. When using a linear motion model in multi-camera trackers, the likelihood that two tracklets are related is often determined by how well they align with the linear motion assumption [95]. However, when there is a change in view angle due to camera motion, this smoothness may be affected as shown in Fig. 5. To correct this, frame matching techniques can be used, but these can be challenging to implement. Other solutions, such as non-linear [96] and relative [95] motion models, have been proposed as more accurate ways to handle this problem.

Yu et al. [155] argued that the existence of imprecise segmentation, prolonged occlusions from stationary objects and targets, and incorrect detections may result in a particular foreground area being insufficiently correlated with one target. As a result, the typical assumption of a one-to-one correspondence between targets and observations in many data association techniques is not satisfied. To address this problem, the authors proposed a solution that frames the problem of visual tracking as the search for the optimal spatial and temporal association of observations, which maximizes the consistency of both motion and appearance of trajectories. In this work, while the linear motion model is used, only the motion and appearance consistency are considered as the exclusion constraints. Similarly, the authors in reference [48] also proposed an online-based tracking method that can perform effectively against complex scene and low discriminability between objects appearances. In this work, the authors first define the tracklet confidence using the detectability and continuity of a tracklet and decompose a multi-object tracking problem into small subproblems based on the tracklet confidence. The method combines both appearance and motion models, and linear motion model with constant velocity is used.

In [96], Yang and Nevatia presented an online method to acquire robust appearance models and non-linear motion patterns for multi-target tracking within a tracklet association framework. While many prior approaches employ only linear motion methods, their approach constructs a non-linear motion map online to better capture directional changes and generate more resilient motion affinities among tracklets.

Table 1
Comparison of common distance/similarity measures in visual tracking.

Metric Name	Metric category	Formulae	Equation number	Reference
Sum of Squared Differences	Distance	$\sum_{u,v}(T(u,v) - I(x,y))^2$	2.3	[124–126]
Normalized cross correlation (NCC)	Distance	$\frac{\sum_{i=1}^{MN}(x_i - y_i)^2}{\sqrt{\sum_{i=1}^{MN} x_i \sum_{i=1}^{MN} y_i}}$	2.4	[106,127–131]
Euclidean distance	Distance	$\sqrt{\sum_{i=1}^{MN}(x_i - y_i)^2}$	2.5	[8,86,88,132–135]
The Earth Mover's Distance	Distance	$\frac{min_{F=(f_{ij})\in F(x,y)} \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij}}{min \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}$	2.6	[50,136–141]
Bhattacharyya coefficient	similarity	$\sum_{i=1}^{m} \sqrt{p_i q_i}$	2.7	[65,142–146]
Intersection-over-union (IOU)	similarity	$IOU = \frac{A \cap B}{A \cup B}$	2.8	[89,121,147,148]



(a) (b)

Fig. 5. Position estimation based on camera movement. (a) viewing through the arrow indicating the viewing angle, the location of person with ID #60 can be seen to be far from camera. (b) as the camera moves forward, the target appears closer to the camera.

Sadeghian et al. [11] observed that even if a target's actual motion is linear, the sequence of coordinates and velocities can be non-linear due to the presence of noisy detections, particularly when reasoning in the image plane. To capture these non-linearities through data-driven learning, they employed a Long Short-Term Memory (LSTM) network to train on trajectories of 2D velocities affected by noise (extracted through their motion feature extractor).

(c) *Target Dynamic Interaction Model*

Many tracking techniques assume that the movement of each pedestrian is independent, disregarding the intricate and crucial interaction between the target and other targets in the scene. In reality, a pedestrian's movement is determined not only by their own past motion alone but also by the motion/behavior of adjacent subjects [11,13]. Recent research has shown that tracking methods that take into consideration the interaction between targets in the scene can give much better results in complex scenes [11,15,156]. To account for the influence of nearby objects in determining target motion, existing studies in tracking integrate the dynamic interaction model of the target into the problem formulation via group model [13,15,157], crowd motion pattern [156], social LSTM [11,158], and many more. Fig. 6 shows how targets adjust their trajectories based on interaction with other targets in the scene.

In [15], Pellegrini et al. proposed a dynamic social behavior model that take into account the important behavior of pedestrians in collision anticipation and early-stage trajectories adjustment for collision avoidance. This work is mostly regarded as one of the very first efforts made at integrating dynamic interaction of target in tracking and was motivated by various studies on crowd simulation. However, besides this approach being designed for only short-term trajectory prediction, recent studies have also shown that collision and avoidance alone is not sufficient to model target interaction. Similarly, to improve the tracking performance, the authors in [156] introduced an agent-based behavioral model of targets. The model presented in that paper adopts an energy minimization approach, where the objective is to minimize the overall energy required to make a decision. The target in this model is considered decision-making agents who take into account multiple personal, social, and environmental factors to determine their next course of action. Leal-Taixe et al., [13] proposed an approach that integrates interaction of pedestrians in tracking by marrying two different concepts: global optimization and social and group behaviors. In this work, the author formulates multi-target tracking task as a problem of minimum-cost network flow. The author observed that pedestrians preemptively plan their movements to steer clear of collisions. Therefore, the author suggested that incorporating social and grouping behavior into a graph model is an ideal approach.

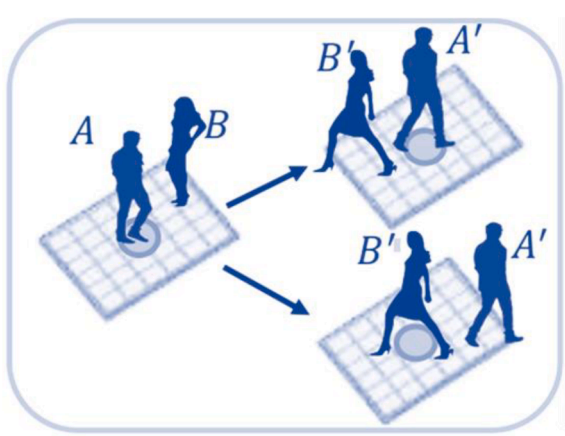


Fig. 6. Illustration of target dynamic interaction. The Figure shows how two targets (A and B) adjust their trajectories to avoid collision.

The authors in [158] introduced the concept of social LSTM, which utilizes Long Short-Term Memory networks to collectively learn the movements of multiple individuals. Their proposed architecture enables the prediction of the future paths of all targets over an extended period. Sadeghian [11] et al. utilized an LSTM-based architecture to capture prolonged temporal connections across various cues, without the requirement of manual specification of parameters or weights. Unlike reference [158], their interaction model is trained through data-driven methods to solve the reidentification task, as opposed to long-term prediction. The authors in reference [159] put forth the argument that in order to develop models capable of replicating the exceptional human ability to navigate through intricate and busy environments, the research community requires access to high-quality data as a prerequisite. With this objective in mind, the authors have created an extensive new dataset consisting of videos showcasing a range of target types. In addition to assessing cutting-edge prediction and tracking techniques, the paper introduces a fresh characterization method that delineates the degree of "social sensitivity" involved in the interaction between two targets.

In summary, existing association based MCMT methods often adopt appearance, motion, and/or dynamic interaction information to efficiently and effectively associate tracklets assessed to be similar. They are effective in handling extended occlusions between targets and the complexity increases proportionally with the number of targets. However, distinguishing between different targets based on appearance remains a major challenge that limits the performance of these methods. It can be difficult to identify unique characteristics of targets in crowded environments with frequent occlusions and similar appearances. Motion descriptors typically use speeds and distances between tracklet pairs, while appearance descriptors use global or part-based color histograms

to differentiate between targets. In recent years, various dynamic interaction models have also been proposed including those considering only spatial interactions alone as well as the ones taking into account both the spatial and temporal continuity.

2.2.1.3. Optimization method. Numerous recent advancements in the tracking of multiple targets seek to identify an (almost) optimal collection of trajectories within a given time frame. To manage the vast universe of potential trajectory hypotheses, it is frequently discretized to a finite set in a data-driven or regular manner. Most of this research sees target tracking as a problem of minimization or maximization (e.g., maximizing the joint posterior of several dependent variables) that can be solved with either a local or global search strategy. Effective techniques that run in polynomial time, such as the Hungarian algorithm [160], weighted bipartite graph [161], and binary integer program [162] have been employed successfully by framing the problem as an optimal assignment [29,120] or optimal matching [5,42]. The basic concept is that each detection must be matched to a subsequent detection and that chaining these matches, results in tracks [163]. In the paragraphs that follow, we present an overview of different optimization approaches used in MCT.

(a) Global search approach

The emergence of global optimization trackers, which can function on longer time scales and use high-level reasoning to resolve ambiguities, was prompted by the understanding that local knowledge may not always be sufficient to make the right decision. These methods aim to find the global optimal solution for the entire sequence of frames. Some examples of global optimization methods used in MCMOT include:

i. *Integer Linear Programming (ILP)* - This method has been adopted by several recent literature [23,28,43,164] and it formulates the problem as a linear program solved by an integer linear solver. It considers the entire sequence of frames and finds the optimal solution. This method is known to be computationally expensive, but it can find the global optimum.
ii. *Dynamic Programming* - This method uses a recursive approach to find the optimal solution. It breaks the problem down into subproblems and solves them one by one. It is a powerful method, but it can be computationally expensive. This approach is followed by the authors in [161,165,166].
iii. *Belief Propagation* - This method uses a message-passing algorithm to propagate information between the variables in the problem. It can be used to solve problems with a large number of variables and can find the global optimum. MCT algorithm proposed in [167–171] utilized this method to find an effective global solution to its data association problem.
iv. *Relaxation Labeling* - This method relaxes the problem by easing the constraint on the variables and then finds the optimal solution by iteratively tightening the constraint. This is the approach followed by researchers in [117,172,173].
v. *Branch and Bound* - This method is a type of depth-first search algorithm. It explores the solution space by branching on the decision variables and pruning branches that cannot lead to the optimal solution. This method is often used to solve problems with a large number of variables.
vi. *Multi*-commodity network flow: This algorithm is frequently for solving data association problems formulated as network flow. It generates multiple layers within the network-flow tracking graph, with each layer corresponding to a distinct identity group. This is the approach followed by Ben et al. in [174].

Global optimization methods are a class of optimization techniques that are commonly used in multi-camera multi-object tracking (MCMOT) systems. For example, in [89], a global search approach was suggested as an addition to an earlier work that made target tracking into an integer linear program by dividing the target location space into smaller pieces (discretized). The work showed how dynamic models can be used well in tracking formulations with objective functions that can be globally optimized. In their research, Zhang et al. [175] introduced an optimization technique based on network flow for solving the multi-object tracking data association problem. They mapped the maximum-a-posteriori (MAP) data association problem onto a cost-flow network, which included a constraint preventing trajectory overlap. Using a min-cost flow algorithm, the authors compute the optimal data association in the network. Similarly, the authors in [52] suggest an improvement to existing works that cast object tracking as a high-level global optimization problem. Their study primarily aims at resolving the long-standing issue of accurately modelling split and merge events in a global optimization problem that can be solved in polynomial time. Over the years, several other studies [6,94–95] also adopt this kind of optimization approach to seek their intended solution to multitarget tracking. However, in the field of multitarget tracking, standard descriptions of desirable qualities do not result in models that can be globally optimized [51,53]. Moreover, global optimization methods are computationally expensive, but they can find the global optimum. The choice of global optimization method depends on the specific requirements of the application and the trade-off between computational complexity and accuracy.

(b) Local search approach

Using this method, local search algorithms typically operate on a single candidate solution and include incrementally making tiny modifications to the candidate solution and evaluating the change to see if it results in an improvement and is taken as the new candidate solution. The goal of a local search algorithm is to find a local optimum. It is well suited to traversing a particular region of the search space and getting close to (or finding exactly) the extrema of the function in that region. Specifically, these methods focus on finding the optimal solution for the current frame or a small set of frames, rather than considering the entire sequence. Some examples of local optimization methods used in MCMOT include:

i. *Greedy algorithm* - This method selects the best detection-to-track association at each step based on a predefined cost function. This algorithm is simple and fast, but it may not find the global optimum. This optimization solution was followed by the authors in [10,176–178].
ii. *Evolutionary algorithm* - This method mimics the process of natural evolution to find the optimal solution. It starts with a population of randomly generated solutions, and iteratively applies nature-inspired operators (e.g., crossover and mutation) to evolve the population towards the optimal solution. For instance, Tian et al. [179] solved data association problem by formulating the step as joint probabilistic problem involving particle filter (JPDA-PF) optimized with firefly algorithm.
iii. *Simulated Annealing* - This method is inspired by the process of annealing in metallurgy. It starts with a random initial solution and iteratively applies small random perturbations to the solution. The perturbations are accepted or rejected based on a predefined acceptance criterion, such as the Metropolis criterion.
iv. *Branch and price* - This method is a type of depth-first search algorithm. It explores the solution space by branching on the decision variables and pruning branches that cannot lead to the optimal solution. This method is often used to solve problems with a large number of variables. This method is followed by the authors in [180,181].

MCMT approaches solely built on local optimization methods are

relative as compared to their counterpart based on global search methods. Among the rare works adopting this method is that proposed in [100]. In this paper, the authors proposed an MCMOT algorithm that combined bottom-up and top-down tracking methods using cooperative particle filter and artificial potential field. However, to achieve better accuracy under more crowded scenarios, the author also employed the result of global optimization. Thus, local optimization methods are often used in combination with global optimization techniques to improve the overall performance of MCMOT systems. The choice of local optimization method depends on the specific requirements of the application and the trade-off between computational complexity and accuracy.

2.3. Mutual exclusion constraints

Mutual exclusion constraints (MECs) between various tracklets ensure likely between-target interaction [52], and consequently effective data association that yield accurate tracking performance. One common as well as fundamental physical constraint in all tracking problems is that two objects cannot exist in the same location at the same time [118]. This constraint is pivotal to the well-known collision avoidance that is frequently used by most multi-target tracking task for plausible target interaction modelling [52]. Another exclusion limitation entails that any given observation can only pertain to a single target. This constraint prevents the problem of identity switching, where a detection is associated with the wrong track and causes the identities of the tracks to become confused. To incorporate mutual exclusion constraints into MTMCT, several approaches have been followed in the literature, these include:

i. *Graph-based optimization* - MECs can be incorporated into the graph-based optimization by adding constraints on the edges of the graph, ensuring that no two edges are connected to the same node.
ii. *Integer linear programming (ILP)* - MECs can be incorporated into the ILP formulation as binary constraints on the decision variables, ensuring that no two tracks can be associated with the same detection.
iii. *Particle filtering* - MECs can be incorporated into particle filtering by applying the constraints during the resampling step, ensuring that no two particles represent the same object.
iv. *Deep Learning-based methods* - MECs can be incorporated into deep learning-based methods using the loss functions, this could be done by using the “triplet loss” or “max-margin loss” functions that encourage the model to learn the similarity between objects and dissimilarity between different objects.

Over the years, researchers have emphasized the need to include various mutual exclusion constraints for efficient tracklet matching. These constraints may include collision avoidance [52,118], spatial locality [12,182], temporal locality [11,12], motion consistency [155], appearance consistency [155] and grouping consistency. Andriyenko and Schindler [52] noted that the simplest form of mutual exclusion constraint, which has been widely used in multi-target tracking, is the collision avoidance. Also, they solved a data association problem by formulating it as an integer linear program that includes collision avoidance as the mutual exclusion constraint. Their approach also utilizes appearance and motion models. Similarly, the authors in reference [118] integrate collision avoidance into a continuous energy function by introducing a continuous exclusion component. This factor penalizes configurations in which two targets approach too closely, with the penalty reaching infinity when they occupy the same position. The term also enforces a distinct data association at all times. The mutual exclusion and spatial locality constraints of identity-aware tracking were modeled by Yu et al. [182] using a $\ell 0$ norm constraints, while formulating the multi-target tracking problem as a quadratic program. In this work, the solution path algorithm was applied to finding the global optimal result. To sum up, MECs are an important component in MTMCT to ensure the correct association of detections to tracks, this is particularly important when dealing with crowded scenes, where multiple objects can have similar appearances and can be occluded by other objects.

3. Elements of MCT

In this section, we outline the main elements of an MCT method. As was already noted, the objective of MCT is to simultaneously identify several objects in a single frame and recover identifying information across a continuous frame sequence from the same and/or various cameras—in other words, trajectory, from a given sequence from the same and/or different cameras. Two key questions should be considered when developing MCT techniques. The first question is how to calculate object similarity across frames and cameras. The second problem is determining how to retrieve identification information based on similarity measurements between objects in frames from within and across cameras. These two issues are commonly considered as a joint SCT and ICT problem and can thus be viewed as a problem combining modeling of appearance, motion, interaction, occlusion, exclusion, and data inference. The following subsection reviews current developments with reference to both subjects.

3.1. Single camera tracking

In the context of single-camera tracking, which typically involves a relatively straightforward scenario of tracking a single object of interest within a camera’s field of view, the tracking process can be simplified. It entails searching for the desired object within the expected region of the scene and establishing a coherent trajectory by connecting the object’s positions over time. By employing this approach, the system can effectively track and monitor the movement of the targeted object within the captured video footage. However, the estimation of the spatio-temporal information of objects is more dependable in multi-object scenario. When a greater, and frequently unknown, number of targets are detected simultaneously, the task becomes exponentially more complex because it is no longer evident which object belongs to which detection [29,51,183]. Tracking under single-camera scenarios, and the appearance of the objects does not change significantly throughout tracking [53,63]. However, due to the presence of numerous challenging scene constraints, there exist reliability concerns for the predicted spatio-temporal information in multi-camera multi-object tracking (MCMOT). Thereby making the problem of multi-object tracking under single-camera settings (SCMOT) relatively simpler than MCMOT. Fig. 7 Illustrate instance of outputs from MCMOT framework. In many recent SCT methods, the targets are first identified in a preprocessing step, typically by background subtraction [184] or by using a discriminative classifier [96,99], from which the trajectories are subsequently estimated [51].

Tracking-by-detection remains the prevalent tracking framework followed by most researchers in tracking. This is due to its enhanced drift resistance and the potential for tracking failure recovery. In recent years, several research efforts [185–189] have also attested to the unprecedented performance improvement in SCT, particularly brought by this tracking paradigm. For instance, while adopting the tracking-by-detection scheme, Yin et al. [190] proposed a SCT approach that unified into a single network, the object motion and metric learning model for target association. In that study, the MOT method proposed was able to conserve additional computational overhead typically associated with repetitive feature extraction for SOT and affinity computation via its utilization of multi-task learning. Similarly, the authors in [188] also addressed the problem of feature discrepancies between detection and re-identification in a category of tracking-by-detection called joint detection and embedding (JDE). In this work, the author introduced a channel-based Decoupling module and Cross-direction Transformer as a

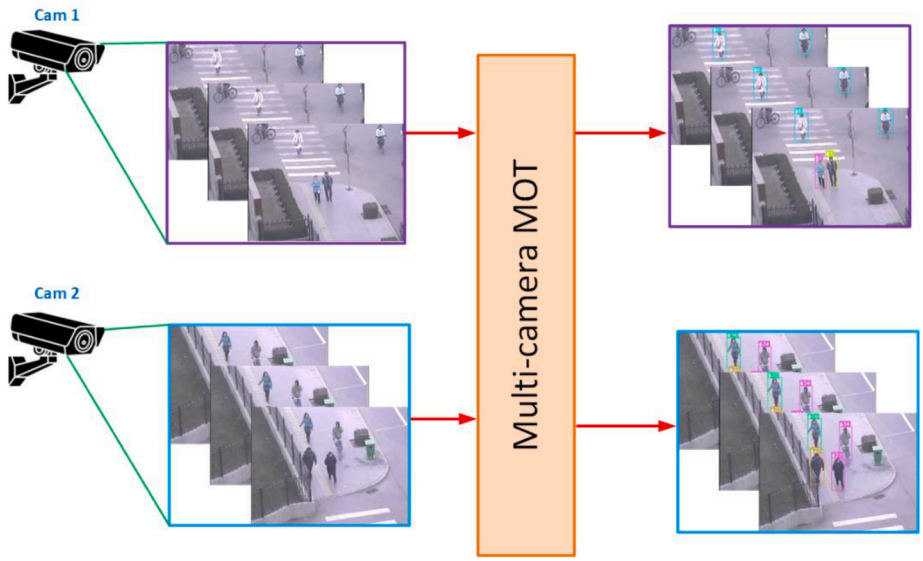


Fig. 7. Illustration of the output of a multicamera multi-object tracking (MCMOT) algorithm. Each output bounding box has a unique number which identifies a particular object in the video.

solution to address feature misalignment. While the focus of this paper is not delved into the extensive advancements of MOT under single camera setup, this subsection aims to concentrate on prior research in visual multi-object tracking within a multi-camera setup. These approaches can generally be categorized based on various factors such as online or offline methods, detection-based or detection-free approaches, recursive or non-recursive techniques, and so on (depending on the specific modality under consideration).

3.1.1. Online-based SCT methods

The aim of online multi-object tracking is to generate accurate tracks of multiple objects using only the information available at the present time. It remains a difficult task in complex environments due to issues such as similar object appearances, frequent obstruction by other objects or background, and various other photometric and geometric factors. Among the most simple and successful MOT approaches based on this technique is the Simple online multi-object tracking (MOT) methods, such as SORT [71], DeepSORT [69], ByteTrack [72], and FairMOT [49]. These methods have been proposed to address various challenges of tracking multiple objects in real-time under single camera setting. They rely on using only the history information of the objects, meaning that the information from previous frames is used to predict the location of the objects in the current frame. Recently, several authors provided crucial improvements to these models, and extend their application to tracking under multiple cameras setup [29,30,120,148,191–194].

Among the key components of these methods is an accurate appearance re-identification model [191], and precise within- as well as across-camera association frameworks [23,30,86]. The appearance model is used to identify objects based on their appearance, such as colour, shape, and texture. By matching the appearance of an object in the current frame with its appearance in previous frames, the re-identification model is able to link object detections across multiple frames, forming a track [20,31,49,88,148,192]. These short tracklets are then associated across multiple cameras (global tracklets) according to the movement and/or interaction of the targets. However, online-based approach has its limitations when dealing with occlusion, which occurs when one object blocks the view of another object or prolonged disappearance in blind-spot [63,195]. In these cases, the appearance of the occluded object may be difficult to match with its previous detections, leading to a loss of track or identity switch [32,148,196]. To overcome this limitation, these methods rely on the accurate appearance re-identification model for long-term tracking, meaning that the model must be able to accurately identify an object across a long period of time, even when it is occluded.

3.1.2. Offline-based SCT methods

Offline-based multi-target tracking refers to a method for tracking multiple objects in a video sequence, which relies on information from both past and future frames. A common approach in offline-based target tracking is to construct a graph that captures the visual and spatio-temporal similarities among the objects being tracked [191]. The graph is then optimized to find the best solution for tracking the objects. Moreover, the batch optimization technique is frequently followed in this kind of trackers [197]. One of the key advantages of offline-based multi-target tracking is that it often achieves better performance than other tracking techniques [198]. This is because it considers both visual and spatio-temporal similarities contained in past, current, and future frames, which can provide a more robust solution for tracking multiple objects.

When it comes to MTSC tracking, offline methods typically yield superior performance because all the tracklets that have been aggregated can be utilized for data association. Over the years, several MTMCT methods [23,197] utilize this category of SCT to performance intra-camera tracking in MTMCT. Such MTSCT include RCT [199], TAO [200], TPM [201], FWT [202] and many more. For instance in [23], the authors formulate tracking in multi-camera scenario as a maximum clique problem based on graph theory, and adopt the offline-based approach for learning a robust solution. In [197], the authors note that the difference between local association processes and the global nature of re-identification appearance cues can significantly impact the performance of MTMCT. To address this issue, they adopt an offline-based solution approach and adjust the local matching procedure accordingly.

Re-ID features can greatly improve the performance of MTMC tracking, but the reliability of these appearance features is reduced by orientation and appearance variation, making direct application of Re-ID unfeasible. To overcome this challenge, Li et al. [195] proposed an approach that employs MTMC_ReID and integrates an offline single-object tracking algorithm [66] to generate tracklets. Owing to the intense computational requirement of graph optimization on which majority of offline trackers are based, their applicability especially for real-time scenario is limited. Overall, offline-based multi-target tracking

is a powerful technique for tracking multiple objects in a video, but it comes at the cost of increased computation time. It can be used in applications where computational resources are available and a high level of performance is required.

3.1.3. Detection-based SCT

Detection-based multi-camera tracking refers to a method of tracking an object or person across multiple cameras that relies on results from detection algorithm (i.e., detection responses). In this approach, the object or person is first detected in each camera's field of view, and then the detection responses are linked together to track the object or person as they move across the different cameras. Visual tracking approaches based on this scheme are generally referred to as the Tracking-by-detection method. They have improved a lot in recent years, especially with the development of object detection algorithms [49,69,71,72,199–201] that work well even in crowded scenes. The tracking-by-detection methods can roughly be classified into batch and online-based techniques [203,204].

i. *Batch methods:* These methods process all the frames of the sequence at once and find the optimal solution for the entire sequence. The optimization is done after all the frames have been processed. These methods are often used when the entire video sequence is available in advance and the computational resources are sufficient. Integer Linear Programming (ILP) and Dynamic Programming are common optimization techniques adopted by trackers based on these methods. Trackers contributed by authors in [28,51,178,205] belong to this category.
ii. *Online methods:* These methods process the frames one-by-one in real-time and update the solution as new frames arrive. These methods are often used in real-time applications where the entire video sequence is not available in advance and the computational resources are limited. Greedy Algorithm, Particle Filtering and Kalman Filter are examples of online methods, and the approach described in [151,203,204,206] can be categorized into this class.

Batch methods are known for providing a global optimal solution for the entire sequence but at the cost of high computational resources and time, online methods are known for providing real-time solutions but at the cost of lower accuracy and global optimality. The choice of method depends on the specific requirements of the application and the trade-off between computational complexity, accuracy, and real-time performance.

3.1.4. Detection-free SCT

One of the challenges in tracking multiple objects simultaneously is the similarity of object appearances and the occlusion and interaction between objects [40]. Detection-free multi-camera tracking offers a solution to these challenges by leveraging appearance-based tracking methods across multiple cameras. By utilizing features like colour, texture, or shape, the system can track objects or people without the need for explicit detection algorithms [207–209]. In research conducted on multi-object tracking, it has been observed that single camera-based systems may struggle to accurately track multiple objects due to horizon limitations, object occlusion, and ambiguous appearances [210,211]. To address these limitations, a new approach has now been investigated, particular for multi-object tracking under multi-camera scenario. This method aims to improve tracking accuracy by integrating information from multiple cameras, enabling better object tracking in complex scenarios.

Detection-free multi-camera tracking refers to a method of tracking an object or person across multiple cameras where tracking and detection are performed in one single step. In this approach, the object or person is tracked based on their appearance or discriminative features. This evolving approach can be more efficient and less computationally intensive than detection-based tracking, particularly when the number of objects in the scene is large or the objects are small. There are several methods for detection-free multi-object tracking, including:

i. *Image-based tracking:* This method tracks objects based on the appearance of the objects in consecutive frames of the video [207,208].
ii. *Feature-based tracking:* This method tracks objects based on salient features such as color histograms, texture, or edge information [212,213].
iii. *Motion-based tracking:* This method tracks objects based on their dynamic motion, such as their position, velocity, and acceleration [214,215].

While detection-free multi-camera tracking provides a promising alternative to traditional multi-object tracking methods, this kind of multi-object tracking, particularly under multi-camera setting can be a challenging problem due to variations in lighting, object occlusion, viewpoint variations and changes in object appearance. However, it is a crucial step in achieving accurate and robust multi-object tracking in real-world scenarios, such as surveillance and robotics.

3.2. Inter-camera tracking

Inter-camera tracking (ICT), also referred to as multi-camera tracking (MCT), is a technique used to track objects as they move across multiple cameras. The primary goal of ICT is to identify the same object in various camera's field of view, and then link the observations from each camera to create a consistent "track" of the object's movement across the entire camera network [132]. This method is used to establish "target correspondence" among the observations from multiple cameras, which allows for consistent labeling and tracking of the objects being monitored [99]. Furthermore, it is concerned with re-identifying the target trajectories across multiple cameras by object re-identification (Re-ID). However, this remains a challenging task as objects may move in and out of the field of view of different cameras and may be occluded by other objects or by the environment. Also, the illumination and pose of each camera varies and must be considered during problem formulation. Prior research effort in ICT only attempted to address cross-camera tracking problems, while relying on off-the-self SCT method, which are only built with intra-camera data association scheme. These approaches tackle this problem using either the tracklet-to-tracklet matching method or tracklet-to-target assignment approach.

3.2.1. Tracklet-to-tracklet matching approach

Many studies on cross-camera tracklet matching, also known as inter-camera data association, utilize the tracklet-to-tracklet matching method. Some of these studies either match tracklets between every two neighboring cameras until all cameras have been matched or use expert approaches like greedy matching or hierarchical clustering to iteratively match local tracklets across all cameras [29]. To efficiently associate corresponding tracklets across camera, Bredereck et al. [34] proposed a greedy matching algorithm that relies on multiple judgements from likelihood functions. In this work, the authors also exploit the spatial positioning of the cameras in the system, which is computed via triangulation method. Similarly, Berclaz et al. [28] employed the output of object detectors to generate Probabilistic Occupancy Maps (POM) for estimating the maximum number of objects and determining their locations, which were subsequently used for association through graph matching. The authors in [66] employed the hierarchical clustering approach to match trajectories and tracklets across camera. In this paper, the authors describe a matching process that begins by calculating a distance matrix between each tractlet and trajectory. The matrix takes into account factors such as appearance similarity, overlap, and separation. As the clustering process continues, trajectories are combined, and the distance matrix is updated until the minimum distance exceeds a certain threshold.

In [165], the author proposed a hierarchical composition model for associating detection responses into tracklets. To address the issue of inference in the proposed model, the authors employed the Bayesian formulation for solving maximum a prior problem. Tesfaye et al. [88] suggested a fast-constrained dominant set clustering approach for associating nodes (tracks) in a graph. The proposed approach does not only consider similarities between tracks observed across the camera, but also takes into consideration similarities among tracks from the same camera. Over the years, different techniques such as pruning using camera topology [31] and adaptive attribute selection [165] have been suggested for reducing the search space, thus improving the efficiency of the matching process. Additionally, there are several research efforts that aim to find a global solution for tracklet matching by using probabilistic state space methods. These methods can include a Bayesian formulation [166], graph models [53,180], or updating particle filters [216].

The tracklet-to-tracklet matching scheme has two main issues. Firstly, since different targets are captured by different numbers of cameras, it is difficult to determine how many local tracklets should be grouped into a global trajectory for each target. Secondly, it is challenging to ensure that the matching of local tracklets is consistent and mutually exclusive in practice. Despite various research efforts to impose constraints such as sparsity and time conflict, an elegant solution to impose the matching consistency principle has not yet been found [29].

3.2.2. Tracklet-to-target assignment approach

Tracklet-to-target assignment approach reformulates the cross-camera tracklet matching problem as that involving the assignment of each tracklet to a unique target using a specialized optimal assignment algorithm. This approach aims to address issues earlier identified in tracklet-to-tracklet matching technique. This method was first suggested by He et al. [29]. In this paper, each tracklet is assigned to a specific target and the best assignment is determined using the Restricted Non-negative Matrix Factorization (RNMF) algorithm. In [90], following the two-step approach, the authors suggested a MTMCT approach to robustly track vehicles in a camera network. Following the tracking-by-detection paradigm, the proposed method first generates all local tracklets, and tracklets belonging to the same target are then matched together into a global tracklet across different cameras using the TRACTA approach.

4. Multi-camera tracking evaluation benchmarks

A procedure for the systematic and objective evaluation of a multi-camera tracker's characteristics is a method used to measure and compare the performance of different multi-camera tracking systems. This procedure is important in order to ensure that multi-camera tracking systems are reliable and accurate, and that they can be effectively used in different applications [217]. The evaluation procedure typically involves the use of a set of well-defined metrics, such as tracking accuracy, robustness to occlusions and clutter, and computational efficiency, to evaluate the characteristics of the multi-camera tracking system [218]. This procedure helps researchers and developers to identify the strengths and weaknesses of different multi-camera tracking systems, and in making informed decisions about which systems are best suited for a particular application. Additionally, it aids in the comparison of different approaches, and helps to identify the state of the art in multi-camera tracking research. In the following, we examine different existing metrics for MCT evaluation.

4.1. Performance metrics

The selection of appropriate metrics for evaluating the performance of tracking algorithms is undoubtedly difficult and may be influenced primarily by the application under consideration [219,220]. Furthermore, a comprehensive tracking assessment must evaluate a few factors, for which several metrics have been suggested over the years [164,218,219,221–225]. To evaluate the effectiveness of the proposed tracking algorithm, existing works on multi-camera tracking frequently employed the same metrics adopted by popular single camera tracking evaluation initiatives like the Multiple Object Tracking (MOT) challenge [218,219] and the PETS tracking challenge [220]. These metrics include False Alarms per Frame (FAF), Multiple Object Tracking Precision (MOTP), Multiple Object Tracking Accuracy (MOTA), number of Identity Switches (IDS), ratio of the Mostly Tracked trajectories (MT), ratio of the Mostly Lost trajectories (ML), and number of track Fragmentations (Frag). Specifically, these metrics are defined as belonging to either the classical metric, the CLEAR MOT metric, or the ID score categories. In this subsection, we present a detailed description of each category as well as the metrics belonging to the category.

4.1.1. Trajectory-based metrics

Over the years, several metrics have been proposed for evaluation of MOT methods. However, only a few of them are adopted by benchmarks for ranking the performance of trackers. Among the most common metrics, though only employed as secondary measures on benchmark, are the trajectory-based metrics proposed in [224,226]. The authors in these papers define the metrics for evaluation of multi-target tracking methods as those that underscore the various kinds of failures that a MOT algorithm may commit. They are mainly the basic building blocks of other metrics, and can be computed to illustrate faults like:

i. *ID switches (IDS):* This metric counts the instances in which an object is correctly tracked but its associated identity is incorrectly swapped.
ii. *Mostly Tracked Trajectories (MT):* This metric measures the number of trajectories that are tracked correctly in at least 80% of the frames.
iii. *Mostly Lost Trajectories (ML):* This measures the number of trajectories in more than 80% of the frames that are lost.
iv. *False trajectories*: These are result trajectories that are unrelated to any actual object in the frame or entire video sequence. This metric encompasses both the number of False Positives (FP) and the number of False Negatives (FN).
v. *Fragmentation*: If the tracker changes the identification of a trajectory in frame t but the accompanying ground-truth identity remains the same, a fragmentation has occurred.
vi. *Merge*: This error occurs when the tracker blends two distinct ground truth identities into one between frames t' and t. Merge is the opposite of fragmentation:
vii. *Merger Index*: A pair of truth tracks contain target pairs—one target of each track—that are associated with a common system track. The performance is degraded as the number of merged truth track pairs increases and these track pairs become more severely merged.
viii. *Mean Deviation:* This measures the meaning of the variations between the result location of the system target and its associated ground truth. the performance degrades, as the variations increase.

In summary, metrics belonging to this category are never used independently to rate trackers since they are too simplistic, generally focused on only one form of tracking error apiece, and easily gamed if desired [227]. In Fig. 8, we provide an illustration of some of these tracking faults (where thin lines denote ground truth and line of same color represents tracks of same target).

4.1.2. CLEAR MOT metrics

Motivated by the lack of common metrics for measuring the performance of multiple object trackers, Bernardin and Stiefelhagen [222] proposed two intuitive and general metrics for the objective comparison

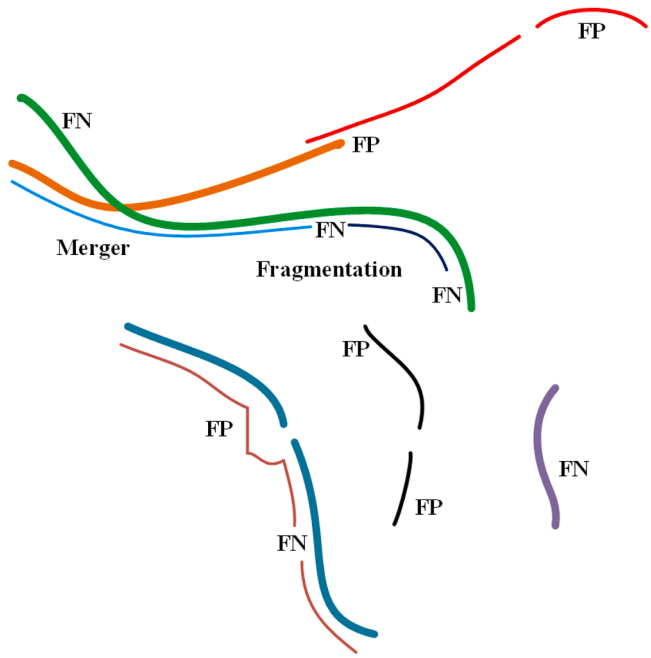


Fig. 8. Illustration of some basic errors in tracking. The errors depicted include False Negative (FN), False Positive (FP), Fragmentation and Merger.

of multiple object trackers' characteristics. These metrics emphasize the tracker's precision in estimating object locations, accuracy in recognizing object configurations, and ability to label objects consistently over time. These metrics were extensively used in two large-scale international evaluations, the Classification of Events, Activities, and Relationships (CLEAR) workshops in 2006 and 2007, to measure and compare the performance of multiple object trackers for a wide range of tracking tasks. These metrics are namely, Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP). They function as a summary of the comparatively simple metrics that comprise them, and can formally be define as follows:

MOTP: This metric demonstrates the tracker's capacity to estimate precise object positions independently of its ability to recognize object configurations, maintain consistent trajectories, and so on. It is defined as the total error in estimated location for true positives and their ground truth, divided by the total number of matches made. This can be computed for bounding box as:

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t m_t} \tag{4.1}$$

where m_t is the total number of matches in frame t, and $d_{t,i}$ denotes bounding box overlap between the detected target i and its assigned ground truth in frame t.

MOTA: Owing to its expressiveness as it combines three tracking error ratios, MOTA is the most widely employed metric for MOT performance evaluation. The MOTA takes into consideration various object configuration faults, false positives, misses, and mismatches across all frames. Formally, it is defined as:

$$MOTA = 1 - \frac{\sum_t (h_t + FP_t + mm_t)}{\sum_t f_t} \tag{4.2}$$

where h_t denotes the number of misses, FP_t is the number of false positive, and mm_t is the number of mismatches, all in frame t. Here we denote the number of objects in each frame as f_t. Moreover, recent years have been characterized by a few improvements proposed to the MOTA metric. These can be found in the newly suggested metrics like PR-MOTA [228] and AMOTA [229] usually employed in 3D MOT evaluation.

4.1.3. Hand-over errors and MCTA

Because they are not specifically introduced for multi-camera tracking evaluation, MOTA and its companions, if extended to evaluating multi-camera tracking methods, perform poorly in accessing across camera errors [164]. This is primarily due to the fact that a trajectory covering n_f frames from n_c cameras has only approximately n_c across-camera detection linkages between subsequent frames and about $n_f - n_c$ within camera one, and $n_f \gg n_c$. A different set of metrics (namely handover failures and multi-camera tracking accuracy (MCTA)) have been developed and employed in literature [29,30,43] over the years, to address this issue.

Handover errors: These errors distinguish between fragmentation and merges that occur when tracking across cameras. Fragments and mergers caused by camera handover can be classified as crossing and returning faults. These more precise handover error scores aid in the understanding of various sorts of tracker failures, and within-camera errors are measured separately using conventional metrics.

MCTA: This metric consolidates all aspects of the system performance, including precision, recall, normalized within camera and across-camera identity mismatches, into a single measure. It is mathematically formulated as:

$$MOTA = 1 - \frac{\sum_t (h_t + FP_t + mm_t)}{\sum_t f_t} \tag{4.3}$$

where P and R represents precision and recall respectively, while Mm^{wc}, T^{wc} Mm^h, and T^h denotes the identity mismatches and total number of true detections within-camera (wc) or during handover (h) respectively.

4.1.4. Identity-based metrics

The previously discussed measures are generally regarded as event-based and are instructive for system component designers since they assist in locating the origin of some failures. However, identity-based metrics are essential in the interest of users in contexts where maintaining identity is critical, such as in sports, security, or surveillance. Metrics in this category assess how well predicted identities correspond to ground-truth identities without concern for where or why errors arise. These metrics include identification recall (ID-Recall), identification precision (ID-Precision) and identification F1 (IDF1) and are suggested particularly for evaluating MCMT methods. They are applicable both for within and across cameras tracking assessment. Both ID-Recall and ID-precision are both ratio of predicted detections that are precisely identified. While the ratio of correctly identified detections to the average number of ground-truth and predicted detections is defined as IDF1. The mathematical expression for each of these metrics is as follows:

$$\text{ID} - \text{Recall} = \frac{\text{IDTP}}{\text{IDTP} + \text{IDFN}} \tag{4.4}$$

$$\text{ID} - \text{Precision} = \frac{\text{IDTP}}{\text{IDTP} + \text{IDFP}} \tag{4.5}$$

$$\text{IDF1} = \frac{\text{IDTP}}{\text{IDTP} + 0.5\text{IDFP} + 0.5\text{IDFN}} \tag{4.6}$$

where IDTP, IDFP, and IDFN is the sum of the numbers of identification true positive, false positive, and false negative respectively, all over the entire frames and cameras.

4.1.5. HOTA metrics

The higher order tracking accuracy (HOTA) was presented by Luiten et al. in [230] to address the issue of overstated importance of either detection or association in existing evaluation metrics. For the purpose of comparing trackers, this metric explicitly balances the impact of carrying out precise detection, localization, and association. As a result

of HOTA's ability to examine each of the five fundamental fault types separately, a clear analysis of tracking performance is possible. While HOTA can be utilized as a single unified metric for assessing trackers, it also decomposes into a family of sub-metrics (refer to Fig. 9) which separately evaluate various characteristics of the trackers. HOTA score ($HOTA_\alpha$) is simply the geometric mean of the detection accuracy score (DetA) and the association accuracy score (AssA) averaged over different localization thresholds. Where DetA is defined as the percentage of aligning detections, and AssA, is the average alignment between matched trajectories, averaged over all detections.

4.1.6. Track-mAP metrics

Track-mAP metric, also known as 3D-IoU, is distinct from other metrics earlier described in that it works on a set of tracks, which are rated according to the tracker's level of confidence that each track exists rather than operating on a set of fixed tracks. These metrics include average precision (AP) and average recall (AR) and were initially used for the ImageNet-video benchmark's evaluation [225]. The area under the precision-recall curve (a curve plotted using the confidence score of each detection) is defined as AP, and it is averaged over multiple intersection-over-union (IoU) thresholds. The maximum recall given some number of detected/segmented instances per video is defined as AR. Both metrics are examined per category and then averaged across all categories. The mean average precision (mAP) has lately been adopted by a number of other benchmarks, including YouTube-VIS [231], VisDrone [232], and TAO [233].

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \tag{4.6}$$

where AP_k is the average precision for tracklets of target k and n is the number of targets.

4.2. Benchmark datasets

Numerous multi-camera tracking datasets have been adopted in multitarget multi-camera tracking research. These include USC Campus [234], PETS2009 [220], DukeMTMC [164], CamNet [235], NLPR MCT [236], etc. CVLAB EPFL sequences [183] such as Passage, Campus etc., and DukeMTMC [164] datasets remained the most frequently employed benchmark datasets for evaluation of MCMT methods. These two datasets are widely used because the video sequences in them are from either of the two different multiple camera arrangements. The former depicts situations in which there is overlap among cameras, and the latter depicts situations in which there is little to no overlap. Recently, for city-scale tracking of vehicles in a camera network, the CityFlow [192] datasets have been contributed, and it is rapidly gaining traction. Like DukeMTMC, this dataset also portrays a non-overlapping arrangement. In Table 2, we compare several existing datasets commonly employed by researchers in multi-target multi-camera tracking. In paragraphs that follow, we will provide a detailed overview of some of these commonly used benchmark datasets.

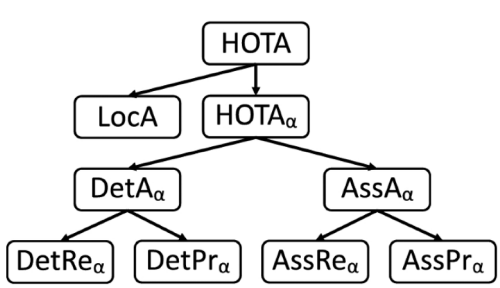


Fig. 9. Diagrammatic illustration of HOTA family of sub-metric [230].

4.2.1. PETS dataset

The Performance Evaluation of Tracking and Surveillance (PETS) dataset is one of the earliest evaluation benchmark datasets in the domain of multi-camera system tracking. The primary objective for initiating PETS editions was the evaluation of on-board surveillance systems for the protection of mobile critical assets such as trucks, trains, and shipping vessels, which could be considered targets for criminals, activists, or even terrorists. The most frequently adopted PETS dataset for evaluation of MCMT methods is the PET2009 dataset (PETS09). This dataset consists of multi-sensor sequences comprising crowd scenario with increasing scene complexity [220]. In Fig. 10, we provide the dataset video sequence layout for better understanding (Table 3).

4.2.2. CVLAB EPFL dataset

The CVLAB EPFL dataset consists primarily of five sequences. All sequences feature several synchronized video streams filming the same area at different angles. All cameras are located about 2 m from the ground. The dataset also contains calibration information for the cameras as well as ground-truth annotation for all targets in each video sequence. Privacy concerns were taken care of in this dataset as all targets on the video sequence consented to their appearance knowing that the dataset will be made public [183].

4.2.3. DukeMTMC dataset

The DukeMTMC dataset contains over 7000 single-camera trajectories and more than 2000 unique identities obtained from eight different cameras deployed in various locations on the campus of Duke University. Each camera used for capturing this dataset contributes 85 min of video, thus the duration of all video total in 10 h. The dataset is one of the largest in this domain. However, owing to privacy concerns, the data has been retracted in recent times and its usage has been seriously warned against.

4.2.4. CityFlow dataset

The CityFlow dataset was contributed for evaluation of Multi-Camera Vehicle Tracking methods. The dataset was originally contributed to enhance research efforts in vehicle tracking and re-identification for Intelligent Transportation System. This dataset has been adopted as the evaluation benchmark in one of the major tracks of AI city challenge workshop since inception. It includes 3.58 h of video footage gathered from 46 cameras located at 16 intersections in a mid-sized U.S. metropolis. The longest simultaneous camera distance is 4 km, and the dataset includes a wide range of place categories, such as intersections, route segments, and highways. The dataset comprises 313,931 bounding boxes for 880 different annotated vehicle identities in total, with only vehicles that passed past at least two cameras being annotated.

4.2.5. LIMA dataset

LIMA dataset was contributed by [113] to bridge the gap between the contemporary research focus and the video surveillance systems end users' demand to address high-level problems. The dataset encompasses the issues of monitoring a typical indoor environment like office or home setting. It contains 4.5 h of RGB-D video from three cameras monitoring a four-room house. The dataset includes target and dresses changes to ensure the global reasoning is a realistic open-set problem, reflecting the challenges of an actual practical application. The authors provide identity annotation for benchmarking in addition to the raw video data.

4.2.6. WILDTRACK dataset

The WILDTRACK dataset is a comprehensive and high-resolution dataset captured using seven stationary cameras in a public open area. The data for this dataset was collected in front of the main building of ETH Zurich, Switzerland, during favorable weather conditions. Each sequence has a resolution of 1920x1080 pixels and was captured at a rate of 60 frames per second. The dataset features unscripted and dense

Table 2
Summary of existing datasets used in MTMCT research.

Dataset	Environment	Scene	Cam	Overlap	FPS	IDs	Resolution	Duration	Class	Year
PETS2009 [220]	–	outdoor	8	yes	7	30	768 × 576	1 min	Person	2009
USC Campus [234]	–	outdoor	3	No	30	146	852 × 480		Person	2010
EPFL [183]	Laboratory	indoor	4	yes	25	4	320 × 240	2.5 min	Person	2008
	Passageway	mixed	4	yes	25	6	320 × 240	1.7 min		
	Campus	outdoor	4	yes	25	4	320 × 240	1.33 min		
	Basketball	indoor	4	yes	25	–	320 × 240	6.23 min		
	Terrace	outdoor	4	yes	25	7	320 × 240	3.33 min		
NLPR MCT [236]	Pavement + Building	Mixed	3	No	20	235	320 × 240	20 min	Person	2015
	Lawn	Mixed	3	No	20	255	320 × 240	20 min		
	Building	indoor	4	yes	25	14	320 × 240	4 min		
	Parking + Builfing	Mixed	5	yes	25	49	320 × 240	25 min		
DukeMTMC [220]	Campus	outdoor	8	No	60	2834	1920 × 1080	85 min	Person	2016
WILDTRACK	Lawn	outdoor	7	Yes	60	–	1920 × 1080	35 min	Person	2018
CityFlowV1 [192]	Urban city	outdoor	40	no	10	666	1280 × 960	195.03 min	vehicle	2019
CityFlowV2 [192]	Urban city	outdoor	46	no	10	880	1280 × 960	215.03 min	Vehicle	2021
MMPTRACK [237]	Office	indoor	5	yes	15	–	1920 × 1080	87 min	Person	2023
	cafe	indoor	4	yes	15	–	1920 × 1080	73 min		
	Retails	indoor	6	yes	15	–	1920 × 1080	172 min		
	Lobby	indoor	4	yes	15	–	1920 × 1080	129 min		
	Industry	indoor	4	yes	15	–	1920 × 1080	117 min		

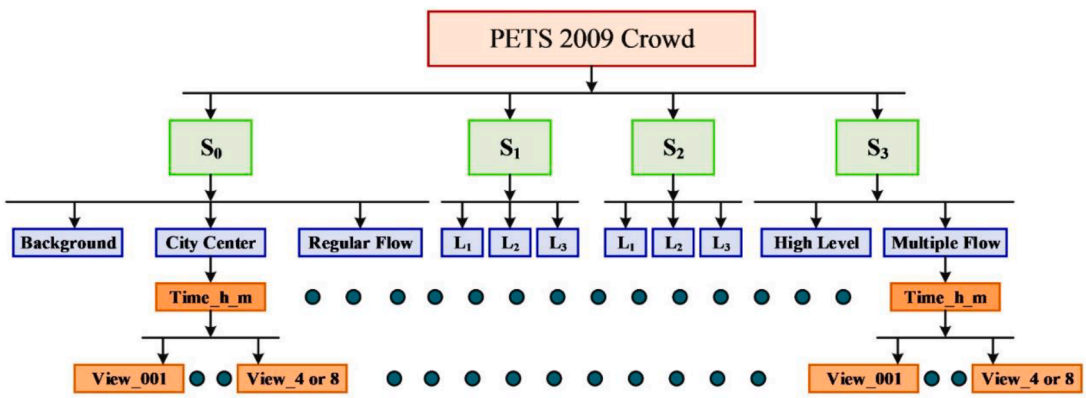


Fig. 10. Organizational structure of PET09 sequences and associated views.

Table 3
Summary of the various video sequences in the dataset.

Environment	Number of cameras	Train (seconds)	Validation (seconds)	Test (seconds)
Office	5	2760	1140	1320
cafe	4	840	1680	1860
Retails	6	5040	2580	2700
Lobby	4	3900	1920	1920
Industry	4	3120	1860	1920
Total	23	15,660	9180	9720

groups of pedestrians standing and walking in the area. Along with the camera frames, the dataset provides accurate joint (extrinsic and intrinsic) camera calibration. Additionally, the dataset includes 7 sets of 400 annotated frames for detection at a rate of 2 fps [149]. This results in over 40,000 bounding boxes that identify every person present in the area of interest, with a total of more than 300 individuals.

4.2.7. MMPTRACK dataset

The MMPTRACK dataset is a comprehensive collection of multi-objects tracking data obtained using 23 highly calibrated, overlapping cameras in various challenging indoor environments. It has a high frame rate of 15 fps and includes over 9.6 h of video footage. The dataset includes over half a million annotated frames per camera view, which were generated using an automatic annotation system that combines depth sensors and RGB cameras to produce accurate 3D tracking results. It is significantly larger than existing datasets in terms of video length and number of annotated frames. The dataset is labeled on a frame-by-frame basis, including full-body bounding boxes, with consistent identification of individuals across all cameras [237].

4.2.8. MTA dataset

Multi-camera Track Auto (MTA), a synthetic dataset generated and recorded using Grand Theft Auto 5 (GTA V), was contributed by Kohl et al. [238]. This dataset was motivated by the need for an appropriate solution to the prevalent retraction of MTMCT datasets due to privacy concerns. The MTA dataset encompasses more than 2,800 individual identities observed across 6 cameras, with each camera providing video footage exceeding 100 min in length, and 41 frames per second (fps). The collection of this dataset poses a notable challenge due to the inherent limitation of GTA V, the video game used for simulation and recording. Being a single-player game, GTA V lacks the capability to create a camera network, which in turn hampers the dataset's ability to achieve essential multi-camera time synchronization.

4.2.9. MultiviewX dataset

The MultiviewX dataset was recently contributed by authors in [239] and serves the specific purpose of facilitating multiview pedestrian detection tasks. Leveraging the capabilities of the Unity engine,[1] the

[1] https://unity.com/.

dataset was meticulously constructed to generate the required scenarios. Notably, the unique pedestrians in the dataset were adopted from the comprehensive collection of human models in PersonX [240]. Covering a slightly smaller area measuring 16 m by 25 m, the MultiviewX dataset comprises 6 cameras with overlapping field-of-view. Each camera captures high-resolution images at a resolution of 1080 × 1920 pixels. To facilitate annotation and analysis, the dataset includes annotations for 400 frames sampled at a rate of 2 frames per second (fps).

4.3. Comparison of some publicly available MCMT algorithms

In this survey we aim to include MCT methods from as diverse origins

Table 4
A comprehensive comparison of the thirty (30) selected MCMT algorithms based on the six crucial aspects.

Tracker Name	Optimization approach	Evaluation benchmark	Association criteria	Solution approach	Tracking method	Camera setting	Publish outlet
GMMCP [23]	Binary Mixed-Integer Linear Program	Duke MTMC and EPFL	Appearance and dynamic motion	Global	Offline	Overlapping and disjoined	arXiv
MYTRACKER [43]	Maximum weighted Independent Set (MWIS)	Duke MTMC and NLPR_MCT	Appearance, motion, interaction	Two-step	online	disjointed	*IET Image Processing,*
DyGLIP [30]	Link prediction	PETS09, CAMPUS, MCT, EPFL, and Cityflow	appearance	Two-step	online	Overlapping and disjoined	*IEEE/CVF CVPR*
TRACTA [29]	Optimal assignment problem	PETS09, CAMPUS, EPFL, MCT	Appearance, motion, camera topology, and location information	Two-step	online	Overlapping and disjoined	IEEE Transactions on Image Processing
B&P [180]	Multi-commodity network flow	PETS09	Appearance	Two-step	online	Overlapping	*IEEE CVPR*
LAAM [197]	Graph optimization	DukeMTMC and CityFlow	appearance	Two-step	offline	Overlapping and disjoined	arXiv
MHT-DAM [99]	Maximum weighted Independent Set (MWIS)	MOT challenge, PETS09	Appearance and motion	Two-step	online	Overlapping	*IEEE TPAMI*
KSP [28]	Integer programming problem	PETS09	Appearance and motion	Two-step	online	Overlapping	IEEE TPAMI
DP [183]	Dynamic programming	PETS09	Appearance and motion	Two-step	online	Overlapping	IEEE TPAMI
DEEPSORT + 2WGMMF + BS [192]	Probabilistic	CityFlow	Appearance and motion	Two-step	online	Disjointed	*IEEE/CVF CVPR*
TC + FVS + BA [192]	Weighted combination of cost	CityFlow	Appearance and motion	Two-step	offline	Disjointed	*IEEE/CVF CVPR*
DeepCC [63]	Correlation clustering	DukeMTMC and NLPR_MCT	Appearance and motion	Two-step	offline	Disjointed	*IEEE/CVF CVPR*
Ddashcam [242]	Non-linear	CityFlow	Appearance and motion	Two-step	offline	Disjointed	*IEEE/CVF CVPR*
UWIPL [121]	Greedy search	CityFlow	Appearance and motion	Two-step	offline	Disjointed	*IEEE/CVF CVPR*
TSCT + TA [148]	clustering	CityFlow	Appearance and motion	Two -step	offline	Overlapping and disjointed	*IEEE Transactions on Image Processing*
MCMT-TBM [241]	Non-linear	EPFL	Appearance and motion	global	online	Overlapping and disjointed	*Neurocomputing*
MOANA + FVS + BA [192]	probabilistic	CityFlow	Appearance and motion	Two-step	online	disjointed	*IEEE/CVF CVPR*
ELECTRICITY [193]	probabilistic	CityFlow	Appearance and motion	Two-step	online	disjointed	*IEEE/CVF CVPR*
TrafficBrain [194]	clustering	CityFlow	Appearance and motion	Two-step	online	disjointed	*IEEE/CVF CVPR*
MTMC_CDSC [132]	Quadratic program	DukeMTMC, MARS	Appearance and motion	Two-step	online	Disjointed	arXiv
BIPCC [164]	Binary integer program	DukeMTMC	Appearance and motion	Two-step	online	Disjointed	ECCV
SCLM + GAE [243]	Hierarchical Clustering	CityFlow	Appearance and topological	Two-step	online	Disjointed	*IEEE/CVF WACV*
TAREIDMTMC [31]	probabilistic	DukeMTMC-Re-ID	Appearance and motion	Two-step	online	Disjointed	*ACM international conference on Multimedia*
OPA [196]	Maximum weighted bipartite matching	DukeMTMC, EPFL, CAMPUS, MOT15, MOT16, and MOT17	Appearance and motion	Two-step	online	Overlapping and disjointed	*IEEE Transactions on Circuits and Systems for Video Technology*
mcmt [191]	–	CityFlow	Appearance	Two-step	online	Disjointed	*IEEE/CVF CVPR*
TRACTA + SAP [90]	Clustering	CityFlow	Appearance and motion	Two-step	online	Disjointed	*IEEE/CVF CVPR*
Fivefive [244]	–	CityFlow	Appearance	Two-step	online	Disjointed	*IEEE/CVF CVPR*
Matcher [241]	Greedy search	CityFlow	Appearance and motion	Two-step	online	Disjointed	*IEEE/CVF CVPR*
BOE [245]	Clustering	CityFlow	Appearance	Two-step	online	Disjointed	*IEEE/CVF CVPR*
FraunhoferiOSB [33]	Clustering	CityFlow	Appearance and motion	Two-step	online	Disjointed	*IEEE/CVF CVPR*

as possible to cover the trends and advances made over the years. Here, we collect 30 multi-camera multi-target tracking methods, that are features at reputable journals or conferences in recent years. The choice of trackers was restricted to the ones for which evaluation was performance with community recognized benchmark and/or code made publicly available. Furthermore, most trackers selected in this study are those with proven reputation, as evidently shown by the frequency of their citations. These state-of-the-art trackers are evaluated on one or more of the existing MCT benchmark datasets. The selected trackers include DyGLIP [30], UWIPL [121], MOANA + FVS + BA [192], TrafficBrain [194], GMMCP [23], MHT [5], TRACTA [24], B&P [114], LAAM [117], KSP [117], MHT-DAM [72], DP [97], OPA [196], DEEPSORT + 2WGMMF + BS [130], DeepCC [63], TSCT + TA [129], TC + FVS + BA [192], Ddashcam [120], ELECTRICITY [193], MTMC_CDSC [132], TAREIDMTMC [31], MYTRACKER [5], MCMT-TBM [144] etc. These trackers have all obtained the best performance on various MCMT benchmarks on which they are evaluated. In Table 4, we provide a comprehensive list of all surveyed trackers, and a detailed comparison of each. From analysis, it is revealed that seventeen out of the thirty trackers surveyed for comparison in this paper were originally evaluated by their designer on AI challenge CityFlow benchmark datasets. After that, DukeMTMC ranked as the second most employed benchmark among all the trackers, as eight of trackers adopted the datasets for evaluation.

The performance of these trackers is mostly ranked based on IDF1, IDP, IDR, and/or MOTA metrics. The recently proposed HOTA metric was only adopted as the performance measure in MCMT-TBM [241]. It turns out that most of the MCT methods follow the assumption that tracking within single camera is an already solved problem, with the majority adopting a ready-made single camera tracker. In summary, this section comprehensively summarizes the most advanced multicamera multi-object trackers. It examines their characteristics through six different perspectives. The analysis revealed that two step hierarchical approach, in which the overall performance of the tracker is determined by SCT outcome, is the most mainstream in MCT. Around 63.33%, 23.33%, and 13.33% of the selected trackers considered tracking in non-overlapping (disjointed), hybrid and overlapping setting respectively. Also, most of the selected trackers follow the online tracking approach representing 63.33% of the overall selected trackers, and the rest are based on offline method. The optimization problem resulting from MCT was formulated as different kinds of optimization problem, for which appropriate (nearly) optimal solution is sought. Among the adopted formulation and solution approaches followed include integer program, dynamic program, clustering, etc. Additionally, the strengths and weaknesses of each of the selected trackers are also shown in Table 5.

4.4. Benchmark results and analysis

In this subsection, we present the result obtained by each of the selected state-of-the-art multi-camera trackers when evaluated on at least one of the two dominant datasets (namely, DukeMTMCT and CityFlow). We begin by comparing the performance of each trackers on DukeMTMC dataset [249], as measured using the IDP metrics, and results as presented in Table 6 and Table 7. Starting with the reference system employed by authors of DukeMTMC [249], BIPCC [50,164] tracker. This tracker achieved a baseline multi-camera ID Precision (IDP) score of 67% and 59.6% on test-easy and test-hard sequences of the dataset, respectively. Similarly, MCMT_CDSC tracker [132] obtained multi-camera IDP score of 68.3% on test-easy sequence and 53.5% on test-hard sequence. The tracker also obtained an average single camera IDP score of 81.4% and 87.6% respectively on test-hard and test-easy sequences. For TAREIDMTMC tracker proposed in [31] and MYTRACKER suggested by authors in [43], multi-camera IDP scores of 71.8% and 71.1% were obtained on test-easy sequence, respectively. The multi-camera IDP score attained on test-hard sequence by these trackers are respectively, 63.2% and 68%. As displayed in Table 6, BIPCC and MCMT_CDSC ranked last on test-easy sequence based on multi-camera IDP scores among other trackers considered. Likewise, BIPCC, MCMT_CDSC, and MYTRACKER rated least on test-hard sequence, with MCMT_CDSC, and MYTRACKER attaining the same score of 63.2% and BIPCC scoring lowest (59.6%) as shown in Table 7. The superior multi-camera IDP score among the selected trackers on test-easy sequence of DukeMTMC dataset is obtained by OPA [196], while LAAM [197] obtained the highest IDP score on test-hard sequence.

The results obtained by selected trackers based on multi-camera ID Recall (IDR) metric are as shown in Table 8 and Table 9. The MTMC_CDSC [132] obtained 53.5% and 42.6% respectively on test-easy and test-hard sequences. The BIPCC [50] and DeepCC [63] trackers, both proposed by Ristani et al. in two independent studies, obtained multi-camera IDR scores of 48.4% and 78.9% on test-easy sequence, correspondingly. On test-hard sequence, these trackers achieved 39.2% and 62.4% multi-camera IDR score, respectively. Similarly, for MYTRACKER [43] and TAREIDMTMC [31], 60.6% and 66% multi-camera IDR scores were recorded on test-easy sequence, respectively. These trackers, however, realized on test-hard sequence, multi-camera IDR score of 43.9% and 55.5%. The best multi-camera IDR scores on test-easy is attained by OPA [196], realizing 86.6%, followed by LAAM [197] trackers that achieved 88.1%. Likewise, the top ranked tracker based on multi-camera IDR score on test-hard sequence is LAAM [197] trackers, reaching a value of 77.8%, followed by OPA [196], which achieved 71.4%.

For evaluation using IDF1 measure, the results obtained for all considered trackers are presented in Tables 10 and 11. The BIPCC [50,164] tracker achieved a baseline multi-camera IDF1 score of 59.77% and 42.92% on test-easy and test-hard sequences, correspondingly. Moreover, DeepCC [63] proposed by the same authors outperformed its baseline and state-of-the-art methods, reaching 82% and 68.5% multi-camera IDF1 score. The MTMC_CDSC proposed by Tesfaye et al. [250,251] achieved multi-camera IDF1 scores of 60% and 50.9% on test-easy and test-hard sequences of DukeMTMC dataset, respectively. Furthermore, The MTMCT approach proposed by the authors in [252] (TAREIDMTMC) also obtained multi-camera IDF1 scores of 68.8% and 61.2% on test-easy and test-hard sequences, respectively. The tracking approach proposed Yoon et al. (MYTRACKER) reached multi-camera IDF1 scores of 65.4% and 50.1% on test-easy and test-hard sequences, accordingly. The GMMCP [23], the only global MCT approach in this category reached the multi-camera IDF1 score of 55.5% on test-easy sequence. The LAAM tracker suggested by Hou et al. in [197] and OPA proposed by authors in [196] achieved as in apiece, multi-camera IDF1 score of 88.6% and 88% on test-easy sequence. While on test-hard sequence, they attained 82.3% and 76.2%, respectively. The best performance on DukeMTMC dataset in terms of multi-camera IDF1 measure was reported by Hou et al. in [197], followed by the approach suggested in [196]. The multi-camera IDF1 score reported in [50] and [132] ranked lowest among the trackers considered in this category.

In Table 12 and Table 13, we present the quantitative performance, specifically based on MOTA measure, of trackers evaluated on test-easy and test-hard sequences of DukeMTMC dataset. The reference system, BIPCC [50,164] employed by Ristani et al. to benchmark performance on DukeMTMC dataset in [249] achieved 59.4% and 54.6% average single-camera MOTA score on test-easy and test-hard sequences of the dataset, respectively. Likewise, DeepCC [63] also attained 47.3% and 28.21% improvement on average single-camera MOTA scores on test-easy and test-hard sequences, accordingly. The MTMC_CDSC proposed by Tesfaye et al. [250,251] achieved average single-camera MOTA scores of 70.9% and 59.6% on test-easy and test-hard sequences, respectively. Furthermore, The TAREIDMTMC approach proposed by the authors in [252] also achieved single-camera MOTA scores of 84.1% and 68% for test-easy and test-hard sequences, respectively. The tracking method proposed by Yoon et al., MYTRACKER realized average single-camera MOTA scores of 78.3% and 59.6% on test-easy and test-hard sequences, correspondingly. The best performance on

Table 5
Summary of strength and weaknesses of the selected trackers.

Reference	Tracker name	Year	Strength	Weaknesses	Target
Tang et al. [192]	TC + FVS + BA	2019	• Uses both appearance and motion features. • Capable of predicting the feature location of a target base on the previous location. • By adding the Batch All (BA) sampling variant, the tracker could process multiple video sequences in parallel, reducing the training time. • Follow the offline tracking approach.	• Requires high computation, which increases the complexity and the cost. • Difficulty in handling occlusion change in appearance in objects. • BA requires a large batch size to perform well.	Vehicles
Li et al. [120]	Ddashcam	• 2019	• The tracker utilizes temporal information besides the commonly employed spatial features for consistent across camera data association. • High accuracy with consistency is obtained as the tracker uses GPS information to complement the spatial–temporal features. • Obtained good accuracy and robustness for several issues in tracking when tested on multiple benchmarks.	• The overall tracking performance depends on accurate detection from every camera, which could be affected by occlusions and different lighting. • Motion of vehicle was assumed to be independent. • The author utilized a proposal-based detector which is generally slow and hinders the real-time speed of the tracker.	Vehicle
• Hsu et al. [121]	UWIPL	2019	• Exploit both temporal and appearance features for tracking and data association. • Utilized temporal attention model to extract highly discriminative feature for tracking. • Robust due to the trajectory-based CLM. • Could handle a larger number of cameras.	• High computational cost due to the deep learning approach. • Depends on accurate detection from every camera, which could be affected by occlusions and different lighting. • CLM requires manual labelling, which is not feasible for large-scale operations.	Vehicle
Tang et al. [192]	MOANA + FVS + BA	2019	• Provides robust appearance feature extraction and motion information. • strong approach to learning the relationship and interaction between objects. • efficient training and stability with deep learning techniques.	• Requires high computation, which increases the complexity and the cost. • Limitations in dealing with dynamic and too complex scenes. Less data training quality. • suspensible to overfitting or underfitting.	Vehicle
He et al. [194]	Traffic Brain	2019	• powerful visual vehicle features by using the clustering loss and trajectory consistency loss. • robust and consistent tracking across different cameras. • can deal with occlusion and change of camera views	• the method had limitation in dealing with images in the same trajectory with different quality for Re-ID as the image with poor quality does not represent the trajectory feature. • additional algorithms need to be utilized for quality judgement. • high computational cost.	Vehicle
Hou et al. [197]	LAAM	2019	• address the issue of inconsistency between the local tracklet association in tracking, and global appearance feature learning in re-ID.	• Depends on accurate detection from every camera, which could be affected by occlusions and different lighting.	Pedestrian and vehicle
Tang et al.	DEEPSORT + 2WGMMF + BS	2019	• Powerful combination to achieve robust and accurate tracking by combining appearance and motion cues. • Suitable for large number of detection and real-time tracking scenarios due to the fast and efficient data association using BS.	• The combination introduces intensive computation cost which could impact the tracking performance. • Sensitive to severe occlusion, fast motion and object complexity.	
You et al. [196]	OPA	2020	• robust to occlusion and effectively model local similarity using pose and optical cue. • accurate and suitable for fast motion target • uses visual information, spatial information, and pose information for MCT.	• the tracking performance depends on the outcome of the SCT, which could be affected by occlusion, poses and lighting. • assume an independent target motion model without considering interaction between objects. • the feature information is extracted using third-party model [63].	Pedestrian
Qian et al. [193]	ELECTRICITY	2020	• superior performance compared to other methods. • create an accurate bounding box using the weighted inter-class non-maximum suppression and association algorithm. • can deal with the variation in appearance due to different camera viewpoints by using aggregation loss for the training. • Easy to train	• Poor performance with low-quality data. • only proposed tracker considered motion and appearance cues during tracklet association. • assumed that the problem of SCT is solved and adopt off-the-shelf tracker [68] • due to the presence of post-processing steps, the tracker is slow.	Vehicle
He et al. [29]	TRACTA	2020	• The combination of semantic attribute and parsing and cross camera tracking improves the robustness of the MCMT. • Provides a better approach to deal with multiple tracklets of single object from multiple cameras	• The centralized tracking paradigm used is limited by the capacity and bandwidth of the cloud server. • It has Limited applicability in 3D as the footprint directly computed from 2D bounding box. • It assumes all persons stays on the ground with zero height.	Vehicle
You et al. [196]	OPA	2020	• approach is robust to occlusion and effectively model local similarity using pose and optical cue. • accurate and suitable for fast motion target.	• the tracking performance depends on the outcome of the SCT, which could be affected by occlusion, poses and lighting. • the authors assume an independent target motion model without considering interaction between objects.	Pedestrian

(*continued on next page*)

Table 5 (*continued*)

Reference	Tracker name	Year	Strength	Weaknesses	Target
			• uses visual information, spatial information, and pose information for MCT.	• the feature information is extracted using third-party model [62].	
Liu et al. [246]	mcmt	2021	• provides high accuracy tracking with the use of crossroad zones. • suitable for large-scale applications. • can deal with difficult occlusions scenarios.	• requires high-quality data, which is not always available in real-world applications. • complex method with high computation cost.	Vehicle
Hsu et al. [148]	TSCT + TA + META	2021	• Robust to occlusion and unreliable detection • viewpoint and illumination variation aware • used metadata information in addition to appearance and motion cue. • emphasized the importance of spatial and temporal cues in associating local tracklets.	• assumed solved the problem of SCT and adopted an off-the-self tracker (TNT [247]). • use of CLM requires manual labelling, which is not feasible for large-scale operations. • Besides, the CLM transition time distribution requires training which is impractical in real word applications	Vehicle
Ye et al. [244]	fivefive	2021	• Provides robust MCT approach based on Cascade R-CNN detector for the real-world vehicle tracking. • Provide an effective appearance feature extraction based on deep learning approach. • Introduce Tracklet-plane matching for effective inter-camera tracklets association.	• It uses a hard constraint for cross camera transition filtering which is not feasible for person tracking. • The use of prior assumption strategy limits the accuracy for unseen data. • The spatio-temporal constraint requires high precision and automation which is difficult to achieve	
Quach et al. [30]	DyGLIP	2021	• Has the ability to recover from lost track through camera transitions. • Provide better feature representation. • Provides high accuracy tracking in overlapping and non-overlapping scenarios	• The two-stage approach reduces the ability of global matching which depend on complicated graph structure. • Inability to hand a scenario of undetected object by one of the cameras. • Additional steps are needed to merge many local IDs [248]	Pedestrian and Vehicle
Hsu et al. [89]	SCLM + GAE	2022	• strong tracking in occluded circumstances and crowded areas due to the capturing ability of GAE • less reliable on manual annotation due to SCLM	• due to graph-optimization, it is computationally expensive and not suitable for real-time applications. • the GAE might depend on the quality of the graph and nodes' accuracy, which might not always be available in real scenario.	Vehicle
Specker et al. [33]	FraunhoferiOSB	2022	• improves SCT tracklets by introducing three additional modules. • Improved MCT performance with utilization of track refinement steps and hierarchical clustering. • Obtained competitive performance on CityFlowV2 benchmark dataset.	• Slow tracking speed as result of additional steps • the tracking performance depends on the outcome of the SCT, which could be affected by occlusion, poses and lighting. • The use of prior assumption strategy limits the accuracy for unseen data.	Vehicle
Li et al. [245]	BOE	2022	• Improved MCT performance due to the introduction of augmented track detection and multi-level data association. • Achieved promising result and impressive performance using the CityFlowV2 benchmark dataset.	• the tracking performance depends on the outcome of the SCT, which could be affected by occlusion, poses and lighting. • Motion model of the target is considered independent and failed to consider the interaction among targets.	Vehicle
Yang et al. [241]	Matcher	2022	• Proved better matching accuracy by including additional features using the reranking process. • Suitable for large scale MTMC in real world application due to the box-grain approach.	• Limited by prolonged occlusion and varying vehicle orientations • Use of constant velocity and linear motion model restrict the trackers capability. • Motion model of the target is considered independent and failed to consider the interaction among targets.	Vehicle

Table 6
IDP obtained by selected trackers on DukeMTMCT Test Easy Sequence.

Trackers	cam_1 (%)	cam_2 (%)	cam_3 (%)	cam_4 (%)	cam_5 (%)	cam_6 (%)	cam_7 (%)	cam_8 (%)	Avg (%)	Multi-cam (%)
MCMT_CDSC [132]	89.1	90.9	76.3	91.2	76.1	91.6	94	89.1	87.6	68.3
BIPCC [50]	91.2	69.3	78.9	88.7	83	87.5	93.6	92.2	83.6	67
DeepCC [63]	95.6	93.6	86.2	96.3	83.6	93.4	93.7	89.4	91.7	84.4
MYTRACKER [43]	89.7	88.9	76.2	84.1	81.4	88.9	91.4	90.8	87.6	71.1
OPA [196]	95.7	95.6	90.6	97.5	84.5	92.7	95	92	92.9	89.4
TAREIDMTMC [31]	77.5	77.3	63	93.6	71.2	81.8	80.2	79.5	77.7	71.8
LAAM [197]	–	–	–	–	–	–	–	–	91.1	87.4

DukeMTMC dataset in terms of the average single-camera MOTA metric on both test-easy and test-hard sequences was achieved in the work of Hou et al. [196], followed by the DeepCC tracker proposed by authors in [63].

The CityFlow 2019 (also known as CityFlowV1) benchmark dataset is one of the largest MTMC city-scale and spatial coverage datasets with 3 h of HD videos from 40 cameras in 10 intersections. This dataset was introduced by Tang et al. in [192], and comprises of videos with 960p resolution and frame rate of 10 fps. In the same article, the authors compared the performance of multiple state-of-the-art MTMCT on the dataset, thus serving as the baseline. For evaluation on the dataset, the authors combined FVS [119], TC [119] and various sampling variants to obtain several MCT trackers with competitive performance. Table 14 summarizes the performance of various selected trackers based on IDF1, IDP and IDR, when evaluated on the CityFlowV1. The baseline MCT approach, FVS + TC + BA suggested by [192] achieved the highest IDF1 score (46.3%), among other baseline approach used to benchmark the dataset. Ranking closely to FVS + TC + BA in the list of baseline

Table 7
IDP obtained by selected trackers on DukeMTMCT Test Hard Sequence.

Trackers	cam_1 (%)	cam_2 (%)	cam_3 (%)	cam_4 (%)	cam_5 (%)	cam_6 (%)	cam_7 (%)	cam_8 (%)	Avg (%)	Multi-cam (%)
MCMT_CDSC [132]	83	78.8	91.1	87.1	91.5	71.2	84.7	90.3	81.4	63.2
BIPCC [50]	92.5	65.7	96.1	86	90.1	81.7	81.2	94.9	81.2	59.6
DeepCC [63]	94.7	77.5	90.8	93	90.9	87	85.2	92.3	87.4	75.9
MYTRACKER [43]	72.2	61.2	86.9	84.4	93.3	70	74.7	93.5	81.4	63.2
OPA [196]	94.4	87.9	88.5	93	90.3	85.3	87.7	93.3	89.3	81.7
TAREIDMTMC [31]	75.2	65.3	80.2	98.1	81.9	71.6	74.9	82.6	74.4	68
LAAM [197]	–	–	–	–	–	–	–	–	93	89
GMMCP [23]	–	–	–	–	–	–	–	–	–	78.89

Table 8
IDR obtained by selected trackers on DukeMTMCT Test Easy Sequence.

	cam_1 (%)	cam_2 (%)	cam_3 (%)	cam_4 (%)	cam_5 (%)	cam_6 (%)	cam_7 (%)	cam_8 (%)	Avg (%)	Multi-cam (%)
MCMT_CDSC [132]	67.7	73.4	56	79	61.9	75.3	72.5	61	68.6	53.5
BIPCC [50]	41.8	67.1	48.8	62.8	65.4	69.1	70.6	59.6	60.4	48.4
DeepCC [63]	93	87.4	77.7	94.4	77.7	92.2	83.7	82.4	86.7	79.8
MYTRACKER [43]	79.6	75.9	63.5	77.6	67.3	78.8	73.5	71.3	74.4	60.6
OPA [196]	94.3	91.5	85.8	95.4	80.6	94.1	87.1	87.8	90	86.6
TAREIDMTMC [31]	75.7	73.7	58.2	90.6	67.2	79.2	73.1	72.5	73.8	66
LAAM [197]	–	–	–	–	–	–	–	–	92	88.1
GMMCP [23]	–	–	–	–	–	–	–	–	–	44.6

Table 9
IDR obtained by selected trackers on DukeMTMCT Test Hard Sequence.

	cam_1 (%)	cam_2 (%)	cam_3 (%)	cam_4 (%)	cam_5 (%)	cam_6 (%)	cam_7 (%)	cam_8 (%)	Avg (%)	Multi-cam (%)
MCMT_CDSC [132]	56.4	53.1	73.7	78.1	78.1	78.1	42.3	47.1	54.7	42.6
BIPCC [50]	36.8	56.1	46.5	82.7	75.1	52.7	47.1	72.8	53.5	39.2
DeepCC [63]	80.1	67.3	87.1	97	85.5	62.2	61.3	87.7	72	62.4
MYTRACKER [43]	58.4	52.6	74.1	82.2	79.2	44.9	47.1	44.4	70.5	43.9
OPA [196]	87	79.5	84.9	97.6	88.2	65.4	70.6	91.2	78	71.4
TAREIDMTMC [31]	65	58.9	77	96.4	79.5	53.5	56.4	77.2	63	55.5
LAAM [197]	–	–	–	–	–	–	–	–	81.1	77.8

Table 10
IDF1 obtained by selected trackers on DukeMTMCT Test Easy Sequence.

	cam_1 (%)	cam_2 (%)	cam_3 (%)	cam_4 (%)	cam_5 (%)	cam_6 (%)	cam_7 (%)	cam_8 (%)	Avg (%)	Multi-cam (%)
MCMT_CDSC [132]	76.9	81.2	64.6	84.7	68.3	82.7	81.8	73	77	60
BIPCC [50]	57.36	66.34	65.74	69.03	75.48	45.91	75.07	51.37	70.1	56.2
DeepCC [63]	94.3	90.4	81.8	95.3	80.6	92.8	88.5	85.8	89.2	82
MYTRACKER [43]	84.3	81.9	69.3	80.7	73.7	83.5	81.5	79.9	80.3	65.4
OPA [196]	91.4	95	93.7	88.1	96.5	82.5	93.4	90.9	89.9	88
TAREIDMTMC [31]	76.5	75.5	60.5	92.1	69.1	80.4	76.4	75.8	83.8	68.8
LAAM [197]	–	–	–	–	–	–	–	–	92.5	**88.6**
GMMCP [23]	–	–	–	–	–	–	–	–	–	55.5

Table 11
IDF1 obtained by selected trackers on DukeMTMCT Test Hard Sequence.

	cam_1 (%)	cam_2 (%)	cam_3 (%)	cam_4 (%)	cam_5 (%)	cam_6 (%)	cam_7 (%)	cam_8 (%)	Avg (%)	Multi-cam (%)
MCMT_CDSC [132]	67.1	63.4	81.5	82.3	82.8	53.1	60.6	81.3	65.4	50.9
BIPCC [50]	52.7	60.6	62.7	84.3	81.9	62.1	59.6	82.4	64.5	47.3
DeepCC [63]	86.8	72	88.9	94.9	88.1	72.5	71.3	89.9	79	68.5
MYTRACKER [43]	64.6	56.6	80	83.3	85.7	54.7	55.7	80.4	63.5	50.1
OPA [196]	90.6	83.5	86.6	95.2	89.2	74.1	78.2	92.3	83.2	76.2
TAREIDMTMC [31]	68.2	69.7	78.6	97.3	80.7	61.2	64.3	79.8	77.9	61.2
LAAM [197]	–	–	–	–	–	–	–	–	85.8	**82.3**

techniques suggested in [192] is the FVS + MOANA + BS reaching a IDF1 score of 46%. Similarly, the tracker obtained by merging DEEPSORT [69], the two-way Gaussian mixture model features (2WGMMF) [253], and batch sample (BS) attained the IDF1 score of 39.8%. However, the performance of this trackers in terms of IDP,and IDR were not reported by authors.

Ddashcam tracker suggested by authors in [242] obtained a IDF1 score of 68.65% on CityFlowV1 dataset, contributing about 48.3% improvement on the state-of-the-art. Similarly, UWIPL tracker suggested by Hsu et al. [254] and TrafficBrain tracker [255] achieved a IDF1 scores of 70.59% and 66.53%, respectively. An IDF1 score of 76.77% was obtained by TSCT-TA-META approach proposed Hsu et al. in [148]. The

Table 12
MOTA obtained by selected trackers on DukeMTMCT Test Easy Sequence.

	cam_1 (%)	cam_2 (%)	cam_3 (%)	cam_4 (%)	cam_5 (%)	cam_6 (%)	cam_7 (%)	cam_8 (%)	Avg (%)
MCMT_CDSC [132]	69.9	71.5	67.4	76.8	68.9	77	73.8	63.4	70.9
BIPCC [50]	43	44.8	57.8	63.2	72.8	73.4	71.4	60.7	59.4
DeepCC [63]	93.3	87.1	79.7	91.8	86.2	88.7	82.2	85	87.5
MYTRACKER [43]	84.9	78.4	65.7	79.8	76.6	82.8	77	71.6	78.3
OPA [196]	95.6	90.4	85.6	93.5	89.8	90	85.4	88.3	90.4
TAREIDMTMC [31]	87.7	81.1	76.6	89.8	83.7	86.9	80.5	83.2	84.1

Table 13
MOTA obtained by selected trackers on DukeMTMCT Test Hard Sequence.

	cam_1 (%)	cam_2 (%)	cam_3 (%)	cam_4 (%)	cam_5 (%)	cam_6 (%)	cam_7 (%)	cam_8 (%)	Avg (%)
MCMT_CDSC [132]	63.2	54.8	68.8	75.6	78.6	53.3	50.8	70	59.6
BIPCC [50]	37.8	47.3	46.7	85.3	78.3	59.4	50.8	73	54.6
DeepCC [63]	79.6	57.9	84.2	90.3	86	63.3	61.4	85	70
MYTRACKER [43]	61.1	50.4	70.3	81.2	81.9	56.1	49.8	71.5	59.6
OPA [196]	87.1	84.3	78.9	90.2	91.4	88.7	69.4	89.2	78.8
TAREIDMTMC [31]	74.2	51.5	88.9	95.4	84.9	62.9	62.6	85.3	68

Table 14
Performance comparison of selected trackers evaluated on CityFlowV1.

Authors	year	Tracking method	IDF1 (%)	IDP (%)	IDR (%)
Tang et al. [192]	2019	FVS + TC + BA	46.3	–	–
		FVS + MOANA + BS	46	–	–
		DEEPSORT + 2WGMMF + BS	39.8	–	–
Li et al. [242]	2019	Ddashcam	68.65	–	–
Hsu et al. [254]	2019	UWIPL	70.59	69.12	72.11
Hsu et al. [148]	2021	TSCT + TA + META	76.77	82.76	71.59
He et al. [255]	2019	Traffic Brain	66.53	–	–
Hsu et al. [89]	2022	CLM + GAE	77.21	82.43	72.62
		SCLM + GAE	75.47	81.30	70.04

CLM-GAE and SCLM-GAE trackers proposed by authors in [256], achieved IDF1 scores of 77.21% and 75.47%, accordingly. The best performance in term of IDF1 is obtained by CLM + GAE tracker, followed by TSCT + TA + META, outperforming the benchmark with 66.8% and 65.8% performance improvement, respectively.

Beginning from 2020, some improvements were made to CityFlow datasets [192], and was formalized as CityFlowV2 datasets in 2021 for AI city challenge. These improvements include an increase in coverage brought about the addition of six extra cameras to make forty-six cameras in total. At the same time, the number of unique vehicle identities was also increased to 880. Over the years, several trackers have been evaluated on this new dataset, and the results obtained by selected ones for this study, are as presented in Table 15. For instance, ELECTRICITY proposed by authors in [256] achieved an IDF1 score of 45.85% on the dataset, while TRACTA [29] obtained a score of 44%. Similarly, mcmt [246], DyGLIP [30], and fivefive [244] realized IDF1 scores of 80.95%, 64.9% and 77.87%, respectively. Recent trackers such as SCLM + GAE [89], Matcher [241], FraunhoferiOSB [33], and BOE [245] also achieved IDF1 scores of 55.56%, 84.86%, 83.48%, and 84.37%, correspondingly. The best performance in terms of IDF1 score on CityFlowV2 is attained by Matcher proposed by authors in [241]. Next to this tracker in terms of performance is the BOE tracker proposed by Li et al. in [245]. Moreover, optimal performance in terms of IDR metric is obtained by FraunhoferiOSB [33], while Matcher also achieved a superior performance with respect to IDP.

Table 15
Performance comparison of selected trackers evaluated on CityFlowV2.

Authors	Year	Tracking method	IDF1 (%)	IDP (%)	IDR (%)
Qian et al. [256]	2020	ELECTRICITY	45.85	–	–
He et al. [29]	2020	TRACTA	44	53.63	32.80
Liu et al. [246]	2021	mcmt	80.95	85.69	76.70
Quach et al. [30]	2021	DyGLIP	64.9	–	–
Ye et al. [244]	2021	fivefive	77.87	–	–
Hsu et al. [89]	2022	SCLM + GAE	55.56	–	–
Yang et al. [241]	2022	Matcher	84.86	91.37	79.21
Specker et al. [33]	2022	FraunhoferiOSB	83.48	86.74	80.46
Li et al. [245]	2022	BOE	84.37	89.0	80.2

To sum up, among the eight selected trackers evaluated on DukeMTMC dataset, OPA [196] and LAAM [197] obtained the highest performance in terms of multi-camera IDP, IDR and IDF1 scores. The performance reported in terms of average single-camera MOTA is also topmost for OPA [196], followed by the DeepCC [63]. The obvious performance improvement observed in OPA, thus outperforming state-of-the-art largely, is primary due to its carefully designed inter-camera data association approach and the inclusion of optical pose method. Similarly, Matcher [241] obtained the topmost performance on the CityFlow dataset [192], as can be seen in Table 15. The tracker obtained an IDF score of 84.86% and top the public leaderboard of the 2022 AI city challenge's track 1. With an IDF score of 83.48%, the FraunhoferiOSB [147] tracker ranked second among the surveyed trackers and was also the second runner-up in the 2022 AI city challenge's track 1. Other top-ranked trackers include mcmt [136] and fivefive [151] trackers with 80.95% and 77.87% IDF1 score respectively.

5. Existing issues and future directions

This paper provides an in-depth examination of methods and challenges associated with the task of Multiple Object Tracking (MOT) in videos captured by multiple camera systems. As the most recent and comprehensive literature review in the past decade, it presents a clear problem formulation and categorization of existing methods, highlights key components of state-of-the-art approaches, and discusses evaluations using various metrics, publicly available datasets, open-source implementations, and benchmark results. Despite significant progress in the field, ongoing challenges and open problems in MCMOT research continue to be an area of active investigation. In the following, we discuss open issues and future research directions.

5.1. Existing issues

Tracking targets using multiple cameras is a novel and growing field of study in comparison to traditional single-camera tracking methods. To advance the research, new datasets and metrics are being developed specifically for multi-camera scenarios. These datasets can include scenarios where multiple cameras are observing the same scene with complete overlap, or where cameras have minimal overlap. However, acquiring these datasets can be difficult as they often raise privacy concerns. Additionally, the current evaluation metrics used to assess the performance of multi-camera tracking algorithms are similar to those used for single-camera tracking, and primarily focus on overall performance rather than the specific elements contributing to the results. Furthermore, the design of most of the current multi-camera tracking systems primarily focuses on creating models that can accurately identify and track targets, while distinguishing them from the background. These models are used to detect, track and re-identify the targets in multiple cameras.

However, in scenarios involving multiple cameras, effectively associating the observations of the same target across cameras is crucial in determining the target's overall trajectory. This is a key challenge in Multi-Camera Multi- Target (MCMT) tracking as the same target may be observed by multiple cameras, and the tracking algorithm must be able to match the observations to determine the target's overall trajectory. Despite this, most existing tracking algorithms focus on single camera data association which may not perform well when dealing with data from multiple cameras [6]. Recent research has shown that MCMT approaches based on global graph methods can improve data association across cameras and perform better with less reliance on object detection models. These methods use a global graph where each camera and target are represented as nodes, and the edges are the likelihood of the target being observed by a camera, allowing for more efficient data association. However, research studies using this approach are still limited in comparison to other methods.

Another major challenges in MCMT are target mismatch. This happens when two objects come too close together and the tracker incorrectly swaps them. Object detection, within-camera tracking, and cross-camera re-identification are all key research areas that play a crucial role in developing better modules for Multi-Camera Multi-Target (MCMT) tracking methods. However, this research is often detached from the actual requirements of a functioning system. In other words, the research is not always aligned with the real-world scenarios that the system is intended to operate in. For example, re-identification benchmarks are often formulated as closed-world issues between two isolated cameras, while in real-world situations, the circumstances are always open-set. This means that the benchmarks are designed to test the system's ability to match targets between two isolated cameras, while in the real-world, the system may have to match targets across multiple cameras, under different lighting conditions, and with different backgrounds. Similarly, tracking benchmarks are often described as short-term problems with little target appearance variation considered or extended periods without any camera observation. In real-world situations, targets of interest may spend prolonged time in locations, with appearance variations, and visible to several cameras, which makes the tracking problem more complex. These variations in appearance and prolonged periods of observation are not always considered in the benchmarks, which can lead to a lack of robustness in the developed tracking methods.

The problem of Multi-Camera Tracking (MTMCT) remains a challenging task and has constantly received significant attention in the field of computer vision and video surveillance. Despite years of efforts, MTMCT remains a largely unsolved problem due to several reasons, which can be summarized as follows:

1. Background clutter and object occlusions cause erroneous or incomplete local tracking results under a single camera. This means that objects in the background or other objects that might be in front of the target can cause confusion for the tracking algorithm, leading to incorrect or incomplete tracking results.
2. Dramatic variations in visual appearance and ambient environment caused by different viewpoints from different cameras make cross-camera local tracklet matching extremely difficult. This means that even if the same target is being tracked by multiple cameras, the different viewpoints and lighting conditions can make it difficult to match the tracklets across cameras.
3. The number of cameras in which each target appears, and the number of targets within the entire multi-camera network are both unknown, making the inference of the global trajectory of each target even more challenging. This means that the tracking algorithm does not have information about how many cameras are being used or how many targets there are, making it difficult to infer the overall trajectory of each target.
4. Currently, techniques such as within-camera tracking and multi-camera optimization are employed in multi-camera multi-target tracking to determine consistent human identification over time and space. While this approach is crucial, it may encounter challenges due to variations in appearance and delayed disappearance, which can make it difficult to seamlessly transfer tracking from one camera to another.
5. The problem becomes even more challenging when a tracker mistakenly generates plural local tracklets for the same target within a single camera. This means that if the tracking algorithm generates multiple tracklets for the same target within a single camera, it becomes even more difficult to infer the global trajectory of that target.
6. Fast camera motion is still a challenge in multicamera tracking because it can make it difficult to accurately track objects across different camera views. When the camera is moving quickly, the objects in the scene may appear to move in different directions or at different speeds than they actually are, leading to problems with object detection and tracking.
7. There is only scant research on the development of some important MCMT sub-module for some applications such as city-scale vehicle tracking in intelligent transport system. For instance, research on vehicle Re-ID is very limited, and thus a significant issue. Building a robust vehicle Re-ID model for the entire process of Multi-Camera Multi-Target Tracking (MTMC) is crucial. Furthermore, GPS information is a significant complement to the MTMC results, and the challenge of how to effectively combine video information with GPS location remains unsolved.

In summary, Object detection, within-camera tracking, and cross-camera re-identification are important research areas that help to improve the performance of MCMT tracking methods. However, the research is often not aligned with the real-world requirements of a functioning system which can lead to a lack of robustness in the developed tracking methods.

5.2. Future research direction

Despite the progress that has been made in MCMT research, the findings of the survey noted that there are still many open problems to be studied and ongoing challenges to be addressed. The following are the identified future research directions:

- **Development of more advanced data association algorithms**: Developing more advanced algorithms for data association between multiple cameras and across frames can improve the performance of multi-camera tracking systems by reducing errors and increasing the accuracy of the tracking results.
- **Improved robustness and accuracy against challenging environmental factors**: Various factors such as low light conditions, occlusions, cluttered scenes etc., have a significant negative

influence on the performance of multi-camera tracking systems. Intensifying research that aims to improve the robustness and accuracy of multi-camera tracking systems to these factors will ensure their reliable and consistent performance.

- **Development of unifying evaluation metrics for MCT:** Developing a specific and unifying evaluation metric for multi-camera tracking is necessary to standardize and effectively assess the performance of tracking algorithms across different camera setups, enabling fair comparisons and advancements in the field.
- **Incorporation of deep learning techniques**: Deep learning techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), diffusion models, and vision transformer models, have been shown to be effective in a wide range of computer vision tasks. Incorporating these techniques into multi-camera tracking systems can improve their performance by enabling them to learn more robust and accurate representations of the objects being tracked.
- **Exploring the use of multi-camera tracking in a variety of applications**: multi-camera tracking systems have a wide range of potential applications, such as surveillance, autonomous vehicles, and robotics. Research in this direction aims to explore the use of multi-camera tracking in different applications, to identify new use cases and to develop methods tailored to the specific requirements of these applications.
- **Investigating the use of multiple cameras with different modalities**: multi-camera tracking systems can be enhanced by using cameras with different modalities, such as thermal and visible light cameras. Research in this direction aims to investigate how the use of multiple cameras with different modalities can improve the performance of multi-camera tracking systems.
- **Developing methods for real-time multi-camera tracking**: In time-sensitive applications, it is important to have a multi-camera tracking system that can process data in real-time. Research in this direction aims to develop methods that enable multi-camera tracking systems to operate in real-time, to meet the requirements of such applications.
- **Investigate the possibility of using unsupervised or semi-supervised learning approaches**: multi-camera tracking systems typically rely on supervised learning methods, which require large amounts of labelled data for training. Research in this direction aims to investigate the possibility of using unsupervised or semi-supervised learning methods, which do not require labelled data, for multi-camera tracking.
- **Fast camera motion adaptiveness:** To tackle the problem of camera motion, researchers have proposed different approaches such as using motion models, motion compensation, or implementing visual odometry to estimate the motion of the camera. However, these methods are not always effective, and more research is needed to improve the performance of multicamera tracking systems under fast camera motion. Investigative study into effective 2D tracking conversion to 3D tracking via odometry is prospective research to address this problem.
- **Developing methods for privacy-preserving multi-camera tracking:** multi-camera tracking systems can collect and process sensitive personal information, which raises privacy concerns. Research in this direction aims to develop methods that can preserve the privacy of individuals while still allowing multi-camera tracking systems to operate effectively. This can be achieved by using privacy-preserving techniques such as homomorphic encryption and differential privacy.
- **Exploring Zero-shot learning in MCT**: In MCT, zero-shot learning can be used to recognize and track objects that were not present in the training data. For example, a zero-shot learning model can be trained to recognize a new type of vehicle, such as a hovercraft, based on its visual features and semantic descriptions. Once the model has learned to recognize the new object, it can be incorporated into the tracking system and used to track the object in real-time.
- **Integrating language-vision models into MCT:** Language-vision models can be used to improve the tracking accuracy by incorporating textual information such as object descriptions, attributes, and relationships. For example, a language-vision model can be trained to understand the relationship between objects and their attributes, such as color, size, and shape. This information can be used to refine the tracking process by incorporating the textual information into the visual tracking algorithm.

6. Conclusion

Tracking of target in a wide area involves the use of multiple cameras, and tracking problem under such configuration is regarded as multi-camera tracking (MCT). Multi-target multi-camera tracking (MTMCT) is a challenging problem in computer vision that has gained increasing attention in recent years. This review article provides a comprehensive overview of the current trends and advances in MCT from six different perspectives, including problem formulation, problem-solving approach, data association criteria, mutual exclusion constraints, benchmark dataset, and performance metrics. This study found that solution methods based on two-step hierarchical techniques are the mainstream technique over the years; with issue of utter dependence of tracking performance on the detection response which is now been addressed by the adoption of global MCT solution technologies. The examination of the tracklet association criteria revealed that most of the existing works assumed an independent motion model in addition to the appearance model with limited consideration of the dynamic target interaction. The study also emphasized the various benefits of mutual exclusion criteria in effective data associations, and their valuable roles in solving target occlusion issue in MCT.

A critical review of various existing works revealed that various benchmark datasets and evaluation metrics have been employed over the years to measure the performance of MCT. Due to its comprehensiveness in terms of the number of unique identity and annotated bounding boxes, DukeMTMC dataset is identified as the most dominant among all MCT datasets. However, owing to it retraction in 2019 due to privacy concern, other datasets such as PETS2009, EPFL, NLPR MCT etc. are gaining attraction. Additionally, this review also identified the need for unifying MCT-specific metrics for effective performance evaluation. To get a good grip of recent trends and advances in multi-camera tracking research, an experimental comparison of 30 MCT models was also conducted. The strengths and weaknesses of these approaches were also highlighted to provide direction for future MCT research. Among all selected trackers for experimental comparison, OPA [196] and LAAM [197] achieved the topmost performance on DukeMTMC dataset in terms of three evaluation metrics. Similarly, Matcher [241] obtained the best performance among all trackers evaluated on CityFlow benchmark dataset, while also toping the public leaderboard of the 2022 AI city challenge's track 1. Finally, the paper identified the need to develop an effective data association approach, contribute privacy-preserving benchmark datasets, as well as developing unifying performance metrics for MCT algorithms evaluation, among others as possible future research in MCT. The review emphasizes the potential of deep learning models, zero-shot learning, and language-vision models in addressing MCT challenges. Overall, this review article provides a valuable resource for researchers and practitioners working on MCT, and it provides insights into the current state-of-the-art and future research directions in this field.

CRediT authorship contribution statement

Temitope Ibrahim Amosa: Conceptualization, Methodology, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Patrick Sebastian:** Conceptualization, Supervision, Project

administration, Funding acquisition, Writing – review & editing. **Lila Iznita Izhar:** Funding acquisition, Resources, Validation, Writing – review & editing. **Oladimeji Ibrahim:** Methodology, Data curation, Writing – review & editing. **Lukman Shehu Ayinla:** Methodology, Writing – review & editing. **Bahashwan Abdulrahman Abdullah Omar:** Methodology, Writing – review & editing. **Abubakar Bala:** Methodology, Writing – review & editing. **Yau Alhaji Samaila:** Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgement

This study was fully supported by the Yayasa UTP Fundamental Research Grant (YUTP-FRG) with cost center 015LC0-441.

References

[1] M. Kristan *et al.*, The ninth visual object tracking vot2021 challenge results, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2711-2738.
[2] A.W. Smeulders, D.M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, M. Shah, Visual tracking: An experimental survey, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2013) 1442–1468.
[3] T. Wang, H. Ling, Gracker: A graph-based planar object tracker, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2017) 1494–1501.
[4] H. Yang, L. Shao, F. Zheng, L. Wang, Z. Song, Recent advances and trends in visual tracking: A review, Neurocomputing 74 (18) (2011) 3823–3831, https://doi.org/10.1016/j.neucom.2011.07.024.
[5] X. Wang, Intelligent multi-camera video surveillance: A review, Pattern Recogn. Lett. 34 (1) (2013) 3–19.
[6] X. Chen, K. Huang, T. Tan, Object tracking across non-overlapping views by learning inter-camera transfer models, Pattern Recogn. 47 (3) (2014) 1126–1137.
[7] H. Aghajan, A. Cavallaro, Multi-Camera Networks: Principles and Applications, Academic Press, 2009.
[8] S. Zhang, Y. Zhu, A. Roy-Chowdhury, Tracking multiple interacting targets in a camera network, Comput. Vis. Image Underst. 134 (2015) 64–73.
[9] T. I. Amosa, P. Sebastian, L. I. B. Izhar, and O. Ibrahim, A Comparative Study on Deep Feature Extraction Approaches for Visual Tracking of Industrial Robots, in *2022 IEEE Industrial Electronics and Applications Conference (IEACon)*, 2022: IEEE, pp. 122-127.
[10] B. Song, A.K. Roy-Chowdhury, Robust tracking in a camera network: a multi-objective optimization framework, IEEE J. Sel. Top. Signal Process. 2 (4) (2008) 582–596.
[11] A. Sadeghian, A. Alahi, S. Savarese, Tracking the untrackable: Learning to track multiple cues with long-term dependencies, in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 300-311.
[12] Y. Huang, H. Bi, Z. Li, T. Mao, Z. Wang, Stgat: Modeling spatial-temporal interactions for human trajectory prediction, in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6272-6281.
[13] L. Leal-Taixé, G. Pons-Moll, B. Rosenhahn, Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker, in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, 2011: IEEE, pp. 120-127.
[14] G. Antonini, S.V. Martinez, M. Bierlaire, J.P. Thiran, Behavioral priors for detection and tracking of pedestrians in video sequences, Int. J. Comput. Vis. 69 (2) (2006) 159–180.
[15] S. Pellegrini, A. Ess, K. Schindler, L. Van Gool, You'll never walk alone: Modeling social behavior for multi-target tracking, in *2009 IEEE 12th international conference on computer vision*, 2009: IEEE, pp. 261-268.
[16] A. Johansson, D. Helbing, Analysis of empirical trajectory data of pedestrians, in: Pedestrian and Evacuation Dynamics, Springer, 2008, pp. 203–214.
[17] P. Scovanner, M. F. Tappen, Learning pedestrian dynamics from the real world, in *2009 IEEE 12th International Conference on Computer Vision*, 2009: IEEE, pp. 381–388.
[18] W. Choi, S. Savarese, A unified framework for multi-target tracking and collective activity recognition, in: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part IV 12, Springer, 2012, pp. 215–230.
[19] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, Trackformer: Multi-object tracking with transformers, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8844-8854.
[20] Y.-J. Cho, S.-A. Kim, J.-H. Park, K. Lee, K.-J. Yoon, Joint person re-identification and camera network topology inference in multiple cameras, Comput. Vis. Image Underst. 180 (2019) 34–46.
[21] O. Javed, K. Shafique, Z. Rasheed, M. Shah, Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views, Comput. Vis. Image Underst. 109 (2) (2008) 146–162.
[22] M.V. Prasad, R. Balakrishnan, Spatio-temporal association rule based deep annotation-free clustering (STAR-DAC) for unsupervised person re-identification, Pattern Recogn. 122 (2022), 108287.
[23] W. Liu, O. Camps, M. Sznaier, Multi-camera multi-object tracking, *arXiv preprint arXiv:1709.07065,* 2017.
[24] M. Chandrajit, R. Girisha, T. Vasudev, Multiple objects tracking in surveillance video using color and hu moments, Signal & Image Process.: Int. J. 7 (3) (2016) 15–27.
[25] A. Yilmaz, O. Javed, M. Shah, Object tracking: a survey, ACM Comput. Surv. 38 (4) (2006), https://doi.org/10.1145/1177352.1177355, pp. 13–es.
[26] T. D. Orazio, P. L. Mazzeo, P. Spagnolo, Color Brightness Transfer Function evaluation for non overlapping multi camera tracking, in *2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 30 Aug.-2 Sept. 2009 2009, pp. 1-6, doi: 10.1109/ICDSC.2009.5289365.
[27] B. Long, Y. Chang, Relevance Ranking for Vertical Search Engines, Newnes, 2014.
[28] J. Berclaz, F. Fleuret, E. Turetken, P. Fua, Multiple object tracking using k-shortest paths optimization, IEEE Trans. Pattern Anal. Mach. Intell. 33 (9) (2011) 1806–1819.
[29] Y. He, X. Wei, X. Hong, W. Shi, Y. Gong, Multi-target multi-camera tracking by tracklet-to-target assignment, IEEE Trans. Image Process. 29 (2020) 5191–5205.
[30] K. G. Quach *et al.*, DyGLIP: A Dynamic Graph Model with Link Prediction for Accurate Multi-Camera Multiple Object Tracking, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20-25 June 2021 2021, pp. 13779-13788, doi: 10.1109/CVPR46437.2021.01357.
[31] N. Jiang, S. Bai, Y. Xu, C. Xing, Z. Zhou, and W. Wu, Online inter-camera trajectory association exploiting person re-identification and camera topology, in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1457-1465.
[32] A. Specker, D. Stadler, L. Florin, J. Beyerer, An Occlusion-aware Multi-target Multi-camera Tracking System, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 19-25 June 2021 2021, pp. 4168-4177, doi: 10.1109/CVPRW53098.2021.00471.
[33] A. Specker, L. Florin, M. Cormier, and J. Beyerer, Improving Multi-Target Multi-Camera Tracking by Track Refinement and Completion, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 19-20 June 2022 2022, pp. 3198-3208, doi: 10.1109/CVPRW56347.2022.00361.
[34] M. Bredereck, X. Jiang, M. Körner, and J. Denzler, Data association for multi-object Tracking-by-Detection in multi-camera networks, in *2012 Sixth International Conference on Distributed Smart Cameras (ICDSC)*, 30 Oct.-2 Nov. 2012 2012, pp. 1-6.
[35] R. Iguernaissi, D. Merad, K. Aziz, P. Drap, People tracking in multi-camera systems: a review, Multimedia Tools Appl. 78 (8) (2019) 10773–10793, https://doi.org/10.1007/s11042-018-6638-5.
[36] A.S. Olagoke, H. Ibrahim, S.S. Teoh, Literature survey on multi-camera system and its application, IEEE Access 8 (2020) 172892–172922, https://doi.org/10.1109/ACCESS.2020.3024568.
[37] P. Li, D. Wang, L. Wang, H. Lu, Deep visual tracking: review and experimental comparison, Pattern Recogn. 76 (2018) 323–338, https://doi.org/10.1016/j.patcog.2017.11.007.
[38] S.M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, S. Kasaei, Deep learning for visual tracking: a comprehensive survey, IEEE Trans. Intell. Transp. Syst. 23 (5) (2022) 3943–3968, https://doi.org/10.1109/TITS.2020.3046478.
[39] S. Javed, M. Danelljan, F.S. Khan, M.H. Khan, M. Felsberg, J. Matas, Visual object tracking with discriminative filters and siamese networks: a survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. 45 (5) (2023) 6552–6574, https://doi.org/10.1109/TPAMI.2022.3212594.
[40] Y. Dai, Z. Hu, S. Zhang, L. Liu, A survey of detection-based video multi-object tracking, Displays 75 (2022), 102317, https://doi.org/10.1016/j.displa.2022.102317.
[41] J. Fu, C. Xu, A survey of single object tracking methods, Nanjing Xinxi Gongcheng Daxue Xuebao 11 (6) (2019) 638–650.
[42] A. Gad, T. Basmaji, M. Yaghi, H. Alheeh, M. Alkhedher, M. Ghazal, Multiple object tracking in robotic applications: trends and challenges, Appl. Sci. 12 (19) (2022) 9408.
[43] K. Yoon, Y.M. Song, M. Jeon, Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views, IET Image Proc. 12 (7) (2018) 1175–1184.
[44] A. Dehghan, Y. Tian, P. H. Torr, M. Shah, Target identity-aware network flow for online multiple target tracking, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1146-1154.
[45] S. Tang, M. Andriluka, B. Andres, B. Schiele, Multiple people tracking by lifted multicut and person re-identification, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3539-3548.
[46] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, M.-H. Yang, Online multi-object tracking with dual matching attention networks, in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 366-382.

[47] P. Chu and H. Ling, Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6172-6181.

[48] S.-H. Bae, K.-J. Yoon, Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking, IEEE Trans. Pattern Anal. Mach. Intell. 40 (3) (2017) 595–610.

[49] Y. Zhang, C. Wang, X. Wang, W. Zeng, W. Liu, Fairmot: On the fairness of detection and re-identification in multiple object tracking, Int. J. Comput. Vis. 129 (2021) 3069–3087.

[50] E. Ristani, C. Tomasi, Tracking multiple people online and in real time, in: Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, NoVember 1–5, 2014, ReVised Selected Papers, Part V 12, Springer, 2015, pp. 444–459.

[51] A. Milan, S. Roth, K. Schindler, Continuous energy minimization for multitarget tracking, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2013) 58–72.

[52] A. Andriyenko, K. Schindler, Globally optimal multi-target tracking on a hexagonal lattice, in: Computer VIsIon–ECCV 2010: 11th European Conference on Computer VIsIon, HeraklIon, Crete, Greece, September 5–11, 2010, ProceedIngs, Part I 11, Springer, 2010, pp. 466–479.

[53] W. Chen, L. Cao, X. Chen, K. Huang, An equalized global graph model-based approach for multicamera object tracking, IEEE Trans. Circuits Syst. Video Technol. 27 (11) (2016) 2367–2381.

[54] A. Dehghan, S. Modiri Assari, M. Shah, GMMCP Tracker: Globally Optimal Generalized Maximum Multi Clique Problem for Multiple Object Tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4091-4099.

[55] S. Khan, M. Shah, Consistent labeling of tracked objects in multiple cameras with overlapping fields of view, IEEE Trans. Pattern Anal. Mach. Intell. 25 (10) (2003) 1355–1360.

[56] R. Cucchiara, Multimedia surveillance systems, in *Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, 2005, pp. 3-10.

[57] R. T. Collins, A. J. Lipton, H. Fujiyoshi, T. Kanade, Algorithms for cooperative multisensor surveillance, *Proceedings of the IEEE,* vol. 89, no. 10, pp. 1456-1477, 2001.

[58] Q. Cai, J.K. Aggarwal, Tracking human motion in structured environments using a distributed-camera system, IEEE Trans. Pattern Anal. Mach. Intell. 21 (11) (1999) 1241–1247.

[59] T.-H. Chang, S. Gong, Tracking multiple people with a multi-camera system, in *Proceedings 2001 ieee workshop on multi-object tracking*, 2001: IEEE, pp. 19–26.

[60] A. Mittal, L.S. Davis, M 2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene, Int. J. Comput. Vis. 51 (2003) 189–203.

[61] L. Lee, R. Romano, G. Stein, Monitoring activities from multiple video streams: Establishing a common coordinate frame, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 758–767.

[62] A.E. Benrazek, B. Farou, H. Seridi, Z. Kouahla, M. Kurulay, Ascending hierarchical classification for camera clustering based on FoV overlaps for WMSN, IET Wireless Sens. Syst. 9 (6) (2019) 382–388.

[63] E. Ristani, C. Tomasi, Features for multi-target multi-camera tracking and re-identification, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6036-6046.

[64] K.-W. Chen, C.-C. Lai, P.-J. Lee, C.-S. Chen, Y.-P. Hung, Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras, IEEE Trans. Multimedia 13 (4) (2011) 625–638.

[65] X. Chen, L. An, B. Bhanu, Multitarget tracking in nonoverlapping cameras using a reference set, IEEE Sens. J. 15 (5) (2015) 2692–2704.

[66] Z. Zhang, J. Wu, X. Zhang, C. Zhang, Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project, *arXiv preprint arXiv:1712.09531,* 2017.

[67] X. Chen, B. Bhanu, Integrating social grouping for multitarget tracking across cameras in a crf model, IEEE Trans. Circuits Syst. Video Technol. 27 (11) (2016) 2382–2394.

[68] M. Hofmann, D. Wolf, G. Rigoll, Hypergraphs for joint multi-view reconstruction and multi-object tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3650-3657.

[69] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, in *2017 IEEE international conference on image processing (ICIP)*, 2017: IEEE, pp. 3645-3649.

[70] Z. Wang, L. Zheng, Y. Liu, Y. Li, S. Wang, Towards real-time multi-object tracking, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, Springer, 2020, pp. 107–122.

[71] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft, Simple online and realtime tracking, in *2016 IEEE international conference on image processing (ICIP)*, 2016: IEEE, pp. 3464-3468.

[72] Y. Zhang, et al., Bytetrack: Multi-object tracking by associating every detection box, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII, Springer, 2022, pp. 1–21.

[73] P. Sun *et al.*, Transtrack: Multiple object tracking with transformer, *arXiv preprint arXiv:2012.15460,* 2020.

[74] Q. Wang, Y. Zheng, P. Pan, Y. Xu, Multiple object tracking with correlation learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3876-3886.

[75] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Proces. Syst. 28 (2015).

[76] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2009) 1627–1645.

[77] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1532–1545.

[78] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Computer VIsIon–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, ProceedIngs, Part I 16, Springer, 2020, pp. 213–229.

[79] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, *arXiv preprint arXiv:2010.04159,* 2020.

[80] W. Liu, et al., Ssd: Single shot multibox detector, in: Computer VIsIon–ECCV 2016: 14th European Conference, Amsterdam, The NeTherlands, October 11–14, 2016, ProceedIngs, Part I 14, Springer, 2016, pp. 21–37.

[81] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779-788.

[82] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: Exceeding yolo series in 2021, *arXiv preprint arXiv:2107.08430,* 2021.

[83] Z. Lu, V. Rathod, R. Votel, J. Huang, Retinatrack: Online single stage joint detection and tracking, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14668-14678.

[84] J. Black, T. Ellis, Multiple camera image tracking, in *Proceedings of 2nd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS'01)*, 2001.

[85] J. Black, T. Ellis, D. Makris, Wide area surveillance with a multi camera network, 2004.

[86] D. M. Nguyen, R. Henschel, B. Rosenhahn, D. Sonntag, P. Swoboda, LMGP: Lifted Multicut Meets Geometry Projections for Multi-Camera Multi-Object Tracking, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8866-8875.

[87] X. Zhou, V. Koltun, P. Krähenbühl, Tracking objects as points, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV, Springer, 2020, pp. 474–490.

[88] Y.T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, M. Shah, Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets, Int. J. Comput. Vis. 127 (9) (2019) 1303–1320, https://doi.org/10.1007/s11263-019-01180-6.

[89] H. M. Hsu, Y. Wang, J. Cai, J. N. Hwang, Multi-Target Multi-Camera Tracking of Vehicles by Graph Auto-Encoder and Self-Supervised Camera Link Model, in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 4-8 Jan. 2022 2022, pp. 489-499, doi: 10.1109/WACVW54805.2022.00055.

[90] Y. He, J. Han, W. Yu, X. Hong, X. Wei, Y. Gong, City-Scale Multi-Camera Vehicle Tracking by Semantic Attribute Parsing and Cross-Camera Tracklet Matching, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 14-19 June 2020 2020, pp. 2456-2465, doi: 10.1109/CVPRW50498.2020.00296.

[91] Y. Xu, X. Liu, Y. Liu, and S. C. Zhu, Multi-view People Tracking via Hierarchical Trajectory Composition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 4256-4265, doi: 10.1109/CVPR.2016.461.

[92] A. R. Zamir, A. Dehghan, and M. Shah, GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs, in *European Conference on Computer Vision*, 2012.

[93] A. Roshan Zamir, A. Dehghan, and M. Shah, Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs, in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*, 2012: Springer, pp. 343-356.

[94] W. Chen, L. Cao, X. Chen, K. Huang, An equalized global graph model-based approach for multicamera object tracking, IEEE Trans. Circuits Syst. Video Technol. 27 (11) (2017) 2367–2381, https://doi.org/10.1109/TCSVT.2016.2589619.

[95] B. Yang and R. Nevatia, An online learned CRF model for multi-target tracking, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: IEEE, pp. 2034-2041.

[96] B. Yang and R. Nevatia, Multi-target tracking by online learning of non-linear motion patterns and robust appearance models, in *2012 IEEE conference on computer vision and pattern recognition*, 2012: IEEE, pp. 1918-1925.

[97] A.A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, W. Hu, Multi-object tracking through simultaneous long occlusions and split-merge conditions, in: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, IEEE, 2006, pp. 666–673.

[98] J. He, Z. Huang, N. Wang, and Z. Zhang, Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5299-5309.

[99] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, Multiple hypothesis tracking revisited, in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4696-4704.

[100] I.J. Cox, S.L. Hingorani, An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking, IEEE Trans. Pattern Anal. Mach. Intell. 18 (2) (1996) 138–150.

[101] D. J. Papageorgiou and M. R. Salpukas, The maximum weight independent set problem for data association in multiple hypothesis tracking, in *Optimization and Cooperative Control Strategies: Proceedings of the 8th International Conference on Cooperative Control and Optimization*, 2009: Springer, pp. 235-255.

[102] Z. Khan, T. Balch, F. Dellaert, MCMC-based particle filtering for tracking a variable number of interacting targets, IEEE Trans. Pattern Anal. Mach. Intell. 27 (11) (2005) 1805–1819.
[103] S. Oh, S. Russell, S. Sastry, Markov chain Monte Carlo data association for multi-target tracking, IEEE Trans. Autom. Control 54 (3) (2009) 481–497.
[104] S.S. Beauchemin, J.L. Barron, The computation of optical flow, ACM Computing Surveys (CSUR) 27 (3) (1995) 433–466.
[105] Y. Li, S. Wang, Q. Tian, X. Ding, A survey of recent advances in visual feature detection, Neurocomputing 149 (2015) 736–751.
[106] M. Shahbazi, M.H. Bayat, B. Tarvirdizadeh, A motion model based on recurrent neural networks for visual object tracking, Image Vis. Comput. 126 (2022), 104533.
[107] J. Canny, A Computational Approach to Edge Detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. PAMI-8, no. 6, pp. 679-698, 1986, doi: 10.1109/TPAMI.1986.4767851.
[108] N. Dodd, Multispectral texture synthesis using fractal concepts, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. PAMI-9, no. 5, pp. 703-707, 1987, doi: 10.1109/TPAMI.1987.4767967.
[109] D. Varga, T. Szirányi, A. Kiss, L. Spórás, L. Havasi, A multi-view pedestrian tracking method in an uncalibrated camera network, in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 37-44.
[110] Z. Zhao, S. Yu, X. Wu, C. Wang, Y. Xu, A multi-target tracking algorithm using texture for real-time surveillance, in *2008 IEEE International Conference on Robotics and Biomimetics*, 2009: IEEE, pp. 2150-2155.
[111] F. Bousetouane, L. Dib, H. Snoussi, Improved mean shift integrating texture and color features for robust real time object tracking, Vis. Comput. 29 (2013) 155–170.
[112] L. Qiangqiang, L. Xiling, and Z. Jun, A New Method of Correspondence for Multiple Cameras Based on Texture Energy, in *2009 Second International Conference on Machine Vision*, 2009: IEEE, pp. 264-269.
[113] R. Layne *et al.*, A dataset for persistent multi-target multi-camera tracking in RGB-D, in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 47-55.
[114] K. Heath and L. Guibas, Multi-person tracking from sparse 3D trajectories in a camera sensor network, in *2008 Second ACM/IEEE International Conference on Distributed Smart Cameras*, 2008: IEEE, pp. 1-9.
[115] B.K. Horn, B.G. Schunck, Determining optical flow, Artif. Intell. 17 (1–3) (1981) 185–203.
[116] W. Choi, Near-online multi-target tracking with aggregated local flow descriptor, in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3029-3037.
[117] A. Andriyenko, K. Schindler, and S. Roth, Discrete-continuous optimization for multi-target tracking, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: IEEE, pp. 1926-1933.
[118] A. Andriyenko and K. Schindler, Multi-target tracking by continuous energy minimization, in *CVPR 2011*, 2011: IEEE, pp. 1265-1272.
[119] Z. Tang, G. Wang, H. Xiao, A. Zheng, and J.-N. Hwang, Single-camera and inter-camera vehicle tracking and 3D speed estimation based on fusion of visual and semantic features, in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 108-115.
[120] P. Li *et al.*, Spatio-temporal Consistency and Hierarchical Matching for Multi-Target Multi-Camera Vehicle Tracking, in *CVPR Workshops*, 2019, pp. 222-230.
[121] H.-M. Hsu, T.-W. Huang, G. Wang, J. Cai, Z. Lei, and J.-N. Hwang, Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models, in *CVPR workshops*, 2019, pp. 416-424.
[122] X. Chen, T. Yang, J. Xu, Multi-gait identification based on multilinear analysis and multi-target tracking, Multimed. Tools Appl. 75 (2016) 6505–6532.
[123] N. M. Al-Shakarji, F. Bunyak, G. Seetharaman, and K. Palaniappan, Robust multi-object tracking for wide area motion imagery, in *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2018: IEEE, pp. 1-5.
[124] G. D. Hager, M. Dewan, and C. V. Stewart, Multiple kernel tracking with SSD," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2004, vol. 1: IEEE, pp. I-I.
[125] J. Prokaj, M. Duchaineau, and G. Medioni, Inferring tracklets for multi-object tracking, in *CVPR 2011 WORKSHOPS*, 2011: IEEE, pp. 37-44.
[126] C. Wang, H. K. Galoogahi, C.-H. Lin, and S. Lucey, Deep-LK for efficient adaptive object tracking, in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018: IEEE, pp. 627-634.
[127] C. Shen, A. Van den Hengel, and A. Dick, Probabilistic multiple cue integration for particle filter based tracking, *Australian Pattern Recognition Society,* vol. 2, 2003.
[128] A. Hii, C.E. Hann, J.G. Chase, E.E. Van Houten, Fast normalized cross correlation for motion tracking using basis functions, Comput. Methods Programs Biomed. 82 (2) (2006) 144–156.
[129] A. Nakhmani, A. Tannenbaum, A new distance measure based on generalized image normalized cross-correlation for robust video tracking and image recognition, Pattern Recogn. Lett. 34 (3) (2013) 315–321.
[130] J. Wu, H.-j. Yue, Y.-y. Cao, Z.-m. Cui, Video object tracking method based on normalized cross-correlation matching, in *2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science*, 2010: IEEE, pp. 523-527.
[131] K. Briechle, U.D. Hanebeck, Template matching using fast normalized cross correlation, in: Optical Pattern Recognition XII, vol. 4387, SPIE, 2001, pp. 95–102.
[132] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah, Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets, *arXiv preprint arXiv:1706.06196,* 2017.
[133] D. Chahyati, M.I. Fanany, A.M. Arymurthy, Tracking people by detection using CNN features, Procedia Comput. Sci. 124 (2017) 167–172.
[134] Y. Qi, et al., Hedging deep features for visual tracking, IEEE Trans. Pattern Anal. Mach. Intell. 41 (5) (2018) 1116–1130.
[135] H. Song, Robust visual tracking via online informative feature selection, Electron. Lett 50 (25) (2014) 1931–1933.
[136] Y. Ling *et al.*, Cross-Modality Earth Mover's Distance for Visible Thermal Person Re-Identification, *arXiv preprint arXiv:2203.01675,* 2022.
[137] Q. Zhao, S. Brennan, and H. Tao, Differential EMD tracking, in *2007 IEEE 11th International Conference on Computer Vision*, 2007: IEEE, pp. 1-8.
[138] V. Karavasilis, C. Nikou, A. Likas, Visual tracking using the Earth Mover's Distance between Gaussian mixtures and Kalman filtering, Image Vis. Comput. 29 (5) (2011) 295–305.
[139] P. Kumar, A. Dick, Adaptive earth movers distance-based Bayesian multi-target tracking, IET Comput. Vis. 7 (4) (2013) 246–257.
[140] Q. Zhao, Z. Yang, H. Tao, Differential earth mover's distance with its applications to visual tracking, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2008) 274–287.
[141] Y. Zeng, X. Fu, L. Gao, J. Zhu, H. Li, Y. Li, Robust multivehicle tracking with wasserstein association metric in surveillance videos, IEEE Access 8 (2020) 47863–47876.
[142] R. Siddiqui and S. Khatibi, Visual tracking using particle swarm optimization, *arXiv preprint arXiv:1401.4648,* 2014.
[143] T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, IEEE Trans. Commun. Technol. 15 (1) (1967) 52–60.
[144] N. Lu, Z. Feng, Mathematical model of blob matching and modified Bhattacharyya coefficient, Image Vis. Comput. 26 (10) (2008) 1421–1434.
[145] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, and L. Van Gool, Color-based object tracking in multi-camera environments, in *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*, 2003: Springer, pp. 591-599.
[146] Z. Jiang, D. Q. Huynh, W. Moran, S. Challa, and N. Spadaccini, Multiple pedestrian tracking using colour and motion models, in *2010 International Conference on Digital Image Computing: Techniques and Applications*, 2010: IEEE, pp. 328-334.
[147] E. Bochinski, T. Senst, and T. Sikora, Extending IOU based multi-object tracking by visual information, in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018: IEEE, pp. 1-6.
[148] H.-M. Hsu, J. Cai, Y. Wang, J.-N. Hwang, K.-J. Kim, Multi-target multi-camera tracking of vehicles using metadata-aided re-id and trajectory-based camera link model, IEEE Trans. Image Process. 30 (2021) 5198–5210.
[149] J. Xing, H. Ai, and S. Lao, Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009: IEEE, pp. 1200-1207.
[150] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism, in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4836-4845.
[151] S.-H. Bae, K.-J. Yoon, Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning, in: *ProceedIngs of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1218–1225.
[152] C. Dicle, O.I. Camps, M. Sznaier, The way they move: Tracking multiple targets with similar appearance, in: *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2304–2311.
[153] J.H. Yoon, C.-R. Lee, M.-H. Yang, K.-J. Yoon, Online multi-object tracking via structural constraint event aggregation, in: *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2016, pp. 1392–1400.
[154] K. Shafique, L. Mun Wai, and N. Haering, A rank constrained continuous formulation of multi-frame multi-target tracking problem, in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 23-28 June 2008 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587577.
[155] Q. Yu, G. Medioni, and I. Cohen, Multiple Target Tracking Using Spatio-Temporal Markov Chain Monte Carlo Data Association, in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 17-22 June 2007 2007, pp. 1-8, doi: 10.1109/CVPR.2007.382991.
[156] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, Who are you with and where are you going?, in *CVPR 2011*, 2011: IEEE, pp. 1345-1352.
[157] D. Helbing, P. Molnar, Social force model for pedestrian dynamics, Phys. Rev. E 51 (5) (1995) 4282.
[158] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, Social lstm: Human trajectory prediction in crowded spaces, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.
[159] A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, Learning social etiquette: Human trajectory understanding in crowded scenes, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14, Springer, 2016, pp. 549–565.
[160] H.W. Kuhn, The Hungarian method for the assignment problem, Naval Res. Logistics Quarterly 2 (1–2) (1955) 83–97.
[161] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, Tracking across multiple cameras with disjoint views, in *Computer Vision, IEEE International Conference on*, 2003, vol. 3: IEEE Computer Society, pp. 952-952.
[162] A. Das, A. Chakraborty, A.K. Roy-Chowdhury, Consistent re-identification in a camera network, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part II 13, Springer, 2014, pp. 330–345.

[163] J. F. Henriques, R. Caseiro, and J. Batista, Globally optimal solution to multi-object tracking with merged measurements, in *2011 International Conference on Computer Vision*, 2011: IEEE, pp. 2470-2477.

[164] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part II, Springer, 2016, pp. 17–35.

[165] Y. Xu, X. Liu, Y. Liu, S.-C. Zhu, Multi-view People Tracking via Hierarchical Trajectory Composition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4256–4265.

[166] Y. Xu, X. Liu, L. Qin, and S.-C. Zhu, Cross-view people tracking by scene-centered spatio-temporal parsing, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, vol. 31, no. 1.

[167] J. Williams, R. Lau, Approximate evaluation of marginal association probabilities with belief propagation, IEEE Trans. Aerosp. Electron. Syst. 50 (4) (2014) 2942–2959.

[168] Q. Li, J. Sun, and W. Sun, An efficient multiple hypothesis tracker using max product belief propagation, in *2017 20th International Conference on Information Fusion (Fusion)*, 2017: IEEE, pp. 1-6.

[169] L. Cao, et al., Convex variational inference for multi-hypothesis fractional belief propagation based data association in multiple target tracking systems, IEEE Sens. J. 21 (17) (2021) 19121–19133.

[170] F. Meyer, P. Braca, P. Willett, F. Hlawatsch, A scalable algorithm for tracking an unknown number of targets using multiple sensors, IEEE Trans. Signal Process. 65 (13) (2017) 3478–3493.

[171] M. Sun, M.E. Davies, I.K. Proudler, J.R. Hopgood, Adaptive kernel Kalman filter based belief propagation algorithm for maneuvering multi-target tracking, IEEE Signal Process Lett. 29 (2022) 1452–1456.

[172] S. Chen, A. Fern, S. Todorovic, Multi-object tracking via constrained sequential labeling, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1130–1137.

[173] A.A. Butt, R.T. Collins, Multi-target tracking by lagrangian relaxation to min-cost network flow, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1846–1853.

[174] H.B. Shitrit, J. Berclaz, F. Fleuret, P. Fua, Multi-commodity network flow for tracking multiple people, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2013) 1614–1627.

[175] L. Zhang, Y. Li, and R. Nevatia, Global data association for multi-object tracking using network flows, in *2008 IEEE conference on computer vision and pattern recognition*, 2008: IEEE, pp. 1-8.

[176] C. Stauffer, Estimating tracking sources and sinks, in *2003 Conference on Computer Vision and Pattern Recognition Workshop*, 2003, vol. 4: IEEE, pp. 35-35.

[177] B. Leibe, K. Schindler, and L. Van Gool, Coupled detection and trajectory estimation for multi-object tracking, in *2007 IEEE 11th International Conference on Computer Vision*, 2007: IEEE, pp. 1-8.

[178] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, Globally-optimal greedy algorithms for tracking a variable number of objects, in *CVPR 2011*, 2011: IEEE, pp. 1201-1208.

[179] M. Tian, Y. Bo, Z. Chen, P. Wu, C. Yue, Multi-target tracking method based on improved firefly algorithm optimized particle filter, Neurocomputing 359 (2019) 438–448.

[180] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn, Branch-and-price global optimization for multi-view multi-target tracking, in: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: IEEE, pp. 1987-1994.

[181] B. Rosenhahn, G. Pons-Moll, and L. Leal-Taixe, Branch-and-price global optimization for multi-view multi-target tracking, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: IEEE Computer Society, pp. 1987-1994.

[182] S.-I. Yu, D. Meng, W. Zuo, A. Hauptmann, The solution path algorithm for identity-aware multi-object tracking, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3871–3879.

[183] F. Fleuret, J. Berclaz, R. Lengagne, P. Fua, Multicamera people tracking with a probabilistic occupancy map, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 267–282, https://doi.org/10.1109/TPAMI.2007.1174.

[184] J. Black, T. Ellis, P. Rosin, Multi view image surveillance and tracking, in: Proceedings, IEEE, 2002, 2002, pp. 169–174.

[185] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, J. Yuan, Track to detect and segment: An online multi-object tracker, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12352–12361.

[186] L. Zheng, M. Tang, Y. Chen, G. Zhu, J. Wang, H. Lu, Improving multiple object tracking with single object tracking, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2453–2462.

[187] Y. Si, Y. Zhang, IAMOT: Multi-object tracking with integrated heads and attention mechanism, Neurocomputing 510 (2022) 95–106.

[188] C. Zhou, M. Jiang, J. Kong, BGTracker: cross-task bidirectional guidance strategy for multiple object tracking, IEEE Trans. Multimedia (2023).

[189] Y. Cheng *et al.*, Segment and track anything, *arXiv preprint arXiv:2305.06558,* 2023.

[190] J. Yin, W. Wang, Q. Meng, R. Yang, J. Shen, A unified object motion and affinity model for online multi-object tracking, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6768–6777.

[191] C. Liu, et al., City-scale multi-camera vehicle tracking guided by crossroad zones, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4129–4137.

[192] Z. Tang, et al., Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8797–8806.

[193] Y. Qian, L. Yu, W. Liu, A.G. Hauptmann, Electricity: An efficient multi-camera vehicle tracking system for intelligent city, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 588–589.

[194] Z. He, Y. Lei, S. Bai, W. Wu, Multi-camera vehicle tracking with powerful visual features and spatial-temporal cue, CVPR Workshops (2019) 203–212.

[195] P. Li, J. Zhang, Z. Zhu, Y. Li, L. Jiang, G. Huang, State-aware re-identification feature for multi-target multi-camera tracking. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[196] S. You, H. Yao, C. Xu, Multi-target multi-camera tracking with optical-based pose association, IEEE Trans. Circuits Syst. Video Technol. 31 (8) (2020) 3105–3117.

[197] Y. Hou, L. Zheng, Z. Wang, and S. Wang, Locality aware appearance metric for multi-target multi-camera tracking, *arXiv preprint arXiv:1911.12037,* 2019.

[198] J. Tang *et al.*, Min-cost network flow and trajectory fix for multiple objects tracking, in *Conference on Computer Vision and Pattern Recognition Workshop*, 2020, vol. 2, p. 3.

[199] T. Mandel, et al., Detection confidence driven multi-object tracking to recover reliable tracks from unreliable detections, Pattern Recogn. 135 (2023), 109107.

[200] F. Du, B. Xu, J. Tang, Y. Zhang, F. Wang, H. Li, 1st place solution to eccv-tao-2020: Detect and represent any object for tracking, *arXiv preprint arXiv: 2101.08040,* 2021.

[201] J. Peng, et al., TPM: Multiple object tracking with tracklet-plane matching, Pattern Recogn. 107 (2020), 107480.

[202] R. Henschel, L. Leal-Taixé, D. Cremers, B. Rosenhahn, Fusion of head and full-body detectors for multi-object tracking, in: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1428–1437.

[203] J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon, Bayesian multi-object tracking using motion context from multiple objects, in *2015 IEEE Winter Conference on Applications of Computer Vision*, 2015: IEEE, pp. 33-40.

[204] W. Hua, D. Mu, Z. Zheng, D. Guo, Online multi-person tracking assist by high-performance detection, J. Supercomput. 76 (2020) 4076–4094.

[205] C. Huang, Y. Li, R. Nevatia, Multiple target tracking by learning-based hierarchical association of detection responses, IEEE Trans. Pattern Anal. Mach. Intell. 35 (4) (2012) 898–910.

[206] B. Wang, G. Wang, K.L. Chan, L. Wang, Tracklet association by online target-specific metric learning and coherent dynamics estimation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (3) (2016) 589–602.

[207] L. Lin, Y. Lu, C. Li, H. Cheng, W. Zuo, Detection-free multiobject tracking by reconfigurable inference with bundle representations, IEEE Trans. Cybern. 46 (11) (2016) 2447–2458, https://doi.org/10.1109/TCYB.2015.2478515.

[208] R. Hoseinnezhad, B.-N. Vo, D. Suter, and B.-T. Vo, Multi-object filtering from image sequence without detection, in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010: IEEE, pp. 1154-1157.

[209] K. Fragkiadaki and J. Shi, Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement, in *CVPR 2011*, 20-25 June 2011 2011, pp. 2073-2080, doi: 10.1109/CVPR.2011.5995366.

[210] Z. Cai, S. Hu, Y. Shi, Q. Wang, D. Zhang, Multiple human tracking based on distributed collaborative cameras, Multimed. Tools Appl. 76 (2017) 1941–1957.

[211] M. Han, W. Xu, H. Tao, and Y. Gong, An algorithm for multiple object trajectory tracking, in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2004, vol. 1: IEEE, pp. I-I.

[212] Q. Jiang, M. Liu, X. Fu, Y. Shang, and H. Ding, Pedestrian Tracking Based on HSV Color Features and Reconstruction by Contributions, in *2018 5th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)*, 2018: IEEE, pp. 7-12.

[213] R. T. Collins, Mean-shift blob tracking through scale space, in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, 2003, vol. 2: IEEE, pp. II-234.

[214] K.-Y. Kim, J.-S. Kwon, and K.-S. Cho, "Multi-object tracker using kemelized correlation filter based on appearance and motion model," in *2017 19th International Conference on Advanced Communication Technology (ICACT)*, 2017: IEEE, pp. 761–764.

[215] S. Khan, O. Javed, and M. Shah, Tracking in uncalibrated cameras with overlapping field of view, in *2nd IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, 2001, vol. 5: IEEE Computer Society Press Los Alamitos.

[216] K. Kim, L.S. Davis, Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering, in: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006, Proceedings, Part III 9, Springer, 2006, pp. 98–109.

[217] R. Kasturi, et al., Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2008) 319–336.

[218] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, MOT16: A benchmark for multi-object tracking, *arXiv preprint arXiv:1603.00831,* 2016.

[219] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking, *arXiv preprint arXiv:1504.01942,* 2015.

[220] J. Ferryman and A. Shahrokni, Pets2009: Dataset and challenge, in *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance*, 2009: IEEE, pp. 1-6.

[221] J. Black, T. Ellis, and P. Rosin, A novel method for video tracking performance evaluation, *Proceedings of the IEEE InternationalWorkshop on Visual Surveillance and*

Performance Evaluation of Tracking and Surveillance (VS-PETS 03), pp. 125-132, 2003.
[222] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the clear mot metrics, EURASIP J. Image Video Processing 2008 (2008) 1–10.
[223] K. Smith, D. Gatica-Perez, J. Odobez, and S. Ba, Evaluating multi-object tracking, in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)-workshops*, 2005: IEEE, pp. 36-36.
[224] B. Wu and R. Nevatia, Tracking of multiple, partially occluded humans based on static body part detection, in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006, vol. 1: IEEE, pp. 951-958.
[225] O. Russakovsky, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.
[226] Y. Li, C. Huang, and R. Nevatia, Learning to associate: Hybridboosted multi-target tracker for crowded scene, in *2009 IEEE conference on computer vision and pattern recognition*, 2009: IEEE, pp. 2953-2960.
[227] I. Leichter, E. Krupka, Monotonicity and error type differentiability in performance measures for target detection and tracking in video, IEEE Trans. Pattern Anal. Mach. Intell. 35 (10) (2013) 2553–2560.
[228] L. Wen, et al., UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking, Comput. Vis. Image Underst. 193 (2020), 102907.
[229] X. Weng, J. Wang, D. Held, and K. Kitani, 3d multi-object tracking: A baseline and new evaluation metrics, in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020: IEEE, pp. 10359-10366.
[230] J. Luiten, et al., Hota: A higher order metric for evaluating multi-object tracking, Int. J. Comput. Vis. 129 (2021) 548–578.
[231] L. Yang, Y. Fan, N. Xu, Video instance segmentation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5188–5197.
[232] P. Zhu, et al., Detection and tracking meet drones challenge, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2021) 7380–7399.
[233] A. Dave, T. Khurana, P. Tokmakov, C. Schmid, D. Ramanan, Tao: A large-scale benchmark for tracking any object, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, 2020, pp. 436–454.
[234] C.-H. Kuo, C. Huang, and R. Nevatia, Inter-camera association of multi-target tracks by on-line learned appearance affinity models, in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, 2010: Citeseer, pp. 383-396.
[235] S. Zhang, E. Staudt, T. Faltemier, and A. K. Roy-Chowdhury, A camera network tracking (camnet) dataset and performance baseline, in *2015 IEEE Winter Conference on Applications of Computer Vision*, 2015: IEEE, pp. 365-372.
[236] L. Cao, W. Chen, X. Chen, S. Zheng, and K. Huang, An equalised global graphical model-based approach for multi-camera object tracking, *arXiv preprint arXiv: 1502.03532,* vol. 8, 2015.
[237] X. Han, et al., Mmptrack: Large-scale densely annotated multi-camera multiple people tracking benchmark, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4860–4869.
[238] P. Kohl, A. Specker, A. Schumann, J. Beyerer, The mta dataset for multi-target multi-camera pedestrian tracking by weighted distance aggregation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1042–1043.
[239] Y. Hou, L. Zheng, S. Gould, Multiview detection with feature perspective transformation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, Springer, 2020, pp. 1–18.
[240] X. Sun, L. Zheng, Dissecting person re-identification from the viewpoint of viewpoint, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 608–617.
[241] X. Yang, et al., Box-grained reranking matching for multi-camera multi-target tracking, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3096–3106.
[242] P. Li *et al.*, Spatio-temporal Consistency and Hierarchical Matching for Multi-Target Multi-Camera Vehicle Tracking, in *CVPR Workshops*, 2019.
[243] H.-M. Hsu, Y. Wang, J. Cai, J.-N. Hwang, Multi-target multi-camera tracking of vehicles by graph auto-encoder and self-supervised camera link model, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 489–499.
[244] J. Ye, et al., A robust mtmc tracking system for ai-city challenge 2021, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4044–4053.
[245] F. Li, et al., Multi-camera vehicle tracking system for AI City Challenge 2022, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3265–3273.
[246] C. Liu *et al.*, City-Scale Multi-Camera Vehicle Tracking Guided by Crossroad Zones, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 19-25 June 2021 2021, pp. 4124-4132, doi: 10.1109/CVPRW53098.2021.00466.
[247] G. Wang, Y. Wang, H. Zhang, R. Gu, J.-N. Hwang, Exploit the connectivity: Multi-object tracking with trackletnet, in: *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 482–490.
[248] P. Nguyen, K. G. Quach, C. N. Duong, S. L. Phung, N. Le, and K. Luu, Multi-Camera Multi-Object Tracking on the Move via Single-Stage Global Association Approach, *arXiv preprint arXiv:2211.09663,* 2022.
[249] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, Performance Measures and a Data Set for Multi-target, Multi-camera Tracking, in *ECCV Workshops*, 2016.
[250] Y.T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, M. Shah, Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets, Int. J. Comput. Vis. (2017) 1–18.
[251] Y.T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, M. Shah, Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets, Int. J. Comput. Vis. 127 (2019) 1303–1320.
[252] N. Jiang, S. Bai, Y. Xu, C. Xing, Z. Zhou, and W. Wu, Online Inter-Camera Trajectory Association Exploiting Person Re-Identification and Camera Topology, presented at the Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 2018. [Online]. Available: https://doi.org/10.1145/3240508.3240663.
[253] Y.-G. Lee, Z. Tang, J.-N. Hwang, Online-learning-based human tracking across non-overlapping cameras, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2017) 2870–2883.
[254] H.-M. Hsu, T.-W. Huang, G. Wang, J. Cai, Z. Lei, and J.-N. Hwang, Multi-Camera Tracking of Vehicles based on Deep Features Re-ID and Trajectory-Based Camera Link Models, in *CVPR Workshops*, 2019.
[255] Z. He, Y. Lei, S. Bai, and W. Wu, Multi-Camera Vehicle Tracking with Powerful Visual Features and Spatial-Temporal Cue, in *CVPR Workshops*, 2019.
[256] Y. Qian, L. Yu, W. Liu, and A. G. Hauptmann, ELECTRICITY: An Efficient Multi-camera Vehicle Tracking System for Intelligent City, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 2511-2519, 2020.

TEMITOPE IBRAHIM AMOSA received the B.Eng. degree in computer engineering from the University of Ilorin, Nigeria, in 2019. He is currently pursuing the MSc. degree with the Department of Electrical and Electronics Engineering, Universiti Teknologi PETRONAS, Malaysia. His research interests lie at the intersection of natural language processing and computer vision for surveillance and biomedical applications. He is a graduate member of The Nigerian Society of Engineers, The Institute of Electrical & Electronics Engineers, the Black in AI and the IEEE computer society

PATRICK SEBASTIAN received the B.Eng. degree in electrical engineering from Universiti Teknologi Malaysia, Malaysia, in 1993, the M.Sc. degree from the University of Minnesota, Minneapolis, MN, USA, in 2003, and the Ph.D. degree in video surveillance from Middlesex University, London, U.K., in 2016. He is currently working as a Senior Lecturer with the Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS (UTP). His current research interests include signal and imaging research, and intelligent video surveillance.

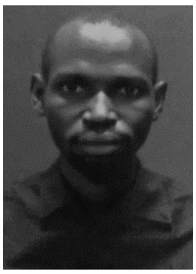
LILA IZNITA IZHAR received the B.Eng. degree in electrical and electronic engineering from the University of the Ryukyus, Japan, M.Sc. and Ph.D. degrees in electrical and electronic engineering from Universiti Teknologi PETRONAS (UTP), and the Diploma degree in electrical and electronic engineering from Imperial College London. She is currently a Senior Lecturer at the Electrical and Electronic Engineering Department, UTP. She has been in the department for more than ten years. She has over 17 years of experience doing research in medical image processing and analysis. This has cultivated her interest in intelligent signal and imaging research for medical/health application. Her work has been published in various journals and she has presented at several papers from her research in various top biomedical/imaging conferences including the IEEE Engineering in Medicine and Biology Society and the European Conference of the International Federation for Medical and Biological Engineering.

OLADIMEJI IBRAHIM received the B.Eng. degree in electrical engineering from the University of Ilorin, Nigeria, in 2005, MSc. in applied instrumentation and control from Glasgow Caledonian University, U.K., in 2009, and Ph.D in electrical and electronic from Universiti Teknologi PETRONAS (UTP). He is a Lecturer with the Department of Electrical and Electronics Engineering, University of Ilorin. His research interests include the dynamics and control of power converters, interface for renewable energy systems, state space control, smart grid, and demand side management. He is a member of The Nigerian Society of Engineers, The Institute of Electrical & Electronics Engineers, and The Institute of Measurement and Control.

AYINLA SHEHU LUKMAN received the B.Eng in electrical and electronics engineering, and M.Eng in Computer Engineering from Nnamdi Azikiwe University, Awka, Nigeria in 2006 and 2017 respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronics Engineering, Universiti Teknologi PETRONAS, Malaysia. He is a Lecturer in the Department of Computer Engineering, University of Ilorin, Ilorin, Nigeria. His research interests include Indoor localization, Wireless Sensor Networking, and Machine Learning applications for wireless communication.

BAHASHWAN ABDULRAHMAN ABDULLAH OMAR received a B.Eng. degree in electrical and electronics engineering from Infrastructure University Kuala Lumpur, Malaysia. He is currently pursuing MSc. degree with the Department of Electrical and Electronics Engineering, Universiti Teknologi PETRONAS, Malaysia. His areas of interest lie in artificial intelligence, predictive maintenance, and computer vision for tracking systems.

ABUBAKAR BALA received the bachelor's degree (Hons.) in computer engineering from Bayero University Kano, Nigeria, in 2011, and the master's degree in computer engineering from the King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, in 2015. He also obtained the Ph.D. degree in electrical engineering from the Universiti Teknologi PETRONAS (UTP), Malaysia, in 2021. He is currently a research scientist in the same University, and a Lecturer with Bayero University Kano (BUK), Nigeria. He is working on developing new algorithms to optimize the echo state network (ESN), a type of recurrent neural network. His research interests include optimization, artificial neural networks, fault predictions, cloud computing, and renewable energy.



YAU ALHAJI SAMAILA received the bachelor's degree (Hons.) in Electrical and Electronics Engineering from the University of Maiduguri, Nigeria, in 2010, and master's degree Electrical and Electronics Engineering from the Bayero University Kano, Nigeria, in 2018 respectively. He is currently pursuing the Ph.D. degree in Electrical and Electronics with the Universiti Teknologi Petronas, Malaysia. His research area of interest is Computer vision and Digital Signal processing.