Multi-Target Multi-Camera Tracking and Re-Identification

from Detection to Tracking in Real-Time Scenarios

Research Project
Study program Computer Science & Engineering
Faculty of Information, Media and Electrical Engineering
Cologne University of Applied Sciences

presented by:	Luca Uckermann
matriculation number:	111 337 75
address:	Elisenstr. 29 51149 Cologne luca_simon.uckermann@smail.th-koeln.de
submitted to:	Prof. Dr. Jan Salmen

Cologne, 2023-11-14

Declaration

I certify that I have written the submitted work independently. All passages taken verbatim or in spirit from the published or unpublished work of others, or from the author's own work, are marked as taken. All sources and tools used in the work are acknowledged. The work has not been submitted to any other examination authority with the same content or in substantial parts.

Place, Date	Signature

Abstract

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Contents

1 Introduction

Multi-Target Multi-Camera Tracking (MTMCT) is an essential field of research in computer vision, with significant applications ranging from video surveillance and traffic monitoring to sports analysis and crowd management. By simultaneously tracking multiple objects across various camera views, MTMCT systems aim to provide a comprehensive understanding of the scene dynamics and interactions.

The advent of deep learning and other advanced algorithms has revolutionized the field of MTMCT, especially in the last years, enabling faster, more accurate and reliable tracking in complex environments. In particular, online and real-time tracking methods have emerged as a critical area of focus, given their potential to provide timely and actionable insights in various real-world applications.

Even though Single-Target Single-Camera Tracking (ST-SCT) as well as Multi-Target Single-Camera Tracking (MT-SCT) has been extensively studied, MTMCT is still a relatively new and challenging, but also promising area of research. The complexity of MTMCT is significantly higher than ST-SCT and MT-SCT, due to the need to simultaneously track multiple objects across multiple cameras.

Single-Target Multi-Camera (ST-MCT) is a insignificant field of research, because if the use-case requires multiple cameras, it is almost always necessary to track multiple targets. Therefore, this project will not cover the special case of ST-MCT.

This research project aims to provide a comprehensive review of the state-of-the-art in MTMCT, with a special focus on online and real-time tracking methods. Latest trends, technologies, and challenges in this field are explored, drawing insights from recent research papers and studies. This review highlights the significant advancements made in MTMCT and identifies the gaps and opportunities for future research.

The rest of this project is structured as follows. Chapter 2 provides an overview of the key challenges and issues in MTMCT, along with a discussion of the datasets, metrics, and components of an MTMCT system. Futhermore it explains the basic concepts of object detection and tracking. Chapter 3 presents a detailed review of the literature on MTMCT, with the main focus on online and real-time tracking methods with static cameras. Chapter 4 compares and contrasts the different methods reviewed in the previous sections, identifies the gaps and limitations in current research, and suggests areas for future research. While considering the ethical and privacy concerns related to MTMCT, it also discusses the need for regulations and guidelines. Finally, Chapter 5 concludes the project with a summary of the key findings and insights, along with stating the future directions and challenges for research in this area.

1.1 Definition of MTMCT

MTMCT is an integration of object detection and tracking methodologies to simultaneously track multiple predefined objects of interest across various camera views. The objective of MTMCT is to maintain a coherent understanding of the identities (IDs) of the objects and trajectories as they move through the fields of view of different cameras. The objects of interest are often people and vehicles, but in theory can be any moving object. The camera setup differs from one application to another, but typically consists of multiple cameras with either overlapping, non-overlapping, or partially overlapping fields of view. The cameras may be static or moving, and may be placed at different heights and angles. The cameras may also have differing technical specifications like resolution, frame rate, and field of view (FOV).

1.2 Importance of MTMCT

MTMCT plays a crucial role in various real-world applications. In video surveillance, it is used to monitor and analyze the movement of individuals or vehicles across different cameras, which can be vital for security and forensic analysis. In sports analysis, MTMCT can provide valuable insights by tracking the movement and interaction of players across different camera angles. In traffic monitoring, MTMCT can help manage traffic flow and detect incidents by tracking vehicles as they move through different camera views.

Furthermore, the need for online and real-time tracking in these applications is imperative. Real-time processing of data streams from multiple cameras and providing instantaneous tracking results are essential to make timely and actionable insights, which is particularly relevant in scenarios like accident prevention, control of traffic flow, crime detection, and real-time sports analysis.

1.3 Objective of Research Project

First, the basics concepts of SCT are explained to provide a foundation for understanding MTMCT. The primary objective of this project is to provide a comprehensive overview of proposed methods and technologies for MTMCT and review the current state-of-the-art in MTMCT, with a special focus on online and real-time tracking methods. Through an extensive literature review, the aim is to explore the latest trends, technologies, and challenges faced in this field, and provide insights drawn from recent research papers and studies. By highlighting the significant advancements made in MTMCT, the intend is to identify the gaps in current research and outline potential avenues for future exploration, while keeping in mind the ethical and privacy concerns related to MTMCT.

1.4 Related Work

The work “Person Re-identification: Past, Present and Future” by Zheng, Yang, and Hauptmann [1] focuses on the history, present and future of person re-identification. Although this paper gives a good overview, it was published in 2016 and therefore does not cover the latest research in this field, which will be covered by this project.

The doctoral thesis of Tian [2, Chapter 5], published in 2019, revolves around the topic of tracking multiple objects and gives a state-of-the-art overview of this field. It does not cover the topic of multi-camera tracking. However, it provides a mathematical insight into the topic of tracking multiple objects.

The most recent and comprehensive review of MTMCT was published in 2023 by Amosa, Sebastian, Izhar, *et al.* [3]. It provides a detailed overview of the state-of-the-art in MTMCT, covering the latest trends, technologies, and challenges in this field. However, the mentioned review gives a broader overview and does not focus on online and real-time tracking methods, which is the main focus of this project. Futhermore, this research project aims to provide an easier introduction to the field of MTMCT by first explaining the basics before diving into the details of the latest research.

2 Background

This chapter provides an overview of the basic concepts of object detection and tracking and the steps of an Multi-Target Multi-Camera Tracking (MTMCT) system, along with a discussion of its key challenges and issues. Also the foundational building blocks of MTMCT are introduced, namely Single-Target Single-Camera Tracking (ST-SCT) and Multi-Target Single-Camera Tracking (MT-SCT). Futhermore it explains the datasets and metrics used to evaluate MTMCT systems.

2.1 Steps of an MTMCT System

An MTMCT system typically consists of the following steps: detection, feature extraction, data association, and tracking. Only the basic and fundamental concepts are explained in this section, more advanced and recent methods, mostly revolving around deep learning, will be discussed in chapter 3.

2.1.1 Detection

Detection refers to the process of identifying objects of interest within video frames. This is typically done using a variety of techniques, ranging from traditional image processing methods to deep learning models. The objective of the detection step is to locate and classify objects in the frame, providing a basis for subsequent steps in the MTMCT process.

2.1.2 Feature Extraction

Feature extraction involves extracting relevant information from detected objects to facilitate tracking. This could include low-level features like color, shape and texture as well as high-level features like object parts and their spatial relationships, speed, and direction of movement. The features extracted from objects are used to identify and distinguish them from other objects in the scene.

2.1.3 Data Association

Data association is the process of associating current detected objects with existing tracks based on similarities in their features. This is done by comparing the features of detected objects with the features of existing tracks and assigning the detected objects to the most similar tracks. This step is critical in maintaining the identity of objects as they move through the scene or even leaving and re-entering the scene, which is called re-identification (re-ID). Commonly the data association step is first performed in a hierarchical manner: first on a single camera view (intra-camera), before the tracks are associated across multiple camera views (inter-camera) and finally being optimized globally.

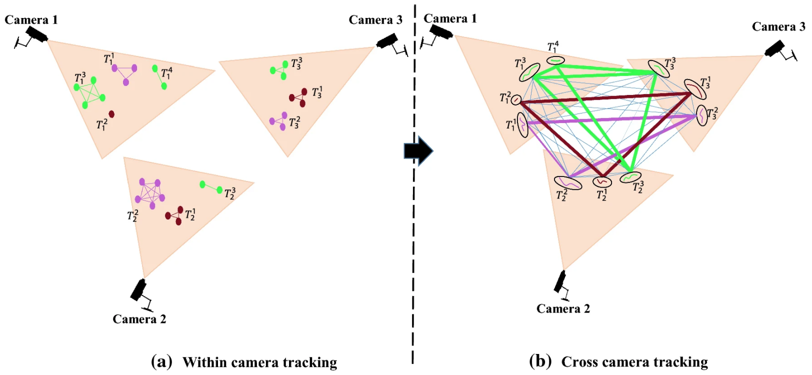

Figure 2.1: Intra- and inter-camera tracking [4, Fig. 1]

The two steps of data association of three non-overlapping camera views are illustrated in Figure 2.1. The first step is intra-camera tracking, where the tracks are associated independently within the three camera views. The second step is inter-camera tracking, where the tracks are associated across the three camera views and the IDs of the objects are maintained. This simple example can be extended to any number of camera views, overlapping or non-overlapping.

2.1.4 Tracking

Tracking refers to the step of maintaining the trajectory of detected objects over time. This involves predicting the future location of an object based on its past movements and updating its trajectory as new observations, so the next frame of a video, become available. To sum it up tracking is responsible for maintaining and managing the tracks and IDs of objects as they move through the scene and to ensure consistent global IDs across multiple camera views.

2.2 Fundamental Concepts

This section briefly describes the preliminary concepts of MTMCT, which are essential to follow the progression from basic object tracking methods to advanced MTMCT techniques.

2.2.1 Single-Target Single-Camera Tracking (ST-SCT)

ST-SCT is the simplest form of object tracking and involves tracking a single target in the field of view of a single camera. The primary goal of ST-SCT is to maintain the identity (ID) and trajectory of the target as it moves through the view of the camera.

2.2.2 Multi-Target Single-Camera Tracking (MT-SCT)

MT-SCT builds upon the principles of ST-SCT but introduces the added complexity of dealing with multiple targets in a view of a single-camera. It aims to track multiple objects simultaneously while maintaining the ID of each target and avoiding ID switches. This requires sophisticated algorithms that can handle occlusions, interactions between targets, and other challenges that especially arise in crowded scenes.

The progression from ST-SCT to MT-SCT, and ultimately to MTMCT, reflects the increasing complexity and capability of tracking systems to handle more complex scenarios. This evolution is possible, due to advances in computer vision and machine learning, which provide the tools necessary to tackle the challenges associated with tracking multiple targets across multiple camera views.

2.3 Challenges and Issues

The process of tracking multiple objects across various camera views requires careful consideration of various factors that can significantly affect the performance and accuracy of the tracking system. Some of the main challenges and issues faced in MTMCT are discussed in the following sections.

2.3.1 Occlusion

Occlusion occurs when an object is partially or completely blocked from view, making it difficult to accurately track its position and identity. This can happen when objects overlap with each other or are obstructed by other elements in the scene, such as buildings or trees. Occlusion is a common challenge in crowded environments, such as public spaces and sporting events, where multiple objects are often in close proximity to each other.

2.3.2 Varying Lighting Conditions

Lighting conditions can have a significant impact on the performance of an MTMCT system. Variations in lighting, such as changes in natural light throughout the day or artificial lighting when a tracked object enters a building, can affect the appearance of objects and make it challenging to maintain consistent tracking. The presence of shadows and reflections can also complicate the tracking process.

2.3.3 Camera Specifications

The specifications of the cameras used in an MTMCT system can have a significant impact on its performance. When multiple cameras are used, they may have different:

- **Resolution:** The number of pixels in the image
- **Frame rate:** The number of frames captured per second
- **Field of view (FOV):** The area captured by the camera
- **Angle:** The angle from which the camera captures the scene

This can make it challenging to maintain consistent tracking across different camera views, especially when objects move from one camera to another. Objects may appear differently when viewed from different cameras, and their size and shape can be distorted. Achieving accurate tracking requires the system to account for these variations and correctly align objects across different camera views.

2.3.4 Uncertainties

In a MTMCT system, the number of present objects in the entire camera network, in a single camera view, and the number of camera views in which a tracked object is present at a given time are all unknown. This uncertainty complicates the precise tracking of objects across multiple camera views.

2.4 Datasets

Datasets are a fundamental aspect of MTMCT research, they are the resource for the training, evaluation, and comparison of various tracking methods. A diverse array of datasets exists to fullfil requirements of MTMCT research, each offering unique challenges and scenarios.

Commonly utilized datasets to train object detectors are:

- **Microsoft COCO (Common Objects in Context):** Comprehensive dataset utilized for object detection, segmentation, and captioning. COCO comprises a diverse range of objects [5].
- **ImageNet:** Vast dataset employed for image classification and object detection. Object detectors trained on ImageNet are able to recognize an broad range of objects [6].

Beside these datasets, there are several datasets specifically designed for MTMCT research. These datasets are discussed in subsection 3.2.5.

2.5 Metrics and Evaluation

Evaluating the performance of a MTMCT system is critical to understand its effectiveness and reliability. Beside well known metrics like accuracy, precision and recall, there are several metrics specifically designed for multi-target and multi-camera systems. These metrics are discussed in the this section.

2.5.1 MOTP and MOTA

The Multiple Object Tracking Precision (MOTP) and Multiple Object Tracking Accuracy (MOTA) are two standard metrics used for evaluating multi-target tracking systems. MOTP measures the accuracy of the object localization, while MOTA combines three types of errors into a single metric to provide a comprehensive evaluation of the tracking performance. Both of these metrics were introduced by Bernardin and Stiefelhagen [7] in 2008.

$$\text{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \quad [7, \text{Eq. } 1] \tag{2.1}$$

Equation 2.1 provides a measure of the average error in estimated positions of the tracked objects. In this equation, d_t^i represents the distance between the predicted position and the ground truth position of object i at frame t, and c_t is the number of correctly matched objects (the true positives) in frame t. The distances for all matched objects across all frames is divided by the total number of matched objects across all frames. MOTP ranges from 0 to 1, a lower MOTP value indicates higher precision in the object localization.

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \quad [7, \text{Eq. } 2] \tag{2.2}$$

Equation 2.2 combines three types of errors to give a single performance measure. In this equation, m_t is the number of misses (true objects not detected), fp_t is the number of false positives (spurious object detections), mme_t is the number of mismatch errors (identity switches) and g_t is the total number of true objects present in frame t. The MOTA score is 1 minus the sum of all errors divided by the total number of true objects across all frames. MOTA ranges from $-\infty$ to 1, a higher MOTA value indicates better tracking accuracy.

2.5.2 IDF1

The IDF1 score is another important metric for evaluating MTMCT systems. It represents the harmonic mean of the identification precision and recall, providing a balanced measure that accounts for both the ratio of correctly identified detections and the average number of ground-truth and computed detections. This metric was introduced by Ristani, Solera, Zou, *et al.* in their widely referenced paper "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking" [8].

$$\text{IDF}_1 = \frac{2 \times \text{IDTP}}{2 \times \text{IDTP} + \text{IDFP} + \text{IDFN}} \quad [8, \text{Eq. } 11] \tag{2.3}$$

In equation 2.3:

- **IDTP (Identification True Positives)**: Represents the number of detections that were correctly identified.
- **IDFP (Identification False Positives)**: Denotes the number of detections that were wrongly identified (misidentifications).
- **IDFN (Identification False Negatives)**: Indicates the number of actual detections that were missed or not identified.

The IDF1 metric essentially captures the identification precision and recall in multi-object tracking scenarios. The higher the IDF1 score, the better the performance of the tracker in maintaining consistent identities.

2.5.3 MT and ML

The Mostly Tracked (ML) and Mostly Lost (ML) are used to assess the effectiveness of a tracking system in maintaining consistent trajectories for the objects being tracked. The metrics published by Wu and Nevatia [9] in 2006 are commonly used in the MOTChallenge benchmarks to evaluate the performance of tracking systems.

MT measures the proportion of ground truth trajectories that are covered by the tracker for at least 80% of their respective lifetimes, indicating the ability of the system to consistently track objects over time. On the other hand, ML measures the proportion of ground truth trajectories that are covered by the tracker for less than 20% of their respective lifetimes, reflecting the inability of the system to maintain consistent object tracking.

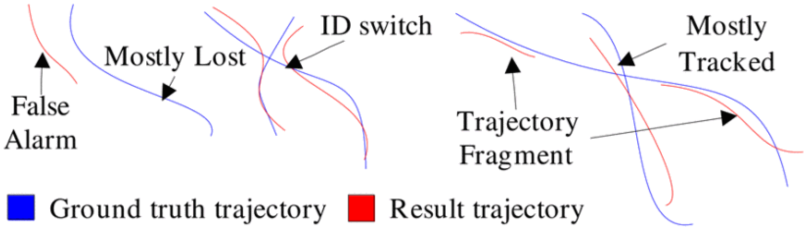

Figure 2.2: MT and ML [9, Fig. 5]

Figure 2.2 illustrates various scenarios encountered in multi-target tracking evaluations:

- **Ground Truth Trajectory (Blue):** Represents the actual path or movement of an object in the scene.
- **Result Trajectory (Red):** Represents the predicted path of an object by the tracking system.
- **False Alarm:** Points where the tracking system detects an object when there is no one present in the ground truth.
- **ID Switch:** An instance where the tracking ID assigned to an object changes erroneously during tracking.
- **Trajectory Fragment:** A segment of the result trajectory that is shorter than the ground truth, indicating a break or interruption in tracking.
- **Mostly Tracked:** Scenarios where the result trajectory closely follows the ground truth trajectory for the majority of the path of the object ($\geq 80\%$)
- **Mostly Lost:** Scenarios where the result trajectory only briefly aligns or intersects with the ground truth trajectory, indicating the object was not effectively tracked for most of its path ($\leq 20\%$).

3 Literature Review

This chapter reviews the literature on Multi-Target Multi-Camera Tracking (MTMCT) and discusses the trends and advancements as well as the milestones in this field. It will only focus on the latest and state-of-the-art methods and technologies and will not cover the whole history of MTMCT including all the past algorithms and methods. The chapter is structured as follows: Section 3.1 discusses the beginnings of MTMCT, Section 3.2 highlights the milestones in MTMCT, Section 3.3 reviews the methods and algorithms used in MTMCT, and Section 3.4 discusses the strengths and weaknesses of the reviewed methods and algorithms.

3.1 The Beginnings

Back in 1999 and 2001 Cai and Aggarwal [10] and Chang and Gong [11] conducted research in the area of tracking people in an multi-camera system. Also in 2001, Khan, Javed, and Shah [12] proposed a method for tracking people and vehicles with uncalibrated cameras. The system is able to discover spatial relationships between the FOVs of the three cameras used. All three works rely on Bayesian classification and networks [13].

The methods even demonstrated the feasibility of tracking people in real-time, but are in general very limited in their capabilities. For example the work of Chang and Gong is limited to people in upright pose. The algorithm proposed by Cai and Aggarwal lacks robustness compared to the single-camera tracking and Khan, Javed, and Shah approach does not calibrate the cameras correctly and is highly susceptible to errors caused by occlusion. But in the past two decades the field of tracking in multi-camera systems has evolved significantly.

3.2 Milestones

This section highlights significant milestones that have shaped the MTMCT research domain, focusing on the five critical areas: detection, feature extraction, data association, tracking, and datasets (challenges).

3.2.1 Detection

The foundation for modern object detection methods was laid in 1998 by Lecun, Bottou, Bengio, *et al.* with the development of Convolutional Neural Networks (CNNs), which are deep learning models specifically designed to process images [14]. The advent of deep learning in the past quarter-century has led to a significant improvement in object detection performance.

With the introduction of R-CNN [15] in 2014, Girshick, Donahue, Darrell, *et al.* demonstrated that deep learning can be used for object detection. The architecture follows a two-stage process: first, it proposes regions of interest using a selective search and then classifies these regions using CNN features. Due to R-CNN proposing the regions of interest independently, it was computationally intensive. Just one year later improvements were made with Fast R-CNN [16], addressed the inefficiencies of its predecessor by introducing a mechanism to share convolutional computations across region proposals and incorporating a Region of Interest (RoI) pooling layer to extract a fixed-size feature vector from the feature map for each proposal. In 2017 Ren, He, Girshick, *et al.* proposed Faster R-CNN [17], which integrated a Region Proposal Network (RPN) into the architecture that employs anchors, which are predefined reference boxes of various scales and aspect ratios and used as a basis for proposing potential object locations. This allows the generation of region proposals almost cost-free by sharing the convolutional features with the downstream detection network. This end-to-end trainable model marked a significant leap in efficiency and set a new standard for object detection tasks.

Following the success of R-CNN and its successors, the object detection landscape was further revolutionized by the introduction of You-Only-Look-Once (YOLO) [18] and Single Shot MultiBox Detector (SSD) [19], which are designed to be even more efficient and suitable for real-time applications.

The YOLO framework, presented by Redmon, Divvala, Girshick, *et al.* in 2015, revolutionized real-time object detection by predicting bounding boxes and class probabilities directly from full images in just one evaluation. YOLO processes the entire image in a single forward pass through the network, divides the image into a grid, and predicts bounding boxes and probabilities for each grid cell. The strength of YOLO lies in its speed, making it highly suitable for applications where real-time detection is crucial. Even though the original author has stopped working on YOLO, due to ethical concerns, it is still being improved continuously. The latest official version, YOLOv4, was released in 2020 by Bochkovskiy, Wang, and Liao [20].

Liu, Anguelov, Erhan, *et al.* proposed SSD in 2015, another influential single-shot object detector that balances the trade-off between speed and accuracy. Unlike YOLO, SSD operates on multiple feature maps at different resolutions to effectively handle objects of various sizes. The architecture applies a set of convolutional filters to these feature maps to predict both the bounding box offsets and the class probabilities for a fixed set of default bounding boxes, which are distributed over the image. Detecting and tracking objects across different scales and perspectives makes SSD particularly suitable for MTMCT applications.

Table 3.1: Overview Object Detectors

Model	**Speed**	**Accuracy**	**Computational Requirements**
YOLO [18]	Very High	Moderate	Low
Faster R-CNN [17]	Moderate	High	High
SSD [19]	High	High	Moderate

Table 3.1 compares the mentioned prominent object detection models used in MTMCT:

- **Speed:** Refers to the time it takes for the detector to process a single frame, usually measured in frames per second (FPS). High speed is crucial for real-time tracking applications, where it is necessary to process video feeds live or near-live.
- **Accuracy:** Measures the ability to correctly identify and locate objects. It is usually quantified by precision and recall rates, or the average precision (AP) over a dataset.
- **Computational Requirements:** Refers to the resources needed to run the detector, typically measured in terms of the number of floating-point operations (FLOPs) or the memory and processing power required. Efficient use of computational resources is essential for deploying MTMCT systems on hardware with limited capabilities.

3.2.2 Feature Extraction

Early feature extraction techniques relied on hand-crafted descriptors such as Scale-Invariant Feature Transform (SIFT) [21] and Histogram of Oriented Gradients (HOG) [22], which were pivotal in object recognition and re-identification (re-ID) tasks. With the introduction of deep learning, CNNs have enabled the automatic learning of feature representations, greatly enhancing the robustness and power for re-ID [23], [24].

More recently, Siamese networks have emerged as a popular choice for learning discriminative features in a pairwise manner, proving to be highly effective for re-ID tasks [25].

3.2.3 Data Association

Data association in MTMCT involves matching detections of the same object across different frames and camera views, which is essential for maintaining object identity over time. The Hungarian algorithm [26], also known as the Munkres assignment algorithm, has historically been used for optimal assignment in data association, addressing the problem of associating detections to tracks in a globally optimal way.

The complexity of data association increased with the need to handle multiple objects and cameras, giving rise to the development of Joint Probabilistic Data Association Filters (JPDAF) [27] that consider the probabilities of all potential measurement-to-track assignments.

Fleuret, Berclaz, Lengagne, *et al.* introduced the use of Probabilistic Occupancy Map (POM) [28] to model targets into a POM and combine occupancy probabilities with color and motion attributes in the tracking process in 2008. The POM is a ground plane that represents the occupancy probability of each cell with this approach tracking of multiple persons in a complex environment is possible. The POM is still used in recent and state-of-the-art approaches.

The advent of graph-based approaches provided a robust framework for data association, viewing the problem as finding the shortest path in a graph where each node represents a detection and edges represent association costs [29]. This method became especially useful in managing associations over long periods and occlusions.

Recently, with the surge of deep learning, Neural Networks have been employed to learn the data association task, allowing for an end-to-end approach to tracking by directly learning to associate features extracted from raw pixels namely Recurrent Neural Networks (RNNs) [30]. This signifies a shift from traditional methods that require hand-crafted features and heuristics towards data-driven approaches.

The introduction of appearance models using deep learning has significantly improved the association performance in MTMCT by providing discriminative features that can robustly represent an object across different viewpoints and illumination conditions, which are essential for accurate association over multiple cameras [1], [31].

3.2.4 Tracking

The Kalman Filter [32] represents one of the early foundations for object tracking, providing a framework for predicting the future locations of an object.

As tracking scenarios became more complex, approaches like Multiple Hypothesis Tracking (MHT) [33] were developed to manage several potential data association hypotheses, especially in crowded scenes [34].

Graph-based methods are another cornerstone in tracking, framing the tracking task as an optimization problem where the best path in a graph represents the sequence of object detections over time, where the nodes represent detections and and the edges represent the association costs [29].

3.2.5 Datasets and Challenges

Besides the datasets mentioned in section 2.4, which are used for object detection in general, there are also datasets which are more tailored towards MTMCT. Typically revolves around tracking specific object classes, predominantly people and vehicles. Datasets, which fits these requirements are listed in table 3.2.

Table 3.2: Overview of Datasets

Dataset	Environment	Num. of Scenarios	Num. of Cameras (Overlap)	FPS	IDs	Year	Class
MARS[1] [36]	!!!	!!!	!!!	!!!	!!!	2016	Person
MOT16 [37]	Outdoor	14	1	25-30	!!!	2016	Person, Vehicle
DukeMTMC [8]	Outdoor	1	8 (✓)	60	2834	2016	Person
WILDTRACK [38]	Outdoor	1	7 (✓)	60	313	2018	Person
MSMT17 [39]	Mixed	12	15 (✓)	15	4101	2018	Person
CityFlowV1 [40]	Outdoor	5	40 (✓)	10	666	2019	Vehicle
MOT20 [41]	Outdoor	8	1	25	!!!	2020	Person, Vehicle
CityFlowV2 [40]	Outdoor	6	46 (✓)	10	880	2021	Vehicle
MMPTRACK [42]	Indoor	5	23 (✓)	15	!!!	2023	Person
MEVID [43]	Mixed	17	33 (✓)	!!!	158	2023	Person

Table 3.2 provides a summary of various datasets that have significantly contributed to the MTMCT research domain. Each dataset is categorized based on several distinct criteria to reflect its unique characteristics and relevance:

- **Environment**: Setting of data collection, from controlled indoor environments to dynamic outdoor locations.
- **Num. of Scenarios**: Details the number of distinct scenarios or situations represented in the dataset.
- **Num. of Cameras (Overlap)**: Represents the number of cameras involved and indicates if there is an overlap in their views.
- **FPS**: Specifies the frame rate of the dataset, important for real-time processing considerations.
- **IDs**: Enumerates the unique identities present, which can provide a measure of the complexity of the dataset.
- **Year**: States the year of the release, representing the recentness of the dataset.
- **Class**: Identifies the subjects annotated, such as persons or vehicles.

Each dataset listed plays a role in the following sections, the reviewed literature is often evaluated on one or more of these datasets. The datasets are also used to train and test the tracking methods.

In recent years, challenges have been established to encourage research in object detection and tracking, although they have mostly centered on ST-SCT and MT-SCT. Nevertheless, these challenges remain relevant to MTMCT research. The most recent representatives of the primary challenges are:

- **MOT20 Challenge:** Benchmark, which includes crowded environments and variable lighting conditions. Moreover, it provides ground truth data to facilitate evaluation. The MOT datasets are released in conjunction with the MOTChallenge [41].
- **2023 AICity Challenge:** Focuses on AI applications in smart cities and includes multi-object tracking for traffic surveillance and anomaly detection as one of its key components. The CityFlow datasets belong to the AICity Challenges. [44]
- **VOT2022 Challenge (Visual Object Tracking Challenge):** An annual competition that provides a standardized dataset and evaluation framework for single-object tracking. [45]
- **VOTS2023 Challenge (Visual Object Tracking and Segmentation Challenge):** An extension of the VOT Challenge that focuses on multi-object tracking. The challenge, recently published in October 2023, affirms the quickly growing interest in this field. [46]

3.3 Methods

This section reviews the methods and state-of-the-art algorithms used in MTMCT.

3.3.1 Tracking-by-Detection

The most common approach used by MTMCT systems is to first detect the objects in each frame and then data association is performed to link the detections across frames. This Tracking-by-Detection (TbD) implementation as a multi-shot approach and treats detection and association as separate, sequential tasks, allowing for the use of specialized methods tailored for each step.

One of the pioneering works in this domain is the Simple Online and Realtime Tracking (SORT) [47] algorithm proposed by Bewley, Ge, Ott, *et al.* SORT employs a combination of Kalman filters for predicting the motion of objects and the Hungarian algorithm for associating detections over time, based on both predicted locations and detected bounding boxes. Its efficiency and speed make it suitable for real-time applications, though it may struggle with identity switches in crowded scenes due to its reliance on motion cues alone.

Building on the foundation laid by SORT, Wojke, Bewley, and Paulus introduced the DeepSORT [48] algorithm, which enhances the tracking performance by incorporating deep learning techniques for appearance features extraction. DeepSORT extends SORT by adding a neural network that generates a high-dimensional vector representation of the appearance of an object, which can be used to compute similarity scores between detections. This addition significantly improves the robustness of the tracker in scenarios where motion predictions are insufficient, such as occlusions or complex, dynamic environments.

Both SORT and DeepSORT have set benchmarks in the field of object tracking, with the latter demonstrating how the integration of motion and appearance information can lead to improved tracking performance.

- **SORT:** Focuses on speed and simplicity by using motion models for prediction and frame-by-frame data association.
- **DeepSORT:** Improves SORT by adding appearance information into the data association step, thus enhancing tracking accuracy, especially in cases where objects interact closely or are temporarily occluded.

It is important to mention that both tracking frameworks rely on an external object detector to provide bounding box detections, which can be any of the object detection models discussed in section 3.2.1. Also SORT as well as DeepSORT are not

3.3.2 Single-Shot Approaches

In contrast to the TbD implementations, single-shot approaches aim to perform detection and data association simultaneously in a single step. This paradigm, while less common, offers the advantage of speed and simplicity by eliminating the need for separate data association algorithms. Especially in scenarios where computational resources are limited and real-time performance is critical, single-shot approaches can be highly effective.

A notable contribution in this domain is the Single-Shot Multi Object Tracking (SMOT) [49] algorithm proposed by Li, Xiong, Yang, *et al.* in 2020. SMOT is a tracking framework, which is able to convert any single-shot object detector into a multi-object tracker, which is able to simultaneously generate detection and tracking outputs. It is based on work of Bergmann, Meinhardt, and Leal-Taixé [50], who developed a *Tracktor*, an object detector, which is also able to track objects at the same time. The SMOT framework is able to generate tracklets with a almost constant runtime with respect to number of targets, due to the use of a light-weighted linkage algorithm for online tracklet linking.

In the same year Wang, Zheng, Liu, *et al.* published the paper "Towards Real-Time Multi-Object Tracking," which proposes a single deep-network that Jointly learns the Detection and Embedding (JDE) model. Due to reduction of computational cost, the system is able to achieve (near) real-time performance, while being almost as accurate as the models, which are separately trained for detection and embedding. The architecture is based on the Feature Pyramid Network (FPN) [52], which is useful for detecting objects of different sizes. A variation of the triplet loss [31] is used to learn the embedding space, which is used for data association. This variation of the triplet loss is defines as follows:

$$\mathcal{L}_{\text{triplet}} = \sum_i \max\left(0, f^\top f_i^- - f^\top f^+\right) \quad [51, \text{Eq. } 1] \tag{3.1}$$

- $f^\top$: Instance in a mini-batch selected as the anchor
- f^+: Represents a positive instance (same ID as anchor)
- f^-: Represents a negative instance (different ID as anchor)

The triplet loss defined in equation 3.1 is used to learn an embedding space where instances of the same identity are closely mapped to each other while pushing apart the embeddings of dissimilar identities.

An even more recent framework is the FairMOT [53] algorithm proposed by Zhang, Wang, Wang, *et al.* in 2021. It combines the two tasks of object detection and re-ID while addressing the *unfairness* issue in multi-task learning, which arises because re-ID is often treated as a secondary task in existing frameworks and is not given enough attention. The paper raises three key issues with existing multi-task learning frameworks:

1. **Unfairness Caused by Anchors:** Re-ID task is overlooked in the anchor-based detection framework, where the anchors are only optimized for the detection task.
2. **Unfairness Caused by Features:** One-shot trackers share most of their features between the detection and re-ID branches. While detection requires deep features to estimate the object class re-ID requires low-level appearance features to distinguish between different identities, this leads to a conflict between the two tasks.
3. **Unfairness Caused by Feature Dimension:** The features dimension of re-ID features is usually much higher than the detection features, but high-dimensional features notably harm the detection performance.

To jointly train the detection and re-ID branches in the FairMOT network the uncertainty loss proposed by Cipolla, Gal, and Kendall [54] is used. The uncertainty loss is defined as follows:

$$L_{\text{total}} = \frac{1}{2}\left(\frac{1}{e^{w_1}} L_{\text{detection}} + \frac{1}{e^{w_2}} L_{\text{identity}} + w_1 + w_2\right) \quad [53, \text{Eq. } 5] \tag{3.2}$$

The uncertainty loss defined in equation 3.2 is used to jointly train the detection and re-ID tasks by assigning different weights to the two tasks to allow a fair learning process. The weights w_1 and w_2 are used to control the balance between the two tasks and are learned during training. $L_{\text{detection}}$ and L_{identity} are the detection and re-ID losses respectively.

By addressing the three key issues with existing multi-task learning frameworks, the FairMOT framework is able to outperform state-of-the-art methods in terms of both tracking accuracy and speed on the MOT17 dataset.

An important notice is that the term *single-shot* used by those frameworks only refers to the detection and intra-camera tracking, the inter-camera (multi-camera) associations still require an additional separate step.

3.3.3 Tracking-by-Attention

Tracking-by-Attention (TbA) represents a paradigm shift in MTMCT systems by incorporating attention mechanisms that prioritize the most salient features of objects during tracking. The attention paradigm, inspired by the visual ability of humans to focus selectively, has been integrated into tracking frameworks to dynamically emphasize important spatial and temporal features.

3.3.4 Geometrical Approaches

Geometrical approaches are based on the assumption that the cameras are calibrated and the scene is static. The calibration of the cameras is necessary to determine the relative position and orientation

3.3.5 Graph Based Approaches

Graph-based approaches have been widely used in MTMCT, especially for data association. The problem of data association can be formulated as a graph optimization problem, where each node represents a detection and edges represent the association costs. The goal is to find the shortest path in the graph, which represents the sequence of object detections over time. More recently Graph Neural Networks (GNNs) [55] have been employed to learn the data association task, allowing for an end-to-end approach to tracking.

In 2017 Chen, Cao, Chen, *et al.* [56] proposed a pedestrian tracking model, which combines inter- and intra-camera tracking and unifies the two steps into one global graph by considering the initial observations as inputs and directly outputting the final trajectories. Due to the fact that the initial observations contain more information like motion than simple detections, they are more credible for data association. Futhermore, it speeds up computing time, because the number of observations is much smaller than the number of detections. The main focus of this paper is on equalizing the similarity metrics of both tasks to allow unbiased data association. An equalization of metrics is needed, if it is not applied the joint approach would favor objects from the same camera view almost all the time as more similar, because the observations are made under the same circumstances like view angle and illumination. Experimental results show that the proposed joint approach leads to improved performance compared to tackling the association as two independent tasks, especially when the accuracy of intra-camera tracking quality is poor the two step approach is not able to recover at the second step and produces mismatches errors.

Similar to [56] Nguyen, Quach, Duong, *et al.* present a single-stage approach that combines intra- and inter-camera association by reformulating it as a single-global one-to-many assignment problem. With a focus on dynamic (on-the-move) cameras, the method is used in an autonomous vehicle (AV) environment, which is not the focus of this project, but still an interesting concept and worth mentioning. The proposed method is called Fractional Optimal Transport Assignment (FOTA) [57] and can be used in both the tracking-by-detection and tracking-by-attention paradigms. The architecture consists of an encoder, two decoders and a box-matching layer. The encoder extracts features of the current and previous frames from the cameras and encodes the feature maps into keys that are used by the decoders to detect and track object boxes. The box-matching layer is then used to match the boxes and provide the final tracking results. The FOTA method results in a reduction of ID switch errors in a large AV dataset compared to state-of-the-art methods.

The Dynamic Graph Model with Link Prediction (DyGLIP) [58] approach proposed by Quach, Nguyen, Le, *et al.* in 2021 is a graph model that uses link prediction to solve the data association problem. It works for both overlapping and non-overlapping cameras and is tested on both person and vehicle tracking. The main advantage are better feature representations and the ability to recover from lost tracks during camera transitions. DyGLIP combines link prediction in conjunction with a dynamic graph formulation that takes temporal information of an object into account for the first time in MTMCT. Based on this approach Cheng, Qiu, Chiang, *et al.* propose a Reconfigurable Spatial-Temporal Graph Model (ReST) [59] in 2023, that handles data association in two steps. First spatial association matches objects across different views at the same frames. Before the second step, a graph reconfiguration module simplifies and reconfigures the graph. Then, temporal association uses information such as speed and time to build a temporal graph and match objects across different frames. Unlike traditional approaches ReST does not rely on single-camera tracking results, because it directly matches objects across camera views in the first step. Another advantage is that two graph models can be trained separately, so there is no need to compromise between the two tasks of intra- and inter-camera data association. The ReST model achieves state-of-the-art performance on the Wildtrack dataset.

The graph based soccer player tracker published by Komorowski and Kurzejamski [60] directly uses raw detection heat maps of the feet of the players instead of bounding boxes. The feet of the players are detected by the pre-trained detector FootAndBall [61], the detection heat maps from all cameras are transformed onto a bird's eye view plane and stacked together to form a multi-channel tensor. This leads to extraction and aggregation being performed within the tracking network itself instead of using a separate preprocessing step like common approaches do, therefore this approach is called *tracking-by-regression*. The tracking network consists out of a Long Short-Term Memory-based (LSTM) [62] RNN that models the player dynamics and a GNN that is able to learn the interaction between players. The training data is synthetically generated by the Google Research Football Environment (GRF) [63] and the final tracker is compared with a baseline approach, base on a particle filter. Even though the proposed tracker is not able to use visual cues like jersey numbers due to a large distance to the camera, it achieves better accuracy and a lower number of ID switches compared to the baseline approach.

3.3.6 Attention Models and Transformers

3.3.7 Edge Computing

The term *edge computing* refers to the concept of processing data near the source of the data, which is in contrast to the traditional approach of processing data in a centralized cloud. The advantages of edge computing are low latency, reduced bandwidth, and improved security. The major disadvantage is the limited computational resources of edge devices in this case the cameras themselves.

In the paper of the already discussed single-shot approach SMOT [49] it is mentioned that replacing the components of the SMOT framework with faster versions can achieve real-time performance on less powerful machines like edge devices.

3.3.8 Online and Real-Time

In addition to subsection 3.3.2, which deals with single-shot approaches and their relevance for real-time applications, this section focuses on online and real-time implementations, mentioning certain methods.

Unlike most of the methods used in MTMCT, the real-time system Uni-ID [64] follows a distributed concept to ensure that the communication and computing costs of each camera in the network remain almost constant as the number of cameras increases. Therefore, smart stations are installed on the tracked roadside and connected by a wireless multi-hop network. YOLO is used for detection and DeepSORT for tracking. First, intra-camera tracking and feature extraction is performed to assign a local ID to each object. Second, the local ID, features and track information of the target are sent to the adjacent node in the network. Third, the adjacent node performs inter-camera tracking to assign a global ID to the target. The system is tested with three nodes and achieves real-time performance with a relatively low performance GPU for each node.

The work of Wang, Liao, Hsieh, *et al.* [65] focuses on the less attention-grabbing use of fisheye cameras to simulate a checkout-free store, where each person enters or exits the store by scanning a QR code that initializes and terminates the tracking process. Compared to perspective cameras, fisheye cameras are able to cover a larger area with a single camera, reducing the number of cameras needed in the system. In addition, fisheye cameras are less susceptible to occlusion when mounted on a ceiling (top-view). Once a camera is calibrated, the POM of the scene can be created to determine the likelihood of a person being in a particular area and to match the tracks of the same person across different cameras. In a scenario with 5 fisheye cameras and 5 to 10 people in a scene simultaneously, the system achieves real-time performance of about 10 FPS without GPU support.

Tesfaye, Zemene, Prati, *et al.* propose the use of Fast-Constrained Dominant Set Clustering (FCDSC) [4] to solve both intra- and inter-camera simultaneously. The method is orders of magnitudes faster than existing graph-based methods due to instead of considering the whole graph only a sub-graph is considered. The proposed method follows a three-layer hierarchical approach. The first two layers solve the intra-camera tracking and the third layer the inter-camera tracking while merging the tracks of the same person across camera views. The tracking algorithm runs at 18 FPS and is 2000 times fast than CDSC [66] which it is based on.

3.3.9 Further Approaches

3.3.10 State-of-the-Art Approaches

This subsection presents state-of-the-art approaches, which were published in 2022 and 2023. The approaches achieve state-of-the-art performance on MTMCT datasets but are not real-time capable, due to the use of computationally expensive methods.

Lifted Multicut Meets Geometry Projections (LMGP) [67] proposed by Nguyen, Henschel, Rosenhahn, *et al.* follows the traditional TbD paradigm, but with the use of POM for each node in the tracking graph, it integrates concepts from centralized representation methods. A pre-clustering step refines tracklets generated by intra-camera tracking to reduce ID switch errors. For the pre-clustering step the bottom edge center of each bounding box is projected to obtain the 3D coordinates. If the Euclidean distance between two projected ground points is less than a diameter of a person, the two detections may belong to the same person. While solving a global lifted multicut formulation the model takes into account short- and long-range temporal interactions to perform inter-camera matching. Intra-camera tracking is performed by CenterTrack [69] and embedding vectors are extracted by DG-Net [70]. LMGP achieves near perfect state-of-the-art performance on the Wildtrack dataset.

EarlyBird [68] proposes an early-fusion in the bird's eye view (BEV) that means detections are directly performed in the BEV to solve spatial association of pedestrians across cameras. The approach is built on MVDeTr [71] and brings the concept of joint detection and re-ID extraction from FairMOT to MTMCT. The input frames are augmented and fed to a encoder network, the image features are projected to the ground plane and aggregated to receive BEV features (in the BEV space). Finally detections and their corresponding re-ID features are fed through a decoder network to association the detections. The proposed approach is similar to ReST in the sense that it associates spatially on the ground plane but it has the advantage of projecting the complete feature space to the ground plane and associating it with the decoder network. EarlyBird shows that early fusion in the BEV space is able to outperform late fusion in the image space. The disadvantage is a higher computational cost due to simultaneously projecting full images of all camera views to the ground plane. Furthermore, high-quality 3D annotations are required which is costly and rare for real-world data.

Huang, Yang, Jiang, *et al.* [72] achieve the first-place ranking in the AI City Challenge 2023 (Track1) with their anchor-guided clustering approach for inter-camera re-ID enabled by self-camera calibrations to improve tracking accuracy of people with similar appearances. Three steps are performed to achieve the final tracking results. First, intra-camera tracking is performed with BoT-SORT [73] following a standard TbD scheme. Second, the anchor guided clustering step fixes ID switches and assigns a global ID to each trajectory by hierarchically clustering appearance features from each camera view and obtaining anchors. Each anchor contains features that represent the appearance of the same identity under different conditions. Third, human pose with camera self-calibration is utilized to project the tracked objects on a top-down map.

3.4 Strengths and Weaknesses

This section discusses the strengths and weaknesses of the reviewed methods and algorithms.

4 Discussion

4.1 Comparison of Methods

Compare and contrast the different methods reviewed in the previous chapters.

4.2 Gaps and Limitations

Identify the gaps and limitations in current research.

4.3 Future Research

Suggest areas for future research.

4.4 Ethical and Privacy Concerns

Discuss the ethical and privacy concerns related to MTMCT and the need for regulations and guidelines.

5 Conclusion

5.1 Summary

Summarize the main points made in your paper.

Highlight the importance of online and real-time tracking in MTMCT and its potential to revolutionize various applications.

5.2 Future Directions

Conclude by stating the future directions and challenges for research in this area.

6 Structure

6.1 Citations

6.1.1 General

[3]: Current Trends in MCMOT. State of the Art. A lot of basic and advanced knowledge. Good for introduction. Analyzes 30 MCT algorithms.

[2]: General description of multi-camera tracking. State of the Art, Markov Process, graph partition theory, tracking by joint constraints.

[1]: Person re-identification past, present, future.

[12]: Tracking people in multiple uncalibrated cameras. Discover spatial relationships between the camera FOVs. Tested on PETS 2001.

6.1.2 Beginning

[10]: First approaches of tracking humans in multi camera network. Already done in 1999 with real-time tracking. Automatic camera switching. Bayesian classification schema.

[11]: Bayesian modality fusion to track multiple people in an indoor environment. Tries to fix already known occlusion problem.

6.1.3 Real-time

[17]: Faster R-CNN. Towards Real-Time Object Detection. Region Proposal Network (RPN). RPN is trained end-to-end. Attention mechanism. 5-17 fps on GPU. Two modules (first region proposal, second detector). Sharing convolutional features.

[51]: Toward Real-Time. Only multi-object tracking. Introduces JDE (Joint learning of detection and embedding). Very important paper (first real-time MOT system). Single-shot detector

[65]: Indoor scene, multiple top-view **fisheye** cameras. Possible to cover large space, less occlusion among objects. People detection and tracking. Calibrate cameras, real time (FPS of about 10) without GPU support.

[64]: Real-time distributed MCMOT system. City-scale scenario. Keeping communication and computing costs of each device low. Installs smart stations on the roadside and connects them to maintain communication. Decentralized Tracking. Kalman filter and hungarian algorithm. YoloX and DeepSORT.

[53]: FairMOT, one-shot tracker (anchor-free style). Tackles issue of object detection against re-ID. Re-ID often threated as secondary task. Reasons behind failure: anchors, feature sharing, feature dimension.

[4]: Multiple non-overlapping cameras using fast-constrained dominant set clustering (FCDSC). Three-layer hierarchical approach. Orders of magnitudes faster than existing methods. Can be used in conjunction with re-id algorithms. Good graphics in paper.

6.1.4 VOT

[74]: VOT21 Challenge Results. Considers single-camera, single-target, model-free tracking. VOT-RT2021 focuses on real-time RGB tracking. Requires predicting bounding boxes. Top two trackers: TrasT_M and STARK_RT.

[45]: VOT22 Challenge Results. Considers single.camera, single-target. VOT-RT2022 focuses on real-time RGB tracking, VOT-RTs by segmentation, VOT-RTb by bounding boxes. Goes beyond previous challenges (updating datasets). Real-time tracking at 20fps. Top trackers: MS_AOT and OSTrackSTB.

[46] VOTS23 Challenge Results. First year considering multiple-target tracking challenge. Explores short- and long-term at once. Only one challenge for all. Does not distinguish between these scenarios. Success is measured in IoU, tracking Quality mathbfQ, Accuracy, Robustness, NRE, DRE, ADQ. Dataset with challenging situations, wide range and diverse set of objects, object which are a part of other objects. Also longer videos. 77 trackers submitted, 47 valid. Most trackers applied uniform dynamic model, utilized transformers, general segmentation network SAM. Top tracker: DMAOT built upon VOT22 winner AOT. Best segmentation-based trackers outperformed all bound.box trackers.

6.1.5 Dynamic Cameras

[75]: Tracking multiple vehicles in the front view of an onboard monocular camera. Siamese network with a spatial pyramid pooling. Markov decision process. Effective for real-time long-term tracking. Hungarian algorithm, reinforcement learning.

[57]: Single-Stage Global Association Approach. Dynamic MCMOT (moving cameras in vehicle). Solves fragment-tracking issues. Not relevant for static MCMOT.

6.1.6 Person Tracking

[76]: Integrating social grouping behavior for tracking pedestrians. Online learned conditional random field (CRF). Non-overlapping cameras.

[77]: Non-overlapping cameras. Pedestrian Tracking. Fix ID-switching issues with long-term feature extraction. OC-SORT + feature extraction.

[60]: Soccer Players. Raw detection heat maps. Google Research Football Environment. Multi camera, multi targets. Cameras have fixed positions. Do not use bounding boxes, instead raw input with heat maps. Graph Neural Network. No visual cues, such as jersey numbers. Player movement trajectories and interaction between neighborhood players.

[78]: Optical-based Pose Association (OPA). Online data association algorithm. Solve the occlusion problem. Take also human pose (see [79]) and optical flow into account, not only visual and spatial information. OpenPose, Object Keypoint Similarity, PWC-Net, Kunh-Munkras algorithm.

6.1.7 Vehicle Tracking (AI City)

[80]: Multi-camera vehicle tracking. No real-time tracking. Improve single-camera tracklets. 4th place in 2022 AI City Challenge. Track refinement module. Yolov5 pre-trained on COCO. Using GAN to generate synthetic data. Background filtering. Hierarchical clustering, zones, two rounds of clustering (tracklets separately each possible transition between cameras, akk tracks fro adjacent cameras).

[81]: Inspired [80]. First place in 2021 AI City Challenge. Yolov5 pre-trained on COCO. Most important: Introduces two step clustering (inter-zone, inter-camera clustering).

[82]: Fourth place in 2021 AI City Challenge (Track 3). Occlusion-aware tracking system. Inspired by Stadler.

[83]: Second place in 2022 AI City Challenge (Track 1). No new innovations made on first glance.

[84]: First place in 2020 AI City Challenge (Track 3). Electricity. Efficient vehicle tracking system. Aggregation loss and fast multi-target cross-camera tracking strategy. Weighted inter-class non-maximum suppression.

[85]: Graph Auto-Encoder and Self-Supervised Camera Link Model. First implementation of GAE in MTMCT. Very interesting paper. Network topology is learned automatically.

6.1.8 Re-ID, Data Association and Tracklet Matching

[86]: Unsupervised cross-dataset transfer learning for person re-id. Unsupervised multi-task dictionary learning (UMDL) model. Uses latent attributes. Asymmetric multi-task learning approach.

[87]: First time use of hierarchical clustering for person re-id. No online method (needs neighboring frames).

[88]: Online-learning-based person re-id. Fully unsupervised learning method. Systematically builds camera link model. Two-way GMM fitting. Multi-kernel adaptive segmentation. Multi-shot framework.

[89]: Orientation-driven person re-id (ODPR). Leverages the orientation cuest and stable torso features to learn a discriminative representation. Also estimates camera topology.Entry/Exit zones are clustered with GMM.

[90]: Locality aware appearance metric (LAAM). Intra- and inter-camera metric for re-ID. Can be applied on top of globally learned re-ID features. Improves tracking accuracy.

[79]: State-aware Re-ID. Human pose information is adopted to infer the target state including occlusion status and orientation. State-of-the-art result on Duke-MTMCT.

[91]: Proposes Mutual Information Temporal Weight Aggregated Person Re-ID Model (MI-TWA). Person re-identification. New algorithm. Not so interesting.

[58]: Dynamic Graph Model with Link Prediction. Tackles problem of data association with a dynamic graph model. Better feature representations and able to recover from lost tracks during camera transitions. Works for person and vehicle tracking for overlapping and non-overlapping cameras. First time link prediction and dynamic graph are used together for MCMOT. Attention models.

[92]: Metadata-Aided Re-ID. Uses metadata information (car type, brand and color) for re-ID. Traffic-aware single-camera tracking. trajectory-based camera link model. Not so interesting.

[93]: Tracklet-to-Target Assignment. Solves cross-camera tracklet matching problem by TRACTA. Proposes the Restricted Non-negative Matrix Factorization (RNMF) algorithm. Estimates the number of targets in the whole network. Important paper.

6.1.9 Datasets

[8]: Largest annotated calibrated data set for MTMC (DukeMTMC).

[94]: Created MTMCT dataset in GTA V. No privacy issues. 6 cameras over 100 minutes per camera. Largest synthetic dataset for multi camera multi person tracking.

[95]: HOTA as new tracking perfomance measure

6.1.10 Misc

[96]: Tracking framework for multiple interacting targets both overlapping and non-overlapping cameras, raw target trajectory with group state. SVMS, homography-based voting schema, networkflow problem, K-shortest paths algorithm.

[97]: Non-overlapping multiple cameras tracking based on similarity function. Data association method. Similarity based on color appearance and camera topology. Use superpixels for extracting color features generated by Simple Linear Iterative Clustering K-means camera topology learning.

[98]: Multiple hypothesis tracking (MHT) for multi-camera tracking. Track hypothesis trees. Disjoint views. Status: tracking, searching, end-of-track. Real-time online method (15 fps). Also uses pose of person.

[67]: Mathematical multi-camera tracking approach. Pre-clustering obtained from 3D geometry projections.

[59]: Utilizes information regarding spatial and temporal consistency. Reconfigurable graph model. Two step approach: Associate all objects across cameras spatially then reconfig into a temporal graph model. Matching object across different views.

[56]: Equalized Global Graph Model-Based Approach. Improved similarity metric for single- and multiple-camera tracking. SCT and ICT in one step.

[99]: Joint person re-id and camera network topology inference. First framework which jointly solves both problems. Minimal prior knowledge about environment. Multi-shot method implemented as random-forest.

[100]: Joint learning of feature, affinity and multi-dimensional assignment (FAMNet). Online MOT. One deep-network for all three tasks. End-to-end learning.

6.2 Approaches

Single vs Multi Camera Tracking

Static vs Dynamic MCMOT

Single-Stage vs Multi-Stage Tracking

Intra camera vs Inter camera tracking

Local and Global tracklets

Cross-camera tracklet matching problem

Graph Neural Networks, Self-Attention, Transformers

Hierarchical Clustering

Gaussian Mixture Models (GMM)

Tracking by detections (Multi-shot) vs One-shot (Single-shot)

Challenges: Occlusion, perspective changes, changes in lighting, changes in appearances, unknown number of targets in the whole network, unknown number of cameras in which a certain target appears.

Common Pipeline:

- Detection
- Feature Extraction
- Single Camera Tracking
- Cross Camera Association
- Multi Camera Tracking

6.3 Die Beschics

Single Object Detection (SOD)

Multi Object Detection (MOD)

Object Re-Identification (ReID)

Single Camera Tracking (SCT)

Multi Camera Tracking (MCT)

Camera Link Model (CLM)

Trajectories and Tracklets

Fisheye vs Normal Cameras

Online vs Offline Tracking (Online: real-time and frame-by-frame, Offline: post-processing)

Local neighborhood: Single-camera tracking: Consecutive frames. Multi-camera tracking: Neighboring cameras.

Different cameras have different technical characteristics.

Appearance features vs Motion features

Datasets:

- DukeMTMC
- MOTChallenge
- AI City Challenge
- PETS
- CityFlow

6.4 Composition

- Introduction
- Motivation
- Technical Background
- Problem Statement
- State of the Art
- Approaches
- Challenges
- Papers
- Further Research
- Conclusion

6.5 Research

6.6 Mentioned Papers

mentioned in [82]:

D. Stadler and J. Beyerer. Improving multiple pedestrian tracking by track management and occlusion handling. In IEEE Conf. Comput. Vis. Pattern Recog., 2021.

mentioned in [98]:

Ristani, E., Tomasi, C.: Tracking multiple people online and in real time. Proc. Asian Conf. Computer Vision, Singapore, 2014, pp. 444-459

Wei, S.-E., Ramakrishna, V., Kanade, T., et al.: Convolutional pose machines. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Las Vegas, USA, 2016, pp. 4724-4732

mentioned in [87]:

Kuhn, H. W. 2010. The hungarian method for the as- signment problem. In 50 Years of Integer Programming.

Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; and Sun, J. 2017. Aligne-dreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184.

Zhong, Z.; Zheng, L.; Cao, D.; and Li, S. 2017. Re- ranking person re-identification with k-reciprocal encod- ing. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 3652-3661.

mentioned in [56]:

S. Yu, Y. Yang, and A. Hauptmann, “Harry Potters Marauders Map: Localizing and tracking multiple persons-of-interest by nonnegative dis- cretization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2013, pp. 3714-3720.

mentioned in [76]:

X. Chen, K. Huang, and T. Tan, “Object tracking across non-overlapping views by learning inter-camera transfer models,” Pattern Recognit., vol. 47, no. 3, pp. 1126-1137, 2014.

E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” IEEE Comput. Graph. Appl., vol. 21, no. 5, pp. 34-41, Sep./Oct. 2001.

M. Moussaiid, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, The walking behaviour of pedestrian social groups and its impact on crowd dynamics

W. Ge, R. T. Collins, and R. B. Ruback, “Vision-based analysis of small groups in pedestrian crowds,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 5, pp. 1003-1016, May 2012.

D. Helbing and P. Molnar, Social force model for pedestrian dynamics, Phys. Rev. E, vol. 51, pp. 4282-4286, May 1995.

6.6.1 Arising Questions

Online Tracking?

Hungarian algorithm?

Multi Object vs Multi Target (definitions)

Attention mechanisms

Detection Frameworks:

- YOLO
- Faster R-CNN
- R-CNN

Tracking Frameworks:

- OpenCV
- DeepSORT
- SORT
- MOTSA

Questions: - Zahlen ausschreiben oder numerisch? - Zitieren bei den Grafiken und Equations gut? - Ganze Grafiken übernehmen?

List of Figures

List of Tables

Bibliography

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016. arXiv: 1610.02984. [Online]. Available: http://arxiv.org/abs/1610.02984.

[2] W. Tian, "Novel aggregated solutions for robust visual tracking in traffic scenarios," Ph.D. dissertation, Karlsruher Institut für Technologie (KIT), 2019, 146 pp., ISBN: 978-3-7315-0915-8. DOI: 10.5445/KSP/1000091919.

[3] T. I. Amosa, P. Sebastian, L. I. Izhar, *et al.*, "Multi-camera multi-object tracking: A review of current trends and future advances," *Neurocomputing*, vol. 552, p. 126 558, 2023, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2023.126558. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223006811.

[4] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah, "Multi-target tracking in multiple non-overlapping cameras using Fast-Constrained dominant sets," *International Journal of Computer Vision*, vol. 127, no. 9, pp. 1303–1320, Sep. 2019.

[5] T. Lin, M. Maire, S. J. Belongie, *et al.*, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. arXiv: 1405.0312. [Online]. Available: http://arxiv.org/abs/1405.0312.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[7] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP journal on image and video processing*, vol. 2, Art.Nr.: 246309, 2008, ISSN: 1687-5176, 1687-5281. DOI: 10.1155/2008/246309.

[8] 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 *European conference on computer vision*, Springer, 2016, pp. 17–35. arXiv: 1609.01775 [cs.CV].

[9] 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)*, vol. 1, Jun. 2006, pp. 951–958. DOI: 10.1109/CVPR.2006.312.

[10] Q. Cai and J. Aggarwal, "Tracking human motion in structured environments using a distributed-camera system," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 11, pp. 1241–1247, Nov. 1999, ISSN: 1939-3539. DOI: 10.1109/34.809119.

[11] T.-H. Chang and S. Gong, "Tracking multiple people with a multi-camera system," in *Proceedings 2001 IEEE Workshop on Multi-Object Tracking*, Jul. 2001, pp. 19–26. DOI: 10.1109/MOT.2001.937977.

[12] 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*, IEEE Computer Society Press Los Alamitos, vol. 5, 2001.

[13] J. Pearl, "Probabilistic reasoning in intelligent systems (chapters 1-3)," in *Probabilistic Reasoning in Intelligent Systems*, J. Pearl, Ed., San Francisco (CA): Morgan Kaufmann, 1988, pp. 1–141, ISBN: 978-0-08-051489-5. DOI: https://doi.org/10.1016/B978-0-08-051489-5.50007-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080514895500072.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, ISSN: 1558-2256. DOI: 10.1109/5.726791.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 580–587. DOI: 10.1109/CVPR.2014.81.

[16] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1440–1448. DOI: 10.1109/ICCV.2015.169.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2016.2577031.

[18] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015. arXiv: 1506.02640. [Online]. Available: http://arxiv.org/abs/1506.02640.

[19] W. Liu, D. Anguelov, D. Erhan, *et al.*, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015. arXiv: 1512.02325. [Online]. Available: http://arxiv.org/abs/1512.02325.

[20] A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020. arXiv: 2004.10934. [Online]. Available: https://arxiv.org/abs/2004.10934.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004, ISSN: 1573-1405. DOI: `10.1023/B:VISI.0000029664.99615.94`. [Online]. Available: `https://doi.org/10.1023/B:VISI.0000029664.99615.94`.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, Jun. 2005, 886–893 vol. 1. DOI: `10.1109/CVPR.2005.177`.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012. [Online]. Available: `https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. DOI: `10.1109/CVPR.2016.90`.

[25] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," *CoRR*, vol. abs/1607.08378, 2016. arXiv: `1607.08378`. [Online]. Available: `http://arxiv.org/abs/1607.08378`.

[26] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[27] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE Journal of Oceanic Engineering*, vol. 8, no. 3, pp. 173–184, Jun. 1983, ISSN: 1558-1691. DOI: `10.1109/JOE.1983.1145560`.

[28] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, Feb. 2008, ISSN: 1939-3539. DOI: `10.1109/TPAMI.2007.1174`.

[29] 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*, Jun. 2008, pp. 1–8. DOI: `10.1109/CVPR.2008.4587584`.

[30] A. Milan, S. H. Rezatofighi, A. R. Dick, K. Schindler, and I. D. Reid, "Online multi-target tracking using recurrent neural networks," *CoRR*, vol. abs/1604.03635, 2016. arXiv: `1604.03635`. [Online]. Available: `http://arxiv.org/abs/1604.03635`.

[31] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 815–823. DOI: `10.1109/CVPR.2015.7298682`.

[32] R. E. Kalman *et al.*, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45,

[33] S. Blackman, "Multiple hypothesis tracking for multiple target tracking," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 1, pp. 5–18, Jan. 2004, ISSN: 1557-959X. DOI: `10.1109/MAES.2004.1263228`.

[34] D. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, Dec. 1979, ISSN: 1558-2523. DOI: `10.1109/TAC.1979.1102177`.

[35] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1116–1124. DOI: `10.1109/ICCV.2015.133`.

[36] L. Zheng, Z. Bie, Y. Sun, *et al.*, "Mars: A video benchmark for large-scale person re-identification," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 868–884, ISBN: 978-3-319-46466-4.

[37] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831 [cs]*, Mar. 2016, arXiv: 1603.00831. [Online]. Available: `http://arxiv.org/abs/1603.00831`.

[38] T. Chavdarova, P. Baqué, S. Bouquet, *et al.*, "Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5030–5039. DOI: `10.1109/CVPR.2018.00528`.

[39] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 79–88. DOI: `10.1109/CVPR.2018.00016`.

[40] Z. Tang, M. Naphade, M.-Y. Liu, *et al.*, "Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 8789–8798. DOI: `10.1109/CVPR.2019.00900`.

[41] P. Dendorfer, H. Rezatofighi, A. Milan, *et al.*, "Mot20: A benchmark for multi object tracking in crowded scenes," *arXiv:2003.09003[cs]*, Mar. 2020, arXiv: 2003.09003. [Online]. Available: `http://arxiv.org/abs/1906.04567`.

[42] X. Han, Q. You, C. Wang, *et al.*, "Mmptrack: Large-scale densely annotated multi-camera multiple people tracking benchmark," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 4849–4858. DOI: `10.1109/WACV56688.2023.00484`.

[43] D. Davila, D. Du, B. Lewis, *et al.*, “Mevid: Multi-view extended videos with identities for video person re-identification,” in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 1634–1643. DOI: `10.1109/WACV56688.2023.00168`.

[44] M. Naphade, S. Wang, D. C. Anastasiu, *et al.*, “The 7th ai city challenge,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2023.

[45] M. Kristan, A. Leonardis, J. Matas, *et al.*, “The tenth visual object tracking vot2022 challenge results,” in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds., Cham: Springer Nature Switzerland, 2023, pp. 431–460, ISBN: 978-3-031-25085-9.

[46] M. Kristan, J. Matas, M. Danelljan, *et al.*, “The first visual object tracking segmentation vots2023 challenge results,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct. 2023, pp. 1796–1818.

[47] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 3464–3468. DOI: `10.1109/ICIP.2016.7533003`.

[48] N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” in *2017 IEEE International Conference on Image Processing (ICIP)*, Sep. 2017, pp. 3645–3649. DOI: `10.1109/ICIP.2017.8296962`.

[49] W. Li, Y. Xiong, S. Yang, S. Deng, and W. Xia, “SMOT: single-shot multi object tracking,” *CoRR*, vol. abs/2010.16031, 2020. arXiv: `2010.16031`. [Online]. Available: `https://arxiv.org/abs/2010.16031`.

[50] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, “Tracking without bells and whistles,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 941–951. DOI: `10.1109/ICCV.2019.00103`.

[51] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, “Towards real-time multi-object tracking,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 107–122, ISBN: 978-3-030-58621-8.

[52] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 936–944. DOI: `10.1109/CVPR.2017.106`.

[53] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, “Fairmot: On the fairness of detection and re-identification in multiple object tracking,” *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3069–3087, Nov. 1, 2021, ISSN: 1573-1405. DOI: `10.1007/s11263-021-01513-4`. [Online]. Available: `https://doi.org/10.1007/s11263-021-01513-4`.

[54] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7482–7491. DOI: `10.1109/CVPR.2018.00781`.

[55] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, Jan. 2009, ISSN: 1941-0093. DOI: `10.1109/TNN.2008.2005605`.

[56] W. Chen, L. Cao, X. Chen, and K. Huang, "An equalized global graph model-based approach for multicamera object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 11, pp. 2367–2381, Nov. 2017, ISSN: 1558-2205. DOI: `10.1109/TCSVT.2016.2589619`.

[57] 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," 2022. arXiv: `2211.09663 [cs.CV]`.

[58] K. G. Quach, P. Nguyen, H. Le, *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)*, Jun. 2021, pp. 13 779–13 788. DOI: `10.1109/CVPR46437.2021.01357`.

[59] C.-C. Cheng, M.-X. Qiu, C.-K. Chiang, and S.-H. Lai, "Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 051–10 060. arXiv: `2308.13229 [cs.CV]`.

[60] J. Komorowski and G. Kurzejamski, "Graph-based multi-camera soccer player tracker," in *2022 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2022, pp. 1–8. DOI: `10.1109/IJCNN55064.2022.9892562`.

[61] J. Komorowski, G. Kurzejamski, and G. Sarwas, "Footandball: Integrated player and ball detector," *CoRR*, vol. abs/1912.05445, 2019. arXiv: `1912.05445`. [Online]. Available: `http://arxiv.org/abs/1912.05445`.

[62] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

[63] K. Kurach, A. Raichuk, P. Stanczyk, *et al.*, "Google research football: A novel reinforcement learning environment," *CoRR*, vol. abs/1907.11180, 2019. arXiv: `1907.11180`. [Online]. Available: `http://arxiv.org/abs/1907.11180`.

[64] Y. Chen, L. Ma, S. Liu, M. Liu, C. Wu, and M. Li, "A real-time distributed multi-camera multi-object tracking system," in *2022 2nd International Conference on Electrical Engineering and Mechatronics Technology (ICEEMT)*, Jul. 2022, pp. 146–149. DOI: `10.1109/ICEEMT56362.2022.9862731`.

[65] T. Wang, C.-H. Liao, L.-H. Hsieh, A. W. Tsui, and H.-C. Huang, "People detection and tracking using a fisheye camera network," in *2021 International Conference on Visual Communications and Image Processing (VCIP)*, Dec. 2021, pp. 1–5. DOI: 10.1109/VCIP53242.2021.9675451.

[66] E. Zemene and M. Pelillo, "Interactive image segmentation using constrained dominant sets," *CoRR*, vol. abs/1608.00641, 2016. arXiv: 1608.00641. [Online]. Available: http://arxiv.org/abs/1608.00641.

[67] D. M. H. Nguyen, R. Henschel, B. Rosenhahn, D. Sonntag, and P. Swoboda, "Lmgp: Lifted multicut meets geometry projections for multi-camera multi-object tracking," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 8856–8865. DOI: 10.1109/CVPR52688.2022.00866.

[68] T. Teepe, P. Wolters, J. Gilg, F. Herzog, and G. Rigoll, *Earlybird: Early-fusion for multi-view tracking in the bird's eye view*, 2023. arXiv: 2310.13350 [cs.CV].

[69] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," *CoRR*, vol. abs/2004.01177, 2020. arXiv: 2004.01177. [Online]. Available: https://arxiv.org/abs/2004.01177.

[70] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 2133–2142. DOI: 10.1109/CVPR.2019.00224.

[71] Y. Hou and L. Zheng, "Multiview detection with shadow transformer (and view-coherent data augmentation)," *CoRR*, vol. abs/2108.05888, 2021. arXiv: 2108.05888. [Online]. Available: https://arxiv.org/abs/2108.05888.

[72] H.-W. Huang, C.-Y. Yang, Z. Jiang, *et al.*, *Enhancing multi-camera people tracking with anchor-guided clustering and spatio-temporal consistency id re-assignment*, 2023. arXiv: 2304.09471 [cs.CV].

[73] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, *Bot-sort: Robust associations multi-pedestrian tracking*, 2022. arXiv: 2206.14651 [cs.CV].

[74] M. Kristan, J. Matas, A. Leonardis, *et al.*, "The ninth visual object tracking vot2021 challenge results," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2021, pp. 2711–2738. DOI: 10.1109/ICCVW54120.2021.00305.

[75] Y. Zou, W. Zhang, W. Weng, and Z. Meng, "Multi-vehicle tracking via real-time detection probes and a markov decision process policy," *Sensors*, vol. 19, no. 6, 2019, ISSN: 1424-8220. DOI: 10.3390/s19061309. [Online]. Available: https://www.mdpi.com/1424-8220/19/6/1309.

[76] X. Chen and B. Bhanu, "Integrating social grouping for multitarget tracking across cameras in a crf model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 11, pp. 2382–2394, Nov. 2017, ISSN: 1558-2205. DOI: 10.1109/TCSVT.2016.2565978.

[77] D.-J. Huang, P.-Y. Chou, B.-Z. Xie, and C.-H. Lin, "Multi-target multi-camera pedestrian tracking system for non-overlapping cameras," in *2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, Jul. 2023, pp. 629–630. DOI: `10.1109/ICCE-Taiwan58799.2023.10227006`.

[78] S. You, H. Yao, and C. Xu, "Multi-target multi-camera tracking with optical-based pose association," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3105–3117, Aug. 2021, ISSN: 1558-2205. DOI: `10.1109/TCSVT.2020.3036467`.

[79] P. Li, J. Zhang, Z. Zhu, Y. Li, L. Jiang, and G. Huang, "State-aware re-identification feature for multi-target multi-camera tracking," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2019, pp. 1506–1516. DOI: `10.1109/CVPRW.2019.00192`.

[80] 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)*, Jun. 2022, pp. 3198–3208. DOI: `10.1109/CVPRW56347.2022.00361`.

[81] C. Liu, Y. Zhang, H. Luo, *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)*, Jun. 2021, pp. 4124–4132. DOI: `10.1109/CVPRW53098.2021.00466`.

[82] A. Specker, D. Stadler, L. Florin, and J. Beyerer, "An occlusion-aware multi-target multi-camera tracking system," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2021, pp. 4168–4177. DOI: `10.1109/CVPRW53098.2021.00471`.

[83] F. Li, Z. Wang, D. Nie, *et al.*, "Multi-camera vehicle tracking system for ai city challenge 2022," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2022, pp. 3264–3272. DOI: `10.1109/CVPRW56347.2022.00369`.

[84] Y. Qian, L. Yu, W. Liu, and A. G. Hauptmann, "Electricity: An efficient multi-camera vehicle tracking system for intelligent city," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2020, pp. 2511–2519. DOI: `10.1109/CVPRW50498.2020.00302`.

[85] H.-M. Hsu, Y. Wang, J. Cai, and 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)*, Jan. 2022, pp. 489–499. DOI: `10.1109/WACVW54805.2022.00055`.

[86] P. Peng, T. Xiang, Y. Wang, *et al.*, "Unsupervised cross-dataset transfer learning for person re-identification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 1306–1315. DOI: `10.1109/CVPR.2016.146`.

[87] Z. Zhang, J. Wu, X. Zhang, and C. Zhang, "Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project," 2017. arXiv: `1712.09531 [cs.CV]`.

[88] Y.-G. Lee, Z. Tang, and J.-N. Hwang, "Online-learning-based human tracking across non-overlapping cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2870–2883, Oct. 2018, ISSN: 1558-2205. DOI: `10.1109/TCSVT.2017.2707399`.

[89] 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*, ser. MM '18, Seoul, Republic of Korea: Association for Computing Machinery, 2018, pp. 1457–1465, ISBN: 9781450356657. DOI: `10.1145/3240508.3240663`. [Online]. Available: `https://doi.org/10.1145/3240508.3240663`.

[90] Y. Hou, L. Zheng, Z. Wang, and S. Wang, "Locality aware appearance metric for multi-target multi-camera tracking," 2019. arXiv: `1911.12037 [cs.CV]`.

[91] J. Li and Y. Piao, "Multi-target multi-camera tracking based on mutual information-temporal weight aggregation person re-identification," in *2022 IEEE 5th International Conference on Electronic Information and Communication Technology (ICEICT)*, Aug. 2022, pp. 149–151. DOI: `10.1109/ICEICT55736.2022.9908659`.

[92] H.-M. Hsu, J. Cai, Y. Wang, J.-N. Hwang, and K.-J. Kim, "Multi-target multi-camera tracking of vehicles using metadata-aided re-id and trajectory-based camera link model," *IEEE Transactions on Image Processing*, vol. 30, pp. 5198–5210, 2021, ISSN: 1941-0042. DOI: `10.1109/TIP.2021.3078124`.

[93] Y. He, X. Wei, X. Hong, W. Shi, and Y. Gong, "Multi-target multi-camera tracking by tracklet-to-target assignment," *IEEE Transactions on Image Processing*, vol. 29, pp. 5191–5205, 2020, ISSN: 1941-0042. DOI: `10.1109/TIP.2020.2980070`.

[94] P. Köhl, A. Specker, A. Schumann, and J. Beyerer, "The mta dataset for multi target multi camera pedestrian tracking by weighted distance aggregation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2020, pp. 4489–4498. DOI: `10.1109/CVPRW50498.2020.00529`.

[95] J. Luiten, A. Osep, P. Dendorfer, *et al.*, "HOTA: A higher order metric for evaluating multi-object tracking," *CoRR*, vol. abs/2009.07736, 2020. arXiv: `2009.07736`. [Online]. Available: `https://arxiv.org/abs/2009.07736`.

[96] S. Zhang, Y. Zhu, and A. Roy-Chowdhury, "Tracking multiple interacting targets in a camera network," *Computer Vision and Image Understanding*, vol. 134, pp. 64–73, 2015, Image Understanding for Real-world Distributed Video Networks, ISSN: 1077-3142. DOI: `https://doi.org/10.1016/j.cviu.2015.`

01.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314215000168.

[97] H. Choi and M. Jeon, "Data association for non-overlapping multi-camera multi-object tracking based on similarity function," in *2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, Oct. 2016, pp. 1–4. DOI: 10.1109/ICCE-Asia.2016.7804834.

[98] K. Yoon, Y.-m. Song, and M. Jeon, "Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views," *IET Image Processing*, vol. 12, no. 7, pp. 1175–1184, 2018. DOI: https://doi.org/10.1049/iet-ipr.2017.1244. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-ipr.2017.1244. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-ipr.2017.1244.

[99] Y.-J. Cho, S.-A. Kim, J.-H. Park, K. Lee, and K.-J. Yoon, "Joint person re-identification and camera network topology inference in multiple cameras," *Computer Vision and Image Understanding*, vol. 180, pp. 34–46, 2019, ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2019.01.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314219300037.

[100] P. Chu and H. Ling, "Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 6171–6180. DOI: 10.1109/ICCV.2019.00627.