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, 2023-11-18

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 All passages taken verbatim or in spirit from the published or unpublished work of others, or from the author's own work, are marked as taken.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 (STSCT) as well as Multi-Target Single-Camera Tracking (MTSCT) 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 STSCT and MTSCT, due to the need to simultaneously track multiple objects across multiple cameras.

Single-Target Multi-Camera 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 this special case.

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: The following sections of this chapter define MTMCT and its importance as well as the objective of this research project and related work. Chapter 2 provides a brief overview of the basics of MTMCT to provide a foundation for understanding the rest of this project. Chapter 3 mentions the previous milestones and reviews the current state-of-the-art in MTMCT, with a special focus on online and real-time tracking methods. Chapter 4 delves into a critical analysis of the methods, challenges, and future prospects in the field of MTMCT. Chapter 5 concludes this project by summarizing the key findings and outlining potential ways for future research.

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 their paths 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 object detection and tracking 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 previous milestones, 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 of Zheng, Yang, and Hauptmann [1] focuses on the past, present and future of person re-identification (re-ID), that is the task of identifying a person across multiple cameras for example if the person leaves and re-enters the field of view of a camera or a person is lost for a short time. The paper covers hand-crafted algorithms as well as deep learning approaches for both image- and video-based re-ID. Furthermore, important datasets are covered, quickly explained and the approaches are evaluated. 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.

Two years later "People tracking in multi-camera systems: A Review - multimedia tools and applications" [2] was published which gives an overview of multi-camera tracking methods. The review covers the most important methods and dataset in the field of tracking people in a multi-camera system. However, it does not cover the task of tracking vehicles and is limited to approaches that were released until 2018.

A chapter of the doctoral thesis of Tian [3, 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 in a single-camera system.

The survey "Deep Learning for Visual Tracking: A Comprehensive Survey" [4] carried out by Marvasti-Zadeh, Cheng, Ghanei-Yakhdan, *et al.* in 2019 delineates the evolution and the state of deep learning-based visual tracking methods, categorizing these methods based on their network architecture, training processes, and learning procedures. It provides a detailed examination of various deep learning architectures and custom networks, each contributing to the efficiency and robustness of visual trackers. It analyzes the challenges faced by deep learning-based trackers and the solutions proposed to address them. The survey also offers a comprehensive comparison of well-known single-object visual tracking datasets, evaluating and analyzing state-of-the-art deep learning-based methods across a range of tracking scenarios. While this survey focus on single-object tracking, their insights into the advancements in deep learning architectures and methodologies provide a valuable context for the study on MTMCT that also adapts and applies mentioned approaches. This survey thus serves as an important source understanding the broader landscape of deep learning applications in visual tracking in general.

The most recent and comprehensive review of MTMCT was published in 2023 by Amosa, Sebastian, Izhar, *et al.* [5]. 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 a main aspect of this project. Futhermore,

this research project aims to provide an eased introduction to the field of MTMCT by first explaining the basics before diving into the details of the latest research.

The integration of edge computing in IoT, as explored in "A Survey on the Edge Computing for the Internet of Things" [6], provides valuable insights for the MTMCT domain. This survey highlights how edge computing significantly reduces latency and balances network traffic, which are critical for real-time data processing in IoT networks. Such capabilities are directly relevant to the challenges faced in MTMCT, especially when dealing with high volumes of data from multiple cameras. The discussion of the paper on distributed computational nodes and their role in supporting real-time analysis and decision-making offers a parallel to the computational needs in MTMCT.

2 Background

This chapter provides an overview of the basic concepts of object detection and tracking and the steps of a MTMCT system, along with a discussion of its key challenges and issues. Also the foundational building blocks of MTMCT are introduced. 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

To understand the data association step the following terms need to be defined:

- **Tracklets:** Short segments of a path of an object captured within the view of a single camera, formed by connecting successive frame detections.
- **Trajectories:** The complete path of an object over time, often across multiple frames and cameras, created by merging together tracklets.
- **Tracks:** Refined trajectories that represent the validated path of an object after correction for inaccuracies and false detections.

In the research literature, these terms are often used interchangeably and not always consistently, but for the purposes of this project, the above definitions are used.

Data association is the process of associating current detected objects with existing trajectories based on similarities in their features. This is done by comparing the features of detected objects with the features of existing trajectories and assigning the detected objects to the most similar trajectories. 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 trajectories are associated across multiple camera views (inter-camera) and finally being optimized globally.

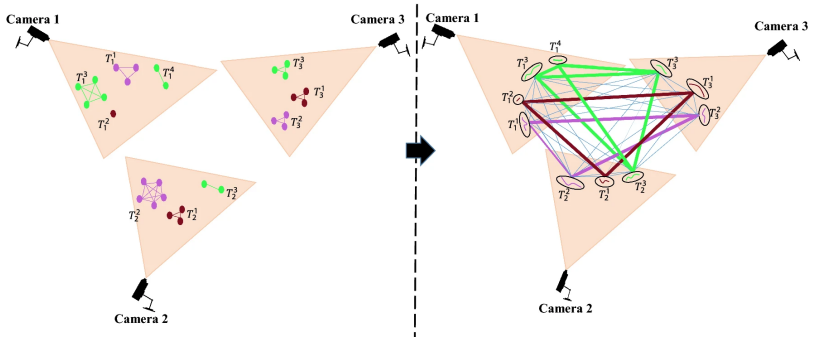

Figure 2.1: Intra- (left) and inter-camera (right) tracking [7, Fig. 1]

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 trajectories are associated independently within the three camera views. The second step is inter-camera tracking, where the trajectories 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 (tracklets), so the next frame of a video, become available. To sum it up tracking is responsible for maintaining and managing the trajectories 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

Single-Target Single-Camera Tracking (STSCT) 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 STSCT 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

Multi-Target Single-Camera Tracking (MTSCT) builds upon the principles of STSCT 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 STSCT to MTSCT, 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 [8].
- **ImageNet:** Vast dataset employed for image classification and object detection. Object detectors trained on ImageNet are able to recognize an broad range of objects [9].

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 [10] in 2008.

$$\text{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \quad [10, \text{Eq. 1}] \tag{2.1}$$

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.

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \quad [10, \text{Eq. 2}] \tag{2.2}$$

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" [11].

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

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 [12] 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.

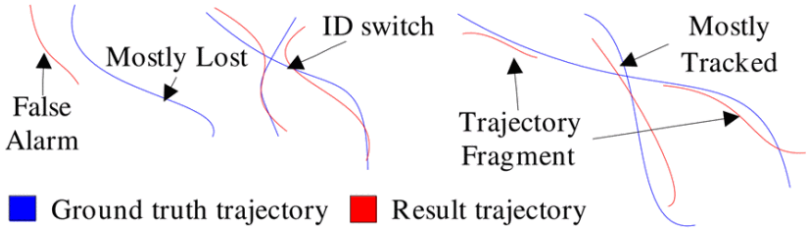

Figure 2.2: MT and ML [12, Fig. 5]

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 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, Section 3.3 reviews the methods and algorithms used in MTMCT.

3.1 The Beginnings

Back in 1999 and 2001 Cai and Aggarwal [13] and Chang and Gong [14] conducted research in the area of tracking people in an multi-camera system. Also in 2001, Khan, Javed, and Shah [15] 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 [16].

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 with 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 [17]. 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 [18] 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 [19], 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 [20], 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) [21] and Single Shot MultiBox Detector (SSD) [22], 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 [23], was released in 2020 by Bochkovskiy, Wang, and Liao. In 2023 Ultralytics released the most recent version YOLOv8 [24], [25].

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 [21]	Very High	Moderate	Low
Faster R-CNN [20]	Moderate	High	High
SSD [22]	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) [26] and Histogram of Oriented Gradients (HOG) [27], 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 [28], [29].

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 [30].

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 [31], 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) [32] that consider the probabilities of all potential measurement-to-track assignments.

Fleuret, Berclaz, Lengagne, *et al.* introduced the use of Probabilistic Occupancy Map (POM) [33] 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 [34]. 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) [35]. 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], [36].

3.2.4 Tracking

The Kalman Filter [37] 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) [38] were developed to manage several potential data association hypotheses, especially in crowded scenes [39].

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
PETS [40]	Outdoor	3	8 (✓)	25	—	2009	Person
MARS[1] [41]	Mixed	Multiple	6 (✓)	—	1261	2016	Person
MOT16 [42]	Outdoor	14	1	25-30	—	2016	Person, Vehicle
DukeMTMC [11]	Outdoor	1	8 (✓)	60	2834	2016	Person
Wildtrack [43]	Outdoor	Multiple	7 (✓)	2	313	2018	Person
MSMT17 [44]	Mixed	12	15 (✓)	15	4101	2018	Person
CityFlowV1 [45]	Outdoor	5	40 (✓)	10	666	2019	Vehicle
MOT20 [46]	Outdoor	8	1	25	—	2020	Person, Vehicle
CityFlowV2 [45]	Outdoor	6	46 (✓)	10	880	2021	Vehicle
MMPTRACK [47]	Indoor	5	23 (✓)	15	—	2023	Person
MEVID [48]	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 STSCT and MTSCT

[1] extension of Market-1501 [49]

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 [46].
- **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. [50]
- **VOT2022 Challenge (Visual Object Tracking Challenge):** An annual competition that provides a standardized dataset and evaluation framework for single-object tracking. [51]
- **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. [52]

3.3 Methods

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

3.3.1 Tracking Paradigms

In the past years various tracking paradigms have been developed, which are used in MTMCT. The most common paradigms are tracking-by-detection, tracking-by-regression, tracking-by-segmentation, and tracking-by-attention. Each of these paradigms has its own advantages and disadvantages, which are discussed in the following sections.

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) [53] 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 [54] 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 able to perform inter-camera tracking, this step has to be performed separately.

Tracking-by-Regression

Tracking-by-Regression (TbR) involves directly estimating the position of the object in each frame of a video sequence. Unlike the TbD approach, regression-based methods predict the changes in position of the object. Most approaches learn a regression function from the input image features to the location coordinates. The advantage of this method lies in its ability to continuously refine the estimated position, making it well-suited for scenarios with smooth movements or predictable trajectory patterns.

Tracking-by-Segmentation

Tracking-by-Segmentation (TbS), on the other hand, focuses on delineating the precise shape of the target object in each frame. This method not only tracks the position of the object but also provides its detailed segmentation, capturing its exact outline and shape. It is particularly useful in complex scenes where the object might change shape, size, or orientation. By combining tracking and segmentation, this approach offers more detailed and accurate object tracking, especially in environments where the distinction between foreground and background is critical.

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.

3.3.2 Single-Shot Approaches

In contrast to all the mentioned tracking paradigms, 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) [55] 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, who developed the Tracktor [56], 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" [57], 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) [58], which is useful

for detecting objects of different sizes. A variation of the triplet loss [36] is used to learn the embedding space, which is used for data association. This variation of the triplet loss is defines as follows:

$$\mathcal{L}_{\text{triplet}} = \sum_i \max\left(0, f^\top f_i^- - f^\top f^+\right) \quad \text{[57, Eq. 1]} \tag{3.1}$$

- $f^\top$: 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 [59] 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:** The 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 [60] is used. The uncertainty loss is defined as follows:

$$L_{\text{total}} = \frac{1}{2}\left(\frac{1}{e^{w_1}} L_{\text{detection}} + \frac{1}{e^{w_2}} L_{\text{identity}} + w_1 + w_2\right) \quad \text{[59, Eq. 5]} \tag{3.2}$$

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 associations still require an additional separate step.

3.3.3 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 [34]. 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) [61] 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.* [62] 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 [62] 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) [63] 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) [64] 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) [65] 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 [66] 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 [67], 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 it is following the TbR paradigm. The tracking network consists out of a Long Short-Term Memory-based (LSTM) [68] 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) [69] 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.

3.3.4 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 due to raw video data not being stored. 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 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.

Wang, Sheng, Zhang, *et al.* introduce a Multi-Camera Multi-Hypothesis Tracking (MC-MHT) framework integrated with a blockchain-based system, Multi-Camera TrackingChain (MCTChain) [70]. This extendable architecture distributes tracking tasks among cameras, improving scalability and security compared to centralized approaches. The architecture consists of three layers: the tracking, blockchain and edge network layer. In the experiment 20 edge cameras are used and the tracking task is performed locally in each camera. Therefore a leader election is implemented in the MCTChain framework to select the camera that is in charge of package transaction. The proposed method achieves real-time performance (24-36 FPS).

Similar to MCTChain the paper "Multi-Camera Vehicle Tracking Using Edge Computing and Low-Power Communication" [71] introduces a decentralized tracking algorithm following the TbD paradigm that on the one hand performs intra-camera tracking locally on the camera and on the other hand uses an ISM-based wireless device-to-device communication for inter-camera tracking.

3.3.5 Online and Real-Time

In addition to subsections 3.3.2 and 3.3.4, which deal with single-shot approaches and edge computing 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 [72] 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 trajectory 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.* [73] 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 trajectories 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) [7] 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 trajectories of the same person across camera views. The tracking algorithm runs at 18 FPS and is 2000 times fast than CDSC [74] which it is based on.

3.3.6 Attention Models and Transformers

Recent advancements in MTMCT have been influenced by the development of attention models and transformers [75] originally applied for natural language processing and conceptualized for enhancing focus in neural networks. Despite their advantages, the high resource demands, particularly in processing power and memory, pose challenges in achieving the low latency required for online and real-time MTMCT tracking applications. Nevertheless, examples of attention models and transformers implementations are at least mentioned in the scope of this project.

The paper “End-to-End Object Detection with Transformers” (DETR) [76] published by Carion, Massa, Synnaeve, *et al.* in 2020 lays the foundational work for using transformers in object detection. It combines a transformer with a set-based global loss, demonstrating significant improvements in accuracy and efficiency. This work paved the way for subsequent transformer-based MOT models.

Expanding the DETR framework, MOTR [77] introduces "track query" to track multiple objects across frames. MOTR updates track queries iteratively, enhancing temporal relation modeling and improving MOT performance. TrackFormer [78] presents an end-to-end trainable model that uses static object queries for new track initialization and autoregressive track queries for existing tracks. TransTrack [79]

introduces a method that simultaneously handles object detection and association. It uses previous frame object features as a query for the current frame, simplifying the tracking process. Furthermore, MotionTrack [80] showcases the application of transformers in an autonomous driving environment with multiple sensor inputs.

The approach Dual Matching Attention Networks (DMAN) [81] consists of both spatial and temporal attention mechanisms. The first generates dual attention maps that allow the network to focus on the matching patterns of the input image pair, while the second adaptively allocates different levels of attention to different samples in the trajectory to suppress noisy observations.

In comparison to the mentioned tracking frameworks that solve the task of intra-camera tracking and are not able to perform inter-camera tracking. MVDeTr [82] and the model of Li, Weng, Xu, *et al.* [83] are able to perform both tasks. MVDeTr focuses on the aggregation of content from multiple camera views. The introduction of the shadow transformer for effective multi-view data fusion is a significant stride towards addressing occlusions and view inconsistencies in MTMCT. [83] leverages transformer-based attention mechanisms for robust person association across different camera views. This approach is instrumental in enhancing the re-ID component of MTMCT, focusing on the challenges posed by uncalibrated and overlapping camera setups.

3.3.7 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.

Hsu, Wang, Cai, *et al.* introduce a Self-supervised Camera Link Model (SCLM) [84] that extracts both appearance and topological features from a Graph Auto-Encoder (GAE) [86] to achieve vehicle tracking in a multi-camera environment. The approach follows the TbD paradigm and advances the Traffic-Aware Single Camera Tracking (TSCT) [87] algorithm, which proposes a zone generation algorithm. After common steps of detecting objects and extracting appearance features, these are used as a node for the GAE to establish the camera link model and generate the tracking results. In addition the intra-camera tracking results are used to generate entry and exit points by using the MeanShift [88] clustering algorithm. The combination of the TSCT and the GAE embeddings with the generation of zones leads to state-of-the-art performance on the CityFlow 2019 and 2020 datasets.

Lifted Multicut Meets Geometry Projections (LMGP) [89] 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 [90] and embedding vectors are extracted by DG-Net [91]. LMGP achieves near perfect state-of-the-art performance on the Wildtrack dataset.

EarlyBird [85] 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 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, Chou, Xie, *et al.* propose a method for non-overlapping multi-camera pedestrian tracking that solves the problem of poor long-term feature storage to allow identifying people correctly even though significant appearance changes like different clothes or lighting conditions occur. The proposed method follows the TbD paradigm and combines an state-of-the-art OC-SORT-based [93] tracker with the person re-ID library Torchreid [94] for feature extraction. The feature extraction is performed as an averaging of the features while only taking frames into account where the person is not occluded nor about to leave or enter the scene. Once a new person enters the scene and has accumulated enough features the cosine distance between the features of the new person and all the people in the area is calculated and the ID is restored if the distance is below a certain threshold, this works for both matching people in the same camera and across camera views. Furthermore, a new dataset including 40000 frames recorded by three cameras is proposed to evaluate the performance of their method. Results show that the combination of OC-SORT, the proposed long-term feature extraction and Torchreid outperforms state-of-the-art methods on the new dataset. Unfortunately, the proposed method is only tested on the new dataset and not on existing datasets which makes it hard to compare the performance in a broader context.

Huang, Yang, Jiang, *et al.* [95] 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 [96] 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.

The "The First Visual Object Tracking Segmentation VOTS2023 Challenge Results" presents the performance of the 47 submitted trackers for the challenge. Most trackers apply a uniform dynamic model and utilize transformers. Multi- as well as single-shot approaches are used, but the top three trackers are single-shot approaches. The top tracker DMAOT is built upon the VOT22 [51] winner AOT [97] and its successor DeAOT [98]. Although a detailed technical documentation on DMAOT is currently not available, it is known that it stores long-term memories object- instead of frame-wise to predict object masks. Overall the challenge reveals a paradigm shift from bounding-box trackers to segmentation-based trackers, which outperform all bounding-box trackers in the challenge.

3.3.8 Honorable Mentions

In the exploration of advanced tracking methodologies, certain studies stand out for their unique and unconventional approaches. This subsection highlights some of these studies, which are not directly related to MTMCT but are still honorable mentions.

One intriguing development in the field of people tracking is "Harry Potter's Marauder's Map: Localizing and Tracking Multiple Persons-of-Interest by Nonnegative Discretization" [99] from 2013. It draws parallels to the fictional Marauder's Map in the Harry Potter series, this research proposes a framework that follows the TbD paradigm and uses nonnegative discretization for robust localization and tracking of persons in complex environments. Their method handles challenges such as occlusions and sparse surveillance camera coverage, employing a semi-supervised learning framework that integrates cues like color, person detection, face recognition, and non-background information. The application in a real-world nursing home setting captured by 15 cameras demonstrates its effectiveness in indoor scenarios.

Equally intriguing is the paper "The MTA Dataset for Multi Target Multi Camera Pedestrian Tracking by Weighted Distance Aggregation" [100], a unique dataset for MTMCT research, which was captured within the virtual environment of the popular video game Grand Theft Auto 5 (GTA). This creative approach leverages the complex, dynamic world of GTA to provide a rich, diverse dataset for tracking research, highlighting the innovative ways in which simulated environments can contribute to advancements in computer vision without the need of touching someones privacy.

4 Discussion

This chapter delves into a critical analysis of the methods, challenges, and future prospects in the field of MTMCT. It compares the mentioned approaches, discusses gaps and limitations of these methodologies, and mentions the ethical implications and future advancements that could revolutionize the field of MTMCT.

4.1 Summary of Methods

Early research in MTMCT, primarily utilized Bayesian classification and network models. These foundational methods were instrumental in kickstart research in the field, though they were constrained by limitations in robustness and susceptibility to errors, particularly in challenging scenarios like occlusions or varied poses.

The advent of CNNs marked a significant shift, introducing deep learning to object detection and substantially elevating performance. Building upon CNNs, R-CNN and its successors, including Fast R-CNN and Faster R-CNN, enhanced efficiency and reduced computational overhead. They introduced more effective region-of-interest handling.

The introduction of real-time object detection frameworks like YOLO and SSD was a game-changer. YOLO, with its single forward pass image processing, revolutionized real-time detection, and the multiple feature maps of SSD effectively addressed the challenge of varying object sizes.

Data association techniques such as the Hungarian Algorithm, JPDAF, and POM have been pivotal in maintaining object identity over time and across different camera views.

Various tracking paradigms emerged such as tracking-by-detection, tracking-by-regression, tracking-by-segmentation, and tracking-by-attention each addressed specific aspects of tracking with their unique advantages. Furthermore, single-shot approaches like SMOT and JDE streamlined the process by integrating intra-camera detection and tracking in a single step, emphasizing speed and simplicity.

The integration of graph-based approaches and neural networks marked another leap forward, offering robust frameworks for data association, especially beneficial for long-term matching and in challenging scenarios with occlusions. Neural networks introduced an end-to-end approach, eliminating the need for hand-crafted features and enabling more efficient data association.

The most recent transformer-based models brought significant improvements in accuracy and efficiency. These models excel in handling multiple objects across frames.

4.2 Gaps and Limitations

While every mentioned approach has its unique advantages, they also have inherent limitations. The following sections discuss the gaps and limitations of these methodologies.

Feature extraction and data association techniques like SIFT, HOG, the Hungarian algorithm, JPDAF and POM are instrumental in the development of MTMCT systems. However, these methods are not able to handle scenarios which constantly growing in complexity with occlusions and varying poses. Furthermore, these methods are computationally expensive, which is a significant limitation in real-time applications.

The performance of the Kalman Filter significantly degrades in the presence of abrupt motion changes or maneuvering targets. On the other hand, MHT, known for its robustness in handling multiple targets and false alarms, faces computational challenges. As the number of targets and hypotheses increases, the increasing computational complexity of MHT makes it less practical for real-time applications in dense environments.

The existing datasets and challenges still mainly focus on intra-camera tracking and there is no challenge yet that focuses on inter-camera tracking within multiple-camera systems. Furthermore, a challenge that focuses solely on real-time tracking is also missing. Both of these challenges would be beneficial to advance the research in these areas, due to developers would be motivated to develop new methods and algorithms to compete in these challenges.

The intense research on intra-camera detection and tracking frameworks like YOLO, Faster R-CNN, SSD and DeepSORT has led to significant advancements in these areas. However, these methods are not optimized for inter-camera tracking, that needs at least one step further or a completely different approach. The lack of a unified framework for inter-camera tracking is a significant gap in the current research.

Single shot approaches like SMOT and JDE have been instrumental in simplifying the tracking process by integrating detection and tracking in a single step and boosting speed and efficiency. However, these methods lack the robustness and accuracy of two-stage approaches, which is a significant limitation in challenging scenarios.

Graph based approaches like DyGLIP and ReST are powerful frameworks for data association, especially in scenarios with occlusions. However graph-based approaches are computationally intensive that is a significant limitation in real-time applications. With the introduction of FCDSC that only considers a sub-graph each step, the

computational complexity of graph-based approaches has been reduced significantly and finally made them feasible for real-time applications.

Attention models and transformers are still in the early stages of development and have not been extensively explored in MTMCT. While these models have shown promising results in other domains, their potential in MTMCT is yet to be fully realized and especially their deployment in real-time applications is still not feasible yet.

Furthermore most approaches lack in the ability of handling objects of different classes and both non- and overlapping camera views. The ability of handling objects of different classes is especially important in scenarios where both people and vehicles are present. The ability of handling non- and overlapping camera views is important in scenarios where the camera setup is not known in advance and the cameras are not calibrated. So there is no *one size fits all* approach yet. The development of such unified framework could also be boosted by the introduction of a challenge that focuses on these aspects.

4.3 Future Research

The advancement of online and real-time methods is critical, considering the increasing demand for instant and accurate tracking in various real-time applications. To achieve significant progress in this area, several research directions need to be explored.

The algorithms should ideally strike a balance between speed and accuracy, providing precise tracking information swiftly. Emphasis on lightweight neural network architectures could lead to models that maintain high accuracy while reducing computational burden, which is crucial for real-time applications. Additionally, integrating MTMCT systems with edge computing offers a promising avenue to enhance real-time processing. By processing data closer to its source, latency can be substantially reduced. Optimizing MTMCT algorithms for edge devices, which often have limited computational resources, would ensure efficient operation and quicker data processing.

Effective resource management also plays a critical role in real-time MTMCT systems. Developing algorithms for dynamic allocation of computational resources, depending on the complexity of the tracking scene, would ensure the optimal use of available processing power. Handling varying data quality, crowd density, and environmental conditions effectively in real-time is another challenge that needs addressing. Algorithms capable of adapting to these variations in real time, while maintaining accuracy across different scenarios, would significantly enhance the robustness of real-time tracking systems.

Low-latency communication protocols are vital, especially for systems where data synchronization and analysis from multiple cameras are required promptly. Research

in this domain could leverage the potential of advanced technologies like 5G for high-speed data transmission, essential for synchronizing and analyzing data from multiple sources in real time.

Lastly, with growing concerns around privacy, developing real-time tracking systems that respect individual privacy is increasingly important. Techniques such as on-device processing, anonymization of tracking data for example blurring faces, and secure transmission methods could be key areas of research to ensure privacy-preserving real-time tracking.

In conclusion, enhancing online and real-time capabilities in MTMCT involves a multifaceted approach, encompassing algorithmic innovation, hardware optimization, and balancing the demands of speed, accuracy, and privacy. Addressing these research areas will lead to more responsive, efficient, and reliable real-time tracking solutions, aligning with the dynamic needs of modern applications.

4.4 Ethical and Privacy Concerns

Future advancements should balance technological progress with ethical considerations, ensuring privacy and ethical standards are secured. Research in the area of synthetically generated datasets and edge computing could potentially address the privacy concerns. Just to mention two significant examples for the concerns: The developer of the YOLO framework stopped working on the project due to ethical dilemmas fearing that his work could be used for military applications [101]. Similarly, the DukeMTMC dataset, was withdrawn over privacy issues [102]. These cases underscore the complex interplay between technological advancement and ethical responsibility.

5 Conclusion

As this exploration of MTMCT concludes, it is clear that the field is still navigating complex challenges and evolving demands. This journey, through the exploration of various methodologies, their inherent limitations, and prospective future directions, paints a comprehensive picture of the current state and potential evolution of MTMCT especially in the context of online and real-time environments.

While frameworks like R-CNN, YOLO, SORT, and DeepSORT have advanced intra-camera tracking, their limitations in inter-camera environments underscore the specific challenges of MTMCT. These challenges include the need for effective inter-camera tracking, robust data association across varied camera views, and maintaining consistent tracking accuracy in diverse and dynamic environments.

The integration of MTMCT with emerging technologies such as edge computing and IoT presents exciting opportunities to enhance the scope and effectiveness of tracking systems. These integrations could lead to more comprehensive and versatile systems, capable of handling the demands of real-time tracking in various applications, from urban surveillance to traffic management and public safety.

However, the advancement of MTMCT technologies also brings to the forefront the need for careful consideration of ethical and privacy issues. As the capabilities of MTMCT systems expand, ensuring their responsible use and addressing the societal implications of widespread surveillance are important. This involves developing frameworks that respect individual privacy and addressing the broader societal impacts of these technologies.

In summary, MTMCT stands as a field with significant potential and will impact a wider range of sectors as it already does. The path forward involves not only technological innovation but also a collaborative approach that includes researchers, technologists, policymakers, and ethicists. By addressing current limitations and exploring new horizons, MTMCT can achieve new levels of efficiency and accuracy, marking a new era of sophisticated and responsible tracking systems.

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] R. Iguernaissi, D. Merad, K. Aziz, and P. Drap, "People tracking in multi-camera systems: A review - multimedia tools and applications," *SpringerLink*, Sep. 2018. [Online]. Available: https://link.springer.com/article/10.1007/s11042-018-6638-5.

[3] 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.

[4] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: A comprehensive survey," *CoRR*, vol. abs/1912.00535, 2019. arXiv: 1912.00535. [Online]. Available: http://arxiv.org/abs/1912.00535.

[5] 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.

[6] W. Yu, F. Liang, X. He, *et al.*, "A survey on the edge computing for the internet of things," *IEEE Access*, vol. PP, pp. 1–1, Nov. 2017. DOI: 10.1109/ACCESS.2017.2778504.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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].

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] A. C. Glenn Jocher. "Ultralytics yolov8 docs." (2023), [Online]. Available: https://docs.ultralytics.com/ (visited on 11/01/2023).

[25] A. C. Glenn Jocher. "Ultralytics yolov8 github." (2023), [Online]. Available: https://github.com/ultralytics/ultralytics/ (visited on 11/01/2023).

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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,

[38] 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.

[39] 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.

[40] J. Ferryman and A. Shahrokni, "Pets2009: Dataset and challenge," in *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Dec. 2009, pp. 1–6. DOI: 10.1109/PETS-WINTER.2009.5399556.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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].

[64] 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.

[65] 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].

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] S. Wang, H. Sheng, Y. Zhang, D. Yang, J. Shen, and R. Chen, "Blockchain-empowered distributed multi-camera multi-target tracking in edge computing," *IEEE Transactions on Industrial Informatics*, pp. 1–10, 2023, ISSN: 1941-0050. DOI: 10.1109/TII.2023.3261890.

[71] M. Nikodem, M. Słabicki, T. Surmacz, P. Mrówka, and C. Dołęga, "Multi-camera vehicle tracking using edge computing and low-power communication," *Sensors*, vol. 20, no. 11, 2020, ISSN: 1424-8220. DOI: 10.3390/s20113334. [Online]. Available: https://www.mdpi.com/1424-8220/20/11/3334.

[72] 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.

[73] 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.

[74] 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.

[75] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762.

[76] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *CoRR*, vol. abs/2005.12872, 2020. arXiv: 2005.12872. [Online]. Available: https://arxiv.org/abs/2005.12872.

[77] F. Zeng, B. Dong, T. Wang, C. Chen, X. Zhang, and Y. Wei, "MOTR: end-to-end multiple-object tracking with transformer," *CoRR*, vol. abs/2105.03247, 2021. arXiv: 2105.03247. [Online]. Available: https://arxiv.org/abs/2105.03247.

[78] T. Meinhardt, A. Kirillov, L. Leal-Taixé, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 8834–8844. DOI: 10.1109/CVPR52688.2022.00864.

[79] P. Sun, Y. Jiang, R. Zhang, *et al.*, "Transtrack: Multiple-object tracking with transformer," *CoRR*, vol. abs/2012.15460, 2020. arXiv: 2012.15460. [Online]. Available: https://arxiv.org/abs/2012.15460.

[80] C. Zhang, C. Zhang, Y. Guo, L. Chen, and M. Happold, *Motiontrack: End-to-end transformer-based multi-object tracing with lidar-camera fusion*, 2023. arXiv: 2306.17000 [cs.CV].

[81] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang, "Online multi-object tracking with dual matching attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2019, pp. 366–382. arXiv: 1902.00749 [cs.CV].

[82] 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.

[83] Y.-J. Li, X. Weng, Y. Xu, and K. Kitani, "Visio-temporal attention for multi-camera multi-target association," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9814–9824. DOI: 10.1109/ICCV48922.2021.00969.

[84] 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.

[85] 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].

[86] T. N. Kipf and M. Welling, *Variational graph auto-encoders*, 2016. arXiv: 1611.07308 [stat.ML].

[87] H. Hsu, Y. Wang, and J. Hwang, "Traffic-aware multi-camera tracking of vehicles based on reid and camera link model," *CoRR*, vol. abs/2008.09785, 2020. arXiv: 2008.09785. [Online]. Available: https://arxiv.org/abs/2008.09785.

[88] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002, ISSN: 1939-3539. DOI: 10.1109/34.1000236.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, *Observation-centric sort: Rethinking sort for robust multi-object tracking*, 2023. arXiv: 2203.14360 [cs.CV].

[94] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 3701–3711. DOI: 10.1109/ICCV.2019.00380.

[95] 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].

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

[97] Z. Yang, Y. Wei, and Y. Yang, "Associating objects with transformers for video object segmentation," *CoRR*, vol. abs/2106.02638, 2021. arXiv: 2106.02638. [Online]. Available: https://arxiv.org/abs/2106.02638.

[98] Z. Yang and Y. Yang, *Decoupling features in hierarchical propagation for video object segmentation*, 2022. arXiv: 2210.09782 [cs.CV].

[99] S.-I. Yu, Y. Yang, and A. Hauptmann, "Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 3714–3720. DOI: 10.1109/CVPR.2013.476.

[100] 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.

[101] Synced. "Yolo creator says he stopped cv research due to ethical concerns." (2020), [Online]. Available: https://medium.com/syncedreview/yolo-creator-says-he-stopped-cv-research-due-to-ethical-concerns-b55a291ebb29 (visited on 11/01/2023).

[102] J. Harvey Adam. LaPlace. "Exposing.ai dukemtmc." (2019), [Online]. Available: https://exposing.ai/duke_mtmc/ (visited on 11/01/2023).