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, 2023-12-18

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 All passages taken verbatim or in spirit from the published or unpublished work of others, or from the author's own work, are marked as taken.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

This research project provides a comprehensive literature review in the area of Multi-Target Multi-Camera Tracking (MTMCT) and Re-Identification, with a focus on Real-Time Scenarios. The literature review begins with an exploration of the early developments in MTMCT, outlining the basic concepts and initial challenges. It then delves into key milestones that have shaped the field, focusing on critical aspects such as detection, feature extraction, data association, and tracking, along with the datasets and challenges encountered.

The review describes various methods and paradigms used in MTMCT, ranging from traditional tracking approaches to more sophisticated techniques such as single-shot, graph-based approaches, and the integration of edge computing. Special emphasis is given to the adaptation of online and real-time tracking methods, which are crucial in modern applications. In addition, the review covers recent advances in the field.

By examining state-of-the-art approaches, the project provides insight into the evolution and current capabilities of MTMCT systems. This comprehensive analysis not only reflects the significant progress made in the field, but also highlights the ongoing challenges and potential directions for future research, particularly in the development of efficient, robust, and ethically responsible MTMCT systems.

Contents

1 Introduction

Multi-Target Multi-Camera Tracking (MTMCT) is a crucial research area in computer vision with important applications in video surveillance, traffic monitoring, sports analysis and crowd management [1]–[4]. By simultaneously tracking multiple objects across multiple camera views, MTMCT systems aim to provide a comprehensive understanding of scene dynamics and interactions.

The advent of deep learning and other advanced algorithms has revolutionized the field of MTMCT, especially in recent years, enabling faster, more accurate and reliable tracking in complex environments [5]. In particular, online and real-time tracking methods have emerged as a critical area of focus due to their potential to provide timely and actionable insights in various real-world applications [6]–[8].

Even though Single-Target Single-Camera Tracking (STSCT) as well as Multi-Target Single-Camera Tracking (MTSCT) have been extensively studied, MTMCT is still a relatively new and challenging, but promising area of research. The complexity of MTMCT is significantly higher than that of STSCT and MTSCT due to the need to track multiple objects simultaneously using multiple cameras [9].

Single-Target Multi-Camera is an insignificant area 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 particular case.

This research project aims to provide a comprehensive review of the milestones and state-of-the-art in MTMCT, with a special focus on online and real-time tracking methods. The latest trends, technologies and challenges in the field will be explored, drawing insights from recent research papers and studies. It also highlights the significant advances that have been 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 provides a critical analysis of the methods, challenges, and future prospects in the field of MTMCT. Chapter 5 concludes by summarizing the main findings and outlining potential avenues for future research.

1.1 Definition of MTMCT

MTMCT is an integration of object detection and tracking methods to simultaneously track multiple objects of interest across different camera views. The goal 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 (FOV) of different cameras. The objects of interest are often people and vehicles, but could theoretically be any moving object. The camera setup varies from application to application, but typically consists of multiple cameras with either overlapping, non-overlapping, or partially overlapping FOVs. The cameras can be static or moving, and can be placed at different heights and angles. The cameras can also be heterogeneous, with different technical specifications such as resolution, frame rate, and FOV.

1.2 Importance of MTMCT

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

In addition, the need for online and real-time tracking is imperative in these applications. Real-time processing of data streams from multiple cameras and the provision of instant tracking results are essential to provide timely and actionable insights, which is particularly relevant in scenarios such as accident prevention, traffic flow control, crime detection, and real-time sports analysis.

1.3 Objective of Research Project

First, the basic 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 to 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 goal is to explore the previous milestones, recent trends, technologies, and challenges in the field and provide insights from recent research papers and studies. By highlighting the significant advances made in MTMCT, it is intended to identify the gaps in current research and outline potential avenues for future exploration, while keeping in mind the ethical and privacy concerns associated with MTMCT.

As the objectives of this research project are delved into, it is essential to provide an early overview of the various aspects and considerations in MTMCT tracking systems. The following Table 1.1 summarizes various parameters, technologies and challenges that are central to MTMCT. This table is intended to provide a snapshot of the complex landscape of MTMCT systems and to give context to the approaches and analyses explored in the following sections. Although some of the terms and concepts presented in the table are not fully explained at this point, they will be elaborated in the following sections to provide a comprehensive understanding of each aspect and its relevance to the research topic.

1.4 Related Work

The work of Zheng, Yang, and Hauptmann [10] focuses on the past, present and future of person re-identification (re-ID), which is the task of identifying a person across multiple cameras, for example when a person leaves and re-enters the FOV of a camera, or when a person is briefly lost by the detection framework. The paper covers hand-crafted algorithms as well as deep learning approaches for both image- and, more importantly, video-based re-ID. It also quickly explains important datasets and evaluates the approaches. Although this paper gives a good overview, it was published in 2016 and therefore does not cover the latest research in the field, which will be covered by this project.

Two years later, "People tracking in multi-camera systems: A Review - multimedia tools and applications" [11] was published, which gives an overview of multi-camera tracking methods. The review covers the most important methods and datasets 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 [12, Chapter 5], published in 2019, revolves around the topic of tracking multiple objects and gives a state-of-the-art overview of the field. It does not cover the topic of multi-camera tracking, but provides a mathematical insight into the topic of tracking multiple objects in single-camera systems.

The survey *Deep Learning for Visual Tracking: A Comprehensive Survey* [5] conducted by Marvasti-Zadeh, Cheng, Ghanei-Yakhdan, *et al.* in 2022 describes the evolution and 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 of which contributes 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 provides 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 focuses on single-object tracking, its insights into advances in deep learning architectures and methods provide a valuable context for the study of MTMCT, which also adapts and applies the aforementioned approaches. Thus, this survey serves as an important resource for 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.* [9]. It provides a detailed overview of the state-of-the-art in MTMCT, covering the latest trends, technologies, and challenges in the 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. Furthermore, this research project aims to provide an easy 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 into the IoT, as explored in "A Survey on the Edge Computing for the Internet of Things" [13], provides valuable insights into 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 large amounts of data from multiple cameras, it provides a parallel to the computational needs in MTMCT.

Table 1.1: Overview of MTMCT Aspects

Aspect	Parameters	Technologies	Challenges	Notes
Field of View	Fully, Partially, Non-Overlapping Seamlessly Adjacent	Camera Network Synchronization Wide Area Scene Analysis FOV-Aware Algorithms	Occlusion Varying Lighting Uncertainties	Each camera setup needs different technologies and solutions, that makes it challenging and requires advanced data association methods
Camera Calibration	Synchronization, Calibration Parameters Distortion Correction	Homography, Automatic Calibration Ground Plane Estimation 3D Reconstruction	Calibration Accuracy Environmental Changes	Accurate camera calibration is essential for reliable tracking, especially across camera views
Object Class	People Vehicle Others	Object Detection Class-Specific Feature Extraction Multi-Class Trackers	Class Differentiation Detection Accuracy	Differentiating between classes such as people and vehicles is necessary to extract features and perform tracking
Tracking	Multi-, Single-Shot Graph-, Transformer-, Based Hybrid Approaches	Graph Neural Networks Attention Models Multi-View Fusion	Data Fusion Tracking Accuracy	Different scenarios require different tracking techniques; a combination of several techniques is possible
Real-Time Processing	Yes, No, Near-Real-Time	Single-Shot Edge Computing Online Tracking Sub-Graph Approaches	Computational Overhead Latency Trade-off between Accuracy and Speed	Real-time processing requires fast algorithms with low computational overhead; on-device processing may be an option
Dataset and Evaluation	Standardized, Custom Datasets	Official Challenges, Benchmarking Platforms Performance Metrics	Annotation Quality Time-Consuming Labeling Relevance	Datasets and benchmarks for evaluation are crucial for reliable performance assessment and comparison between approaches
Ethical Consideration	Privacy Storage	Privacy-Preserving, Encrypted, On-Device Processing	Privacy Preservation Data Security Computing Costs	Ethical concerns must be addressed, especially in public surveillance applications

2 Background

This chapter provides an overview of the basic concepts of object detection and tracking and the steps of an MTMCT system, along with a discussion of its key challenges and issues. It also introduces the basic building blocks of MTMCT. Finally, the datasets and metrics used to evaluate MTMCT systems are explained.

2.1 Steps of an MTMCT System

Figure 2.1 visualizes the typical steps of an MTMCT system: detection, feature extraction, data association and tracking. The steps are only briefly mentioned and described here, exact methods and algorithms are discussed in the Chapter 3.

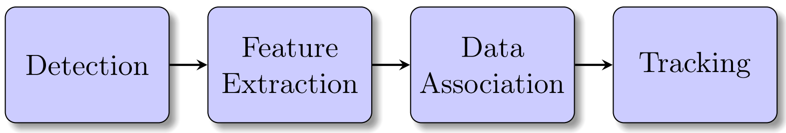

Figure 2.1: Steps of an MTMCT System

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 goal 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 can include low-level features such as color, shape and texture, as well as high-level features such as object parts and their spatial relationships, speed, and direction of motion. 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, it is necessary to define the following terms:

- **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 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 will be used.

Data association is the process of associating currently detected objects with existing trajectories based on similarities in their features. This is done by comparing the features of detected objects and trajectories with the features of existing trajectories and assigning the detected objects to the most similar trajectories. This step is critical for maintaining the IDs of objects as they move through or even leave and re-enter the scene, which is called re-ID. Typically, the data association step is performed hierarchically: first on a single-camera view (intra-camera), before the trajectories are associated across multi-camera views (inter-camera) and finally optimized globally to form tracks.

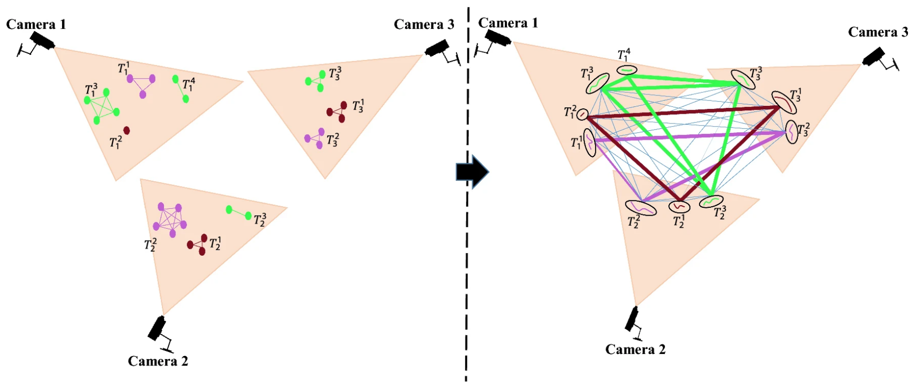

Figure 2.2: Intra- (left) and Inter-Camera (right) Tracking [source image: 8, Fig. 1]

The two steps of data association of three non-overlapping camera views are illustrated in Figure 2.2. The first step is intra-camera tracking, where the trajectories within the three camera views are associated independently. 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.

Figure 2.3 shows a schematic representation of the intra-, inter-camera matching and tracking process of the three camera views. This figure complements Figure 2.2 by visually demonstrating the steps mentioned. First, the people in the three camera views are detected and intra-camera tracking is performed by merging the individual detections into tracklets. Optionally, depending on the algorithm, localization within the world (using a common ground plane) could be performed. Finally, after the intra-camera trajectories have been generated, the inter-camera matching is performed to obtain the final tracks of the people.

2.1.4 Tracking

Tracking refers to the step of maintaining the tracklets and trajectories 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) become available with the next frame of a video. In summary, tracking is responsible for maintaining and managing the trajectories and IDs of objects as they move through the scene and ensuring consistent global IDs across multiple camera views.

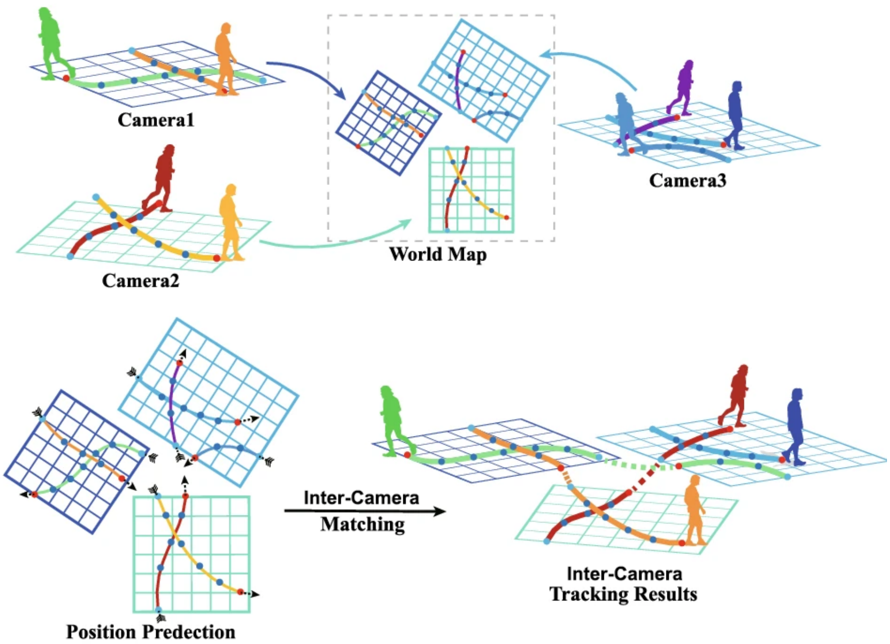

Figure 2.3: Tracking Process [source image: 14, Fig. 1]

2.2 Fundamental Concepts

This section briefly describes the fundamental concepts of MTMCT that are essential for following the progression from basic object tracking methods to advanced MTMCT techniques.

2.2.1 Single-Target Single-Camera Tracking

STSCT is the simplest form of object tracking and involves tracking a single target in the FOV of a single-camera. The primary goal of STSCT is to maintain the ID and trajectory of the target as it moves through the FOV of the camera.

2.2.2 Multi-Target Single-Camera Tracking

MTSCT builds on the principles of STSCT, but introduces the added complexity of dealing with multiple targets in a single-camera view. The goal is to track multiple targets 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 arise especially in crowded scenes.

The evolution from STSCT to MTSCT and ultimately to MTMCT reflects the increasing complexity and ability of tracking systems to handle more complex scenarios. This evolution is made possible by advances in computer vision and machine learning, which provide the tools necessary to overcome the challenges associated with tracking multiple targets across multiple camera views.

2.3 Challenges and Issues

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

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 maintain its ID. This can happen when objects overlap 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.

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 difficult to maintain consistent tracking. The presence of shadows and reflections can also make tracking difficult.

2.3.3 Camera Specifications

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

- **Resolution:** Number of pixels in the image
- **FPS:** Number of frames captured per second
- **FOV:** Area covered by the camera
- **Angle:** Angle from which the camera is capturing the scene

This can make it difficult to maintain consistent tracking across different camera views, especially when objects move from one camera to another. Objects can look different when viewed from different cameras, and their size and shape can be distorted. To achieve accurate tracking, the system must account for these variations and correctly align objects across camera views.

2.3.4 Uncertainties

In an MTMCT system, the number of objects present in the entire camera network, in a single-camera view, and the number of camera views in which a tracked object is present at any given time are all unknown. These uncertainties make it difficult to accurately track objects across multiple camera views.

2.4 Datasets

Datasets are a fundamental aspect of MTMCT research; they are the resource for training, evaluating, and comparing different tracking methods. A wide variety of datasets exist to meet the needs of research, each presenting unique challenges and scenarios.

Commonly used datasets for training object detectors are:

- **Microsoft COCO (Common Objects in Context):** Comprehensive dataset used for object detection and segmentation. COCO includes a wide variety of objects [15].
- **ImageNet:** Large dataset used for image classification and object detection. Object detectors trained on ImageNet are capable of recognizing a wide range of objects [16].

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

2.5 Metrics and Evaluation

Evaluating the performance of an MTMCT system is critical to understanding its effectiveness and reliability. In addition to well-known metrics such as accuracy, precision and recall, there are several metrics specifically designed for MTMCT systems. These metrics are discussed in this section.

2.5.1 MOTP and MOTA

Multiple Object Tracking Precision (MOTP) and Multiple Object Tracking Accuracy (MOTA) are two standard metrics used to evaluate multi-target tracking systems. MOTP measures the accuracy of object localization, while MOTA combines three types of error into a single metric to provide a comprehensive evaluation of tracking performance. Both metrics were introduced by Bernardin and Stiefelhagen [17] in 2008.

$$\text{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \quad [17, \text{Eq. } 1] \tag{2.1}$$

The equation 2.1 provides a measure of the average error in the estimated positions of the tracked objects. In this equation, d_t^i is the distance between the predicted position and the ground truth position of object i in 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 are 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 object localization.

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \quad [17, \text{Eq. } 2] \tag{2.2}$$

The equation 2.2 combines three types of errors into 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 (false object detections), mme_t is the number of mismatch errors (ID switches) and g_t is the total number of true objects present in the frame t. The MOTA score is one 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 identification precision and recall, providing a balanced measure that considers 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" [18].

$$\mathrm{IDF}_1 = \frac{2 \times \mathrm{IDTP}}{2 \times \mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}} \quad [18, \text{Eq. } 11] \tag{2.3}$$

In the Equation 2.3:

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

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

2.5.3 MT and ML

The Mostly Tracked (ML) and Mostly Lost (ML) metrics are used to evaluate the effectiveness of a tracking system in maintaining consistent trajectories for tracked objects. The metrics published by Wu and Nevatia [19] in 2006 are commonly used in 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 lifetime, indicating the ability of the system to track objects consistently 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.

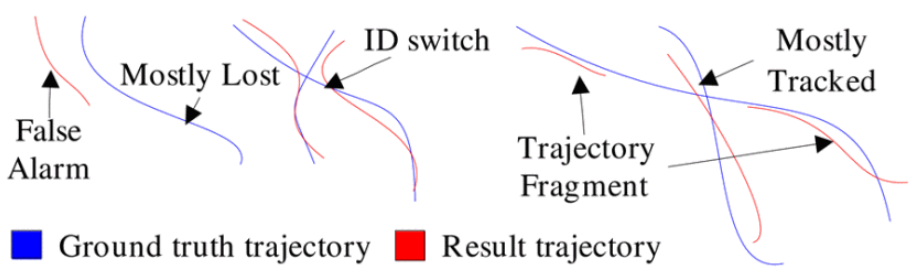

Figure 2.4: MT and ML [source image: 19, Fig. 5]

Figure 2.4 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 object in the ground truth.
- **ID Switch:** An instance where the ID assigned to an object is incorrectly changed during tracking.
- **Trajectory Fragment:** A segment of the resulting trajectory that is shorter than the ground truth, indicating a break or interruption in tracking.
- **Mostly Tracked:** Scenarios where the resulting trajectory closely follows the ground truth trajectory for most of the path of the object ($\geq 80\%$).
- **Mostly Lost:** Scenarios where the resulting trajectory only briefly aligns or intersects with the ground truth trajectory, indicating that 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, advances and milestones in the field. It will focus only on the most recent and state-of-the-art methods and technologies and will not cover the entire history of MTMCT, including all 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 [20] and Chang and Gong [21] did research on tracking people in a multi-camera system. Also in 2001, Khan, Javed, and Shah [22] 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 [23].

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

3.2 Milestones

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

The Figure 3.1 shows the algorithms and techniques in the green boxes (right) that correspond to each milestone area in the blue box (left). The following subsections discuss the techniques in more detail.

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 designed specifically for image processing [24]. The advent of deep learning in the last quarter century has led to a significant improvement in object detection performance.

With the introduction of R-CNN [25] 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 (RoI) using selective search and then classifies these regions using CNN features. Because R-CNN proposed RoI independently, it was computationally

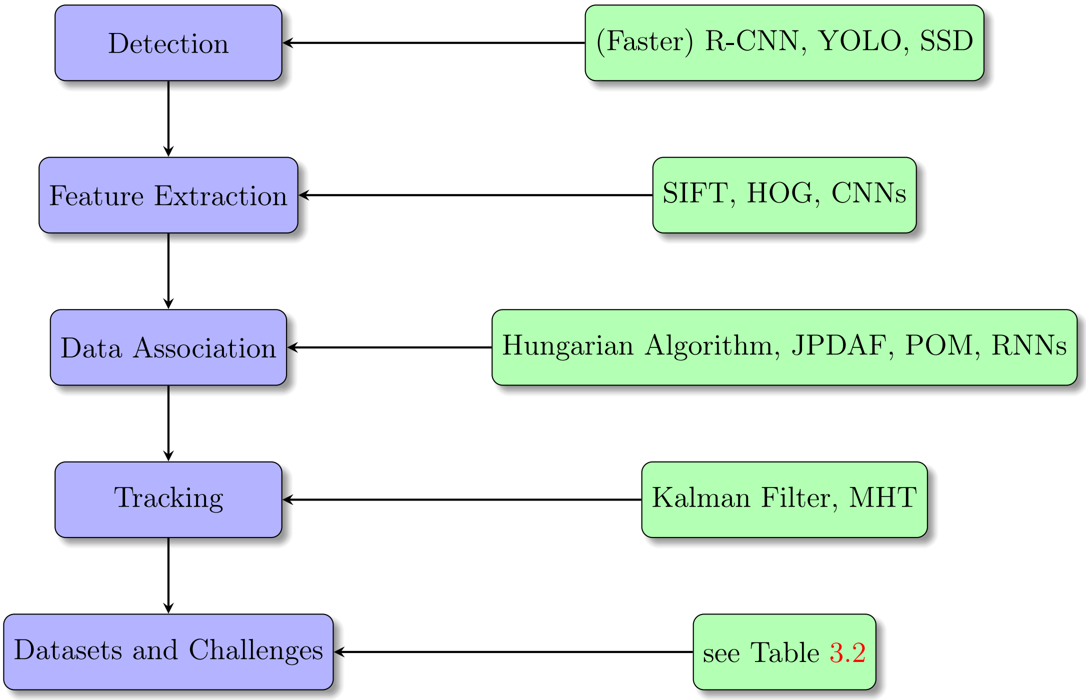

Figure 3.1: Milestones in MTMCT

expensive. Just a year later, improvements were made with Fast R-CNN [26], which addressed the inefficiencies of its predecessor by introducing a mechanism to share convolutional computations across region proposals and by incorporating a 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 [27], which integrates a Region Proposal Network (RPN) into the architecture that uses anchors, which are predefined reference boxes of various scales and aspect ratios, as the basis for proposing potential object locations. This allows the generation of region proposals at almost no cost 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) [28] and Single Shot MultiBox Detector (SSD) [29], which were designed to be even more efficient and suitable for real-time applications.

The YOLO framework, introduced 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 a single 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, which makes it very suitable for applications where real-time detection is critical. Although the original author has stopped working on YOLO due to ethical concerns, it is still being improved. The latest official version, YOLOv4 [30], was released in 2020 by Bochkovskiy, Wang, and Liao. In 2023 Ultralytics released the most recent version YOLOv8 [31], [32].

Liu, Anguelov, Erhan, *et al.* proposed SSD in 2016, another influential single-shot object detector that balances the tradeoff between speed and accuracy. Unlike YOLO, SSD operates on multiple feature maps at different resolutions to effectively handle objects of different sizes. The architecture applies a set of convolutional filters to these feature maps to predict both bounding box offsets and class probabilities for a fixed set of default bounding boxes distributed across the image. The ability to detect and track objects across multiple scales and perspectives makes SSD particularly well-suited for MTMCT applications.

Table 3.1: Overview Object Detectors

Model	Speed	Accuracy	Computational Requirements
Faster R-CNN [27]	Moderate	High	High
YOLO [28]	Very High	Moderate	Low
SSD [29]	High	High	Moderate

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

- **Speed:** Refers to the time it takes the detector to process a single frame, usually measured in milliseconds. High speed is critical for real-time tracking applications where live or near-live video feeds must be processed.
- **Accuracy:** Measures the ability to correctly identify and locate objects. It is usually quantified in terms of precision and recall rates, or average precision (AP) over a dataset.
- **Computational Requirements:** Refers to the resources required to run the detector, typically measured in terms of floating-point operations (FLOPs) or memory and processing power. 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) [33] and Histogram of Oriented Gradients (HOG) [34], which were crucial for object recognition and re-ID tasks. With the introduction of deep learning, CNNs have enabled automatic learning of feature representations, greatly improving robustness and power for re-ID [35], [36].

More recently, Siamese networks have emerged as a popular choice for learning discriminative features in a pairwise fashion, proving highly effective for re-ID tasks [37].

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 ID over time. The Hungarian Algorithm [38], also known as the Munkres assignment algorithm, has historically been used for optimal assignment in data association, addressing the problem of associating detections to trajectories in a globally optimal way.

The complexity of data association increased with the need to handle multiple objects and cameras, leading to the development of Joint Probabilistic Data Association Filters (JPDAF) [39], which consider the probabilities of all potential measurement-to-track assignments.

Fleuret, Berclaz, Lengagne, *et al.* introduced the use of Probabilistic Occupancy Map (POM) [40] to model targets in 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 it is possible to track multiple people in a complex environment. 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 by viewing the problem as finding the shortest path in a graph, where each node represents a detection and edges represent association costs [41]. This method became particularly useful for managing associations over long-time periods and occlusions.

More recently, with the rise of deep learning, Milan, Rezatofighi, Dick, *et al.* [42] use Recurrent Neural Networks (RNNs) [43] to learn the data association task, enabling an end-to-end approach to online multi-target tracking.

3.2.4 Tracking

The Kalman Filter [44] is one of the early foundations of object tracking, providing a framework for predicting the future locations of an object.

As tracking scenarios became more complex, approaches such as Multiple Hypothesis Tracking (MHT) [45] were developed to manage multiple potential data association hypotheses, especially in cluttered scenes [46].

3.2.5 Datasets and Challenges

In addition to the datasets mentioned in the Section 2.4, which are used for object detection in general, there are also datasets that are more specific to MTMCT. Typically, these are for tracking specific object classes, mainly people and vehicles. Datasets that meet these requirements are listed in the Table 3.2.

Table 3.2: Overview of Datasets

Dataset	Environment	Num. of Scenarios	Num. of Cameras (Overlap)	FPS	IDs	Year	Class
PETS [47]	Outdoor	3	8 (✓)	25	—	2009	Person
MARS[1] [48]	Mixed	Multiple	6 (✓)	—	1261	2016	Person
MOT16 [49]	Outdoor	14	1	25-30	—	2016	Person, Vehicle
DukeMTMC [18]	Outdoor	1	8 (✓)	60	2834	2016	Person
MOT17 [49]	Outdoor	14	1	25-30	—	2018	Person
Wildtrack [50]	Outdoor	Multiple	7 (✓)	2	313	2018	Person
MSMT17 [51]	Mixed	12	15 (✓)	15	4101	2018	Person
CityFlowV1 [1]	Outdoor	5	40 (✓)	10	666	2019	Vehicle
MOT20 [4]	Outdoor	8	1	25	—	2020	Person, Vehicle
CityFlowV2 [1]	Outdoor	6	46 (✓)	10	880	2021	Vehicle
MMPTRACK [52]	Indoor	5	23 (✓)	15	—	2023	Person
MEVID [53]	Mixed	17	33 (✓)	—	158	2023	Person

The Table 3.2 provides a summary of various datasets that have made significant contributions to the MTMCT research field. Each dataset is categorized based on several different criteria to reflect its unique characteristics and relevance:

- **Environment:** Data collection setting, from controlled indoor environments to dynamic outdoor locations.
- **Num. of Scenarios:** Specifies the number of distinct scenarios or situations represented in the dataset.
- **Num. of Cameras (Overlap):** Returns the number of cameras involved and indicates whether there is overlap in their views.
- **FPS:** Indicates the frame rate of the dataset, important for real-time processing considerations.

[1] extension of Market-1501 [54]

- **IDs:** Enumerates the unique IDs present, which can provide a measure of the complexity of the dataset.
- **Year:** Returns the year of publication, which represents the timeliness of the dataset.
- **Class:** Identifies the annotated classes, such as persons or vehicles.

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

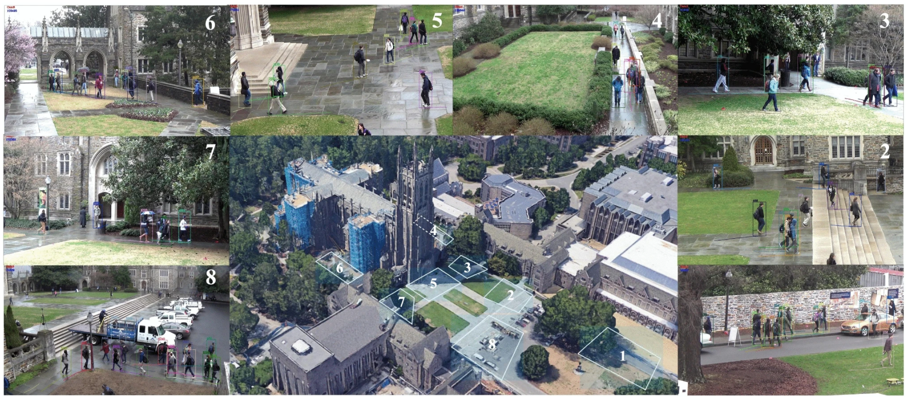

Figure 3.2: DukeMTMC [source image: 14, Fig. 2]

Figure 3.2 visualizes part of the DukeMTMC dataset as an example of one of the MTMCT datasets. It shows the camera views of the Duke University campus used in the dataset, illustrating the variety of scenes and perspectives for tracking individuals across different locations and conditions. The central aerial view shows the spatial arrangement of the cameras, with their respective FOVs highlighted. All eight camera views are shown in a given frame, including the bounding boxes of the tracked individuals.

In recent years, challenges have been established to encourage research in object detection and tracking, although they have mostly focused on STSCT and MTSCT. 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. It also provides ground truth data to facilitate evaluation. The MOT datasets are released in conjunction with the MOTChallenge [4].
- **2023 AICity Challenge:** Focuses on AI applications in smart cities and includes multi-object tracking for traffic monitoring and anomaly detection as one of its key components. The CityFlow datasets are part of the AICity Challenges [2].
- **VOT2022 Challenge (Visual Object Tracking Challenge):** An annual competition that provides a standardized dataset and evaluation framework for tracking single objects [55].
- **VOTS2023 Challenge (Visual Object Tracking and Segmentation Challenge):** An extension of the VOT Challenge that focuses on multi-object tracking. Recently published in October 2023, the challenge confirms the rapidly growing interest in this area [56].

3.3 Methods

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

3.3.1 Tracking Paradigms

In recent years, several tracking paradigms have been developed and 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 perform data association to link the detections across frames. This Tracking-by-Detection (TbD) implementation is a multi-shot approach that treats detection and association as separate, sequential tasks, allowing the use of specialized methods tailored for each step.

Bewley, Ge, Ott, *et al.* proposed one of the pioneering works in this area, the Simple Online and Realtime Tracking (SORT) [57] algorithm. SORT uses a combination of Kalman Filters to predict the motion of objects and the Hungarian Algorithm to associate detections over time based on both predicted locations and detected bounding boxes. Its efficiency and speed make it suitable for real-time applications, although it may struggle with ID switches in crowded scenes due to its reliance on motion cues alone.

Building on the foundation laid by SORT, the DeepSORT [58] algorithm was introduced by Wojke, Bewley, and Paulus, which improves tracking performance by incorporating deep learning techniques for extracting of appearance features. 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 prediction is insufficient, such as occlusions or complex, dynamic environments.

Both SORT and DeepSORT have set benchmarks in object tracking, 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:** Enhances SORT by adding appearance information to the data association step, improving tracking accuracy, especially in cases where objects interact closely or are temporarily occluded.

It is important to note 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 the Section 3.2.1. Neither SORT nor DeepSORT can perform inter-camera tracking; this step must 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. In contrast to the TbD approach, regression-based methods predict changes in position of the object. Most approaches learn a regression function from the input image features to the position coordinates. The advantage of this method is its ability to continuously refine the estimated position, making it well-suited for scenarios with smooth motion 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 may change shape, size, or orientation. By combining tracking and segmentation, this approach provides more detailed and accurate object tracking, especially in environments where distinguishing 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 selectively focus, has been integrated into tracking frameworks to dynamically highlight important spatial and temporal features.

3.3.2 Single-Shot Approaches

In contrast to all of the above tracking paradigms, single-shot approaches aim to perform detection and data association simultaneously in a single-step. While less common, this approach 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 very effective.

A notable contribution in this area is the Single-Shot Multi Object Tracking (SMOT) [59] algorithm proposed by Li, Xiong, Yang, *et al.* in 2020. SMOT is a tracking framework capable of converting any single-shot object detector into a multi-object tracker, which is able to simultaneously generate detection and tracking outputs. It is based on the work of Bergmann, Meinhardt, and Leal-Taixé, who developed the Tracktor [60], an object detector that is also able to track objects at the same time. The SMOT framework is able to generate tracklets with an almost constant runtime with respect to the number of targets, due to the use of a lightweight linking algorithm for online tracklet linking.

In the same year, Wang, Zheng, Liu, *et al.* published the paper "Towards Real-Time Multi-Object Tracking" [61], which proposes a single deep network that Jointly learns the Detection and Embedding (JDE) model. By reducing the computational cost, the system is able to achieve (near) real-time performance while being nearly as accurate as the models trained separately for detection and embedding. The architecture is based on the Feature Pyramid Network (FPN) [62], which is useful for detecting objects of different sizes. A variation of the triplet loss [63] is used to learn the embedding space, which is used for data association. This variation of the triplet loss is defined as follows:

$$\mathcal{L}_{\text{triplet}} = \sum_{i} \max\left(0, f^{\top} f_i^{-} - f^{\top} f^{+}\right) \quad [61, \text{Eq. } 1] \tag{3.1}$$

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

The triplet loss defined in the Equation 3.1 is used to learn an embedding space in which instances of the same ID are closely mapped to each other, while the embeddings of different IDs are pushed apart.

A more recent framework is the FairMOT [64] 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 multitask learning, which arises because re-ID is often treated as a secondary task in existing frameworks and not given enough attention. The paper raises three key issues with existing multitask learning frameworks:

1. **Unfairness Caused by Anchors:** The re-ID task is overlooked in the anchor-based detection framework, where anchors are optimized only 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 IDs, leading to a conflict between the two tasks.
3. **Unfairness Caused by Feature Dimension:** The feature dimension of re-ID features is usually much higher than that of detection features, but high-dimensional features significantly degrade detection performance.

To jointly train the detection and re-ID branches in the FairMOT network, the uncertainty loss proposed by Cipolla, Gal, and Kendall [65] is used. The uncertainty loss is defined as:

$$L_{\text{total}} = \frac{1}{2}\left(\frac{1}{e^{w_1}} L_{\text{detection}} + \frac{1}{e^{w_2}} L_{\text{identity}} + w_1 + w_2\right) \quad [64, \text{Eq. } 5] \tag{3.2}$$

The uncertainty loss defined in the Equation 3.2 is used to jointly train the detection and re-ID tasks by assigning different weights to the two tasks to allow for 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 multitask learning frameworks, the FairMOT framework is able to outperform state-of-the-art methods in both tracking accuracy and speed on the MOT17 dataset.

An important note is that the term *single-shot* used by these 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 data association problem can be formulated as a graph optimization problem, where each node represents a detection and the edges represent the association costs [41]. The goal is to find the shortest path in the graph representing the sequence of object detections over time. More recently, Graph Neural Networks (GNNs) [66] have been used to learn the data association task, allowing for an end-to-end approach to tracking.

In 2017, Chen, Cao, Chen, *et al.* [67] proposed a pedestrian tracking model that combines inter- and intra-camera tracking and unifies the two steps into a global graph by considering the initial observations as inputs and directly outputting the final trajectories. Since the initial observations contain more information such as motion than simple detections, they are more credible for data association. Furthermore, it speeds up the computation time because the number of observations is much smaller than the number of detections. The main focus of this paper is to equalize the similarity metrics of both tasks to allow unbiased data association. An equalization of the metrics is necessary because, if it is not applied, the joint approach would almost always favor objects from the same camera view as more similar because the observations are made under the same circumstances such as viewing 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 in the second step and produces mismatches errors.

Similar to [67], 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) [68] and can be used in both TbD and TbA 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) [69] 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 advantages are better feature representation and the ability to recover lost tracklets during camera transitions. DyGLIP combines link prediction with a dynamic graph formulation that, for the first time in MTMCT, takes into account temporal information of an object. Based on this approach, Cheng, Qiu, Chiang, *et al.* propose a Reconfigurable Spatial-Temporal Graph Model (ReST) [70] in 2023, which 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 intra-camera tracking results because it directly matches inter-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 [3] uses raw detection heat maps of the feet of the players directly instead of bounding boxes. The feet of the players are detected by the pre-trained detector FootAndBall [71], the detection heat maps of all cameras are transformed to a bird's eye view plane and stacked to form a multi-channel tensor. This results in the extraction and aggregation being performed within the tracking network itself, rather than in a separate preprocessing step as in conventional approaches, thus following the TbR paradigm. The tracking network consists of a Long Short-Term Memory-based (LSTM) [72] RNN, which models the player dynamics and a GNN, which is able to learn the interaction between players. The training data is synthetically generated by the Google Research Football Environment (GRF) [73] and the final tracker is compared with a baseline approach based on a particle filter. Although the proposed tracker cannot use visual cues such as jersey numbers due to the 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, as opposed to the traditional approach of processing data in a centralized cloud. The advantages of edge computing are low latency, reduced bandwidth and improved security because raw video data is not stored. The main drawback is the limited computational resources of the edge devices, in this case the cameras themselves.

In the paper on the single-shot approach SMOT discussed earlier, it is mentioned that replacing the components of the SMOT framework with faster versions can achieve real-time performance on less powerful machines such as edge devices.

Wang, Sheng, Zhang, *et al.* present a Multi-Camera Multi-Hypothesis Tracking (MC-MHT) framework integrated with a blockchain-based system, Multi-Camera TrackingChain (MCTChain) [74]. This extensible architecture distributes tracking tasks among cameras, improving scalability and security compared to centralized approaches. The architecture consists of three layers: the tracking, the blockchain and the 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 responsible for the 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" [75] introduces a decentralized tracking algorithm following the TbD paradigm, which performs intra-camera tracking locally on the camera and uses 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 cover single-shot approaches and edge computing and their relevance to real-time applications, this subsection focuses on online and real-time implementations, mentioning specific methods.

Unlike most of the methods used in MTMCT, the real-time system Uni-ID [6] follows a distributed concept to ensure that the communication and computing costs of each camera in the network remain nearly 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 are performed to assign a local ID to each object. Second, the local ID, features and trajectory information of the target are sent to the neighboring node in the network. Third, the neighboring 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-power GPU for each node.

The work of Wang, Liao, Hsieh, *et al.* [7] focuses on the conspicuous 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 can 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 five fisheye cameras and five to ten people in a scene simultaneously, the system achieves real-time performance of about ten FPS without GPU support.

Tesfaye, Zemene, Prati, *et al.* propose the use of Fast-Constrained Dominant Set Clustering (FCDSC) [8] to solve both intra- and inter-camera simultaneously. The method is orders of magnitude faster than existing graph-based methods because it considers only a sub-graph instead of the entire graph at each step. The proposed method follows a three-layer hierarchical approach. The first two layers solve intra-camera tracking and the third layer solves inter-camera tracking by merging the trajectories of the same person across camera views. The tracking algorithm runs at 18 FPS and is 2000 times faster than the CDSC [76] on which it is based.

3.3.6 Attention Models and Transformers

Recent advances in MTMCT have been influenced by the development of attention models and transformers [77], originally applied to natural language processing and designed to improve focus in neural networks. Despite their advantages, the high resource requirements, particularly in terms of processing power and memory, pose a challenge in achieving the low latency required for online and real-time MTMCT tracking applications. Nevertheless, examples of attention models and transformer implementations are at least mentioned in the context of this project.

The paper *End-to-End Object Detection with Transformers* (DETR) [78] published by Carion, Massa, Synnaeve, *et al.* in 2020 lays the foundation for the use of transformers in object detection. It combines a transformer with a set-based global loss and demonstrates significant improvements in accuracy and efficiency. This work paved the way for subsequent transformer-based MOT models.

MOTR [79] extends the DETR framework by introducing a "track query" for tracking multiple objects across frames. MOTR updates track queries iteratively, proving temporal relation modeling and MOT performance. TrackFormer [80] presents an end-to-end trainable model that uses static object queries to initialize new tracks and autoregressive track queries for existing tracks. TransTrack [81] introduces a method that simultaneously handles object detection and association. It uses the object features of the previous frame as a query for the current frame, simplifying the tracking process. Furthermore, MotionTrack [82] demonstrates the application of transformers in an autonomous driving environment with multiple sensor inputs.

The Dual Matching Attention Networks (DMAN) [83] approach 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 aforementioned tracking frameworks, which solve the task of intra-camera tracking and are not able to perform inter-camera tracking. MVDeTr [84] and the model of Li, Weng, Xu, *et al.* [85] 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 an important step in addressing occlusions and view inconsistencies in MTMCT. [85] uses transformer-based attention mechanisms for robust person association across different camera views. This approach is instrumental in improving 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 published in 2022 and 2023. The approaches achieve state-of-the-art performance on MTMCT datasets, but are not real-time applicable, due to the use of computationally expensive methods.

Hsu, Wang, Cai, *et al.* introduce a Self-supervised Camera Link Model (SCLM) [86] that extracts both appearance and topological features from a Graph Auto-Encoder (GAE) [88] to achieve vehicle tracking in a multi-camera environment. The approach follows the TbD paradigm and advances the Traffic-Aware Single Camera Tracking (TSCT) [89] algorithm, which proposes a zone generation algorithm. After the common steps object detection and feature extraction, these are used as nodes for the GAE to build 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 [90] clustering algorithm. The combination of the TSCT and GAE embeddings with zone generation results in state-of-the-art performance on the CityFlow 2019 and 2020 datasets.

Lifted Multicut Meets Geometry Projections (LMGP) [91] 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 the 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 considers short- and long-range temporal interactions to perform inter-camera matching. Intra-camera tracking is performed by CenterTrack [92] and embedding vectors are extracted by DG-Net [93]. LMGP achieves near perfect state-of-the-art performance on the Wildtrack dataset.

EarlyBird [87] proposes early fusion in the Bird's Eye View (BEV), i.e., detections are performed directly in the BEV to solve the spatial association of pedestrians across cameras. The approach is based 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 an encoder network, the image features are projected onto the ground plane and aggregated to obtain BEV features (in BEV space). Finally, the detections and their corresponding re-ID features are fed through a decoder network to associate 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 entire feature space onto the ground plane and associating it with the decoder network. EarlyBird shows that early fusion in the BEV space can outperform late fusion in the image space. The disadvantage is higher computational cost due to the simultaneous projection of full images from all camera views onto the ground plane. It also requires high-quality 3D annotations, which are 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 correctly identify people even when significant changes in appearance occur, such as different clothing or lighting conditions. The proposed method follows the TbD paradigm and combines a state-of-the-art OC-SORT-based [95] tracker with the person re-ID library Torchreid [96] for feature extraction. The feature extraction is performed as a feature averaging, taking into account only those frames where the person is not occluded and is not 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 the people in the area is computed and the ID is restored if the distance is below a certain threshold, this works both for matching people in the same camera and across camera views. Furthermore, a new dataset containing 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 difficult to compare the performance in a broader context.

Huang, Yang, Jiang, *et al.* [97] win first place 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 the tracking accuracy of people with similar appearance. Three steps are performed to achieve the final tracking results. First, intra-camera tracking is performed with BoT-SORT [98] 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 ID under different conditions. Third, human pose with camera self-calibration is used to project the tracked objects onto a top-down map.

The "The First Visual Object Tracking Segmentation VOTS2023 Challenge Results" presents the performance of the 47 trackers submitted to the challenge. Most of the trackers use a uniform dynamic model and transformers. Both multi- and single-shot approaches are used, but the top three trackers are single-shot approaches. The top tracker DMAOT is based on the VOT22 [55] winner AOT [99] and its successor DeAOT [100]. Although detailed technical documentation of DMAOT is currently unavailable, it is known to store long-term memory by object rather than by frame to predict object masks. Overall, the challenge shows a paradigm shift from bounding-box trackers to segmentation-based trackers that outperform all bounding-box trackers in the challenge.

3.3.8 Honorable Mentions

In the exploration of advanced tracking methods, certain studies stand out for their unique and unconventional approaches. This subsection highlights two of these studies that are not directly related to real-time MTMCT but still deserve honorable mentions.

An interesting development in the field of people tracking is "Harry Potter's Marauder's Map: Localizing and Tracking Multiple Persons-of-Interest by Nonnegative Discretization" [101] from 2013. Drawing parallels to the fictional "Marauder's Map" in the Harry Potter series, this research proposes a framework that follows the TbD paradigm and uses non-negative discretization for robust localization and tracking of people in complex environments. Their method overcomes challenges such as occlusion and sparse surveillance camera coverage by employing a semi-supervised learning framework that integrates cues such as color, person detection, face recognition, and non-background information. Its application in a real-world nursing home environment, 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" [102], a unique dataset for MTMCT research captured within the virtual environment of the popular video game Grand Theft Auto V (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 the advancement of computer vision without invading the privacy of individuals.

4 Discussion

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

4.1 Summary of Methods

Early research in MTMCT relied primarily on Bayesian classification and simple network models. These basic methods were instrumental in kick-starting research in the field, but they were limited by robustness and error susceptibility, especially in challenging scenarios such as occlusions or varied poses.

The advent of CNNs marked a significant shift, introducing deep learning to object detection and significantly improving performance. Building on CNN, R-CNN and its successors, including Fast R-CNN and Faster R-CNN, improved efficiency and reduced computational overhead. They introduced more effective RoI handling.

The introduction of real-time object detection frameworks such as YOLO and SSD was a game changer. YOLO, with its single forward pass image processing, revolutionized real-time object 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 were critical to maintaining object ID over time and across different camera views.

Various tracking paradigms such as TbD, TbR, TbS, and TbA have emerged, each addressing specific aspects of tracking with their unique advantages. Furthermore, single-shot approaches such as SMOT and JDE have streamlined the process by integrating intra-camera detection and tracking in a single-step, emphasizing speed and simplicity toward real-time tracking. In addition, frameworks such as FairMOT, FCDSC, and JDE have set milestones for real-time tracking applications.

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

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

4.2 Gaps and Limitations

While each of the aforementioned approaches has unique advantages, they also have inherent limitations. The following sections discuss the gaps and limitations of these methods.

Feature extraction and data association techniques such as SIFT, HOG, the Hungarian Algorithm, JPDAF and POM have been instrumental in the development of MTMCT systems. However, these methods are not able to handle scenarios that increase 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 degrades significantly in the presence of abrupt motion changes or maneuvering targets. On the other hand, MHT, known for its robustness to 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.

Existing datasets and challenges still focus primarily on intra-camera tracking. In addition, there are no challenges that focus solely on real-time tracking within a multi-camera system. These challenges would be beneficial for the advancement of research in these areas, as developers would be motivated to develop new methods and algorithms to compete in these challenges.

Intensive research on intra-camera detection and tracking frameworks such as YOLO, Faster R-CNN, SSD and DeepSORT has led to significant progress in these areas. However, these methods are not optimized for inter-camera tracking, which requires 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 current research.

Single-shot approaches such as SMOT and JDE have been instrumental in simplifying the tracking process by integrating detection and tracking into a single-step, increasing speed and efficiency. However, the prioritization of speed compromises accuracy, especially when detecting small or overlapping objects.

Graph-based approaches such as DyGLIP and ReST are powerful frameworks for data association, especially in occlusion scenarios. However, graph-based approaches are computationally intensive, which is a significant limitation in real-time applications. With the introduction of FCDSC, which considers only a sub-graph at a time, the computational complexity of graph-based approaches has been significantly reduced, finally making 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 has not yet been fully realized and in particular, their use in real-time applications is not yet feasible.

While a lot of progress has been made in the area of short-term trackers, long-term trackers have been overlooked, even though they are closer to real-world scenarios. Therefore, Marvasti-Zadeh, Cheng, Ghanei-Yakhdan, *et al.* [5] call for the development of trackers that are capable of re-ID targets over a long period of time. In addition, the authors state that generic visual trackers are needed to quickly adapt to unseen targets in real-world scenarios.

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

4.3 Future Research

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

Algorithms should ideally balance speed and accuracy, providing accurate tracking information quickly. Emphasis on lightweight neural network architectures could lead to models that maintain high accuracy while reducing computational complexity, which is critical for real-time applications. In addition, the integration of MTMCT systems with edge computing offers a promising way to improve real-time processing [13]. By processing data closer to its source, latency can be significantly reduced. Optimizing MTMCT algorithms for edge devices, which often have limited computational resources, would ensure efficient operation and faster data processing.

Amosa, Sebastian, Izhar, *et al.* [9] mention other aspects that need to be addressed in the future. For example, they call for a unified evaluation metric for multi-camera systems. Currently, evaluation is still based on single-camera tracking metrics, which are outdated for evaluating a multi-camera system. They also suggest investigating semi- or unsupervised learning approaches in the context of MTMCT, which would reduce the need for labeled data. They also consider the integration of language-vision models, which could improve tracking accuracy by incorporating textual information into the tracking process.

Effective resource management also plays a critical role in real-time MTMCT systems. Developing algorithms to dynamically allocate computational resources based on the complexity of the tracking scene would ensure optimal use of available processing power. Effectively dealing with varying data quality, crowd density, and environmental conditions in real-time is another challenge that needs to be addressed. Algorithms that can adapt to these variations in real-time, while maintaining accuracy across different scenarios, would significantly improve the robustness of real-time tracking systems.

Low-latency communication protocols are essential, especially for systems that require real-time data synchronization and analysis from multiple cameras. Research in this area could take advantage of the potential of advanced technologies, such as 5G, for high-speed data transmission, which is essential for synchronizing and analyzing data from multiple sources in real-time.

Finally, as privacy concerns grow, it is increasingly important to develop real-time tracking systems that respect individual privacy. Techniques such as on-device processing, anonymization of tracking data such as face blurring, and secure transmission methods could be key areas of research to ensure privacy-preserving real-time tracking.

In summary, improving online and real-time capabilities in MTMCT requires a multifaceted approach that includes 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 that meet the dynamic needs of modern applications.

4.4 Ethical and Privacy Concerns

Future advances should balance technological progress with ethical considerations, ensuring that privacy and ethical standards are protected. Research in the areas of synthetically generated datasets and edge computing could potentially address privacy concerns. To name just two important examples of these concerns: The developer of the YOLO framework stopped work on the project due to ethical dilemmas, fearing that his work could be used for military applications [103]. Similarly, the DukeMTMC dataset was withdrawn due to privacy concerns [104]. These cases highlight the complex interplay between technological progress and ethical responsibility.

5 Conclusion

As this exploration of MTMCT concludes, it is clear that the field continues to navigate complex challenges and evolving requirements. By exploring different methodologies, their inherent limitations, and possible future directions, insight is provided into the current state and potential evolution of MTMCT, particularly in the context of online and real-time environments.

While frameworks such as 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 multiple 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 improve the scope and effectiveness of tracking systems. These integrations could lead to more comprehensive and versatile systems capable of meeting real-time tracking requirements in applications ranging 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, it is important to ensure their responsible use and address the societal implications of widespread surveillance. This includes developing frameworks that respect individual privacy and address the broader societal implications of these technologies.

In conclusion, MTMCT is an area of significant potential and will, as it already does, impact a wide range of sectors. The way forward requires not only technological innovation, but also a collaborative approach involving researchers, technologists, policy makers and ethicists. By addressing current limitations and exploring new horizons, MTMCT can achieve new levels of efficiency and accuracy, ushering in a new era of sophisticated and responsible real-time tracking systems.

List of Figures

List of Tables

Bibliography

[1] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. Hwang, "Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8789–8798. DOI: 10.1109/CVPR.2019.00900.

[2] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. S. Rahman, M. S. Arya, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, S. Prajapati, A. Li, S. Li, K. Kunadharaju, S. Jiang, and R. Chellappa, "The 7th ai city challenge," in *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2023. DOI: 10.48550/arXiv.2304.07500.

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

[4] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, *Mot20: A benchmark for multi object tracking in crowded scenes*, 2020. DOI: 10.48550/arXiv.2003.09003.

[5] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, *Deep learning for visual tracking: A comprehensive survey*, 2022. DOI: 10.1109/TITS.2020.3046478.

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

[7] 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 *International Conference on Visual Communications and Image Processing*, 2021, pp. 1–5. DOI: 10.1109/VCIP53242.2021.9675451.

[8] 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, 2019. DOI: 10.1007/s11263-019-01180-6.

[9] T. I. Amosa, P. Sebastian, L. I. Izhar, O. Ibrahim, L. S. Ayinla, A. A. Bahashwan, A. Bala, and Y. A. Samaila, "Multi-camera multi-object tracking: A review of current trends and future advances," *Neurocomputing*, vol. 552, 2023. DOI: 10.1016/j.neucom.2023.126558.

[10] L. Zheng, Y. Yang, and A. G. Hauptmann, *Person re-identification: Past, present and future*, 2016. DOI: 10.48550/arXiv.1610.02984.

[11] R. Iguernaissi, D. Merad, K. Aziz, and P. Drap, "People tracking in multi-camera systems: A review - multimedia tools and applications," *Springer*, 2018. DOI: 10.1007/s11042-018-6638-5.

[12] W. Tian, "Novel aggregated solutions for robust visual tracking in traffic scenarios," 2019. DOI: 10.5445/KSP/1000091919.

[13] W. Yu, F. Liang, X. He, W. Hatcher, C. Lu, J. Lin, and X. Yang, "A survey on the edge computing for the internet of things," 2017. DOI: 10.1109/ACCESS.2017.2778504.

[14] C. Ma, F. Yang, Y. Li, H. Jia, X. Xie, and W. Gao, "Deep trajectory post-processing and position projection for single & multiple camera multiple object tracking," vol. 129, 2021. DOI: 10.1007/s11263-021-01527-y.

[15] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft COCO: common objects in context*, 2014. DOI: 10.48550/arXiv.1405.0312.

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

[17] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," vol. 2, 2008. DOI: 10.1155/2008/246309.

[18] 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*, 2016, pp. 17–35. DOI: 10.48550/arXiv.1609.01775.

[19] B. Wu and R. Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 951–958. DOI: 10.1109/CVPR.2006.312.

[20] 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, 1999. DOI: 10.1109/34.809119.

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

[22] S. Khan, O. Javed, and M. Shah, "Tracking in uncalibrated cameras with overlapping field of view," in *IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, vol. 5, 2001.

[23] J. Pearl, "Probabilistic reasoning in intelligent systems," in *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, 1988, pp. 1–141. DOI: 10.1016/B978-0-08-051489-5.50007-2.

[24] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.

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

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

[27] 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, 2017. DOI: 10.1109/TPAMI.2016.2577031.

[28] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, *You only look once: Unified, real-time object detection*, 2015. DOI: 10.48550/arXiv.1506.02640.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Computer Vision – ECCV*, Springer International Publishing, 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0_2.

[30] A. Bochkovskiy, C. Wang, and H. M. Liao, *Yolov4: Optimal speed and accuracy of object detection*, 2020. DOI: 10.48550/arXiv.2004.10934.

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

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

[33] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," vol. 60, no. 2, pp. 91–110, 2004. DOI: 10.1023/B:VISI.0000029664.99615.94.

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

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," vol. 25, 2012. DOI: 10.1145/3065386.

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

[37] R. R. Varior, M. Haloi, and G. Wang, *Gated siamese convolutional neural network architecture for human re-identification*, 2016. DOI: 10.48550/arXiv.1607.08378.

[38] H. W. Kuhn, "The hungarian method for the assignment problem," vol. 2, no. 1–2, pp. 83–97, 1955. DOI: 10.1002/nav.3800020109.

[39] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," vol. 8, no. 3, pp. 173–184, 1983. DOI: 10.1109/JOE.1983.1145560.

[40] 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, 2008. DOI: 10.1109/TPAMI.2007.1174.

[41] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587584.

[42] A. Milan, S. H. Rezatofighi, A. R. Dick, K. Schindler, and I. D. Reid, *Online multi-target tracking using recurrent neural networks*, 2016. DOI: 10.48550/arXiv.1604.03635.

[43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," vol. 323, no. 6088, pp. 533–536, 1986. DOI: 10.1038/323533a0.

[44] R. E. Kalman, "A new approach to linear filtering and prediction problems," vol. 82, no. 1, pp. 35–45, 1960. DOI: 10.1115/1.3662552.

[45] S. Blackman, "Multiple hypothesis tracking for multiple target tracking," vol. 19, no. 1, pp. 5–18, 2004. DOI: 10.1109/MAES.2004.1263228.

[46] D. Reid, "An algorithm for tracking multiple targets," vol. 24, no. 6, pp. 843–854, 1979. DOI: 10.1109/TAC.1979.1102177.

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

[48] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *Computer Vision – ECCV*, Springer International Publishing, 2016, pp. 868–884. DOI: 10.1007/978-3-319-46466-4_52.

[49] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, *MOT16: A benchmark for multi-object tracking*, 2016. DOI: 10.48550/arXiv.1603.00831.

[50] T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret, "Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5030–5039. DOI: 10.1109/CVPR.2018.00528.

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

[52] X. Han, Q. You, C. Wang, Z. Zhang, P. Chu, H. Hu, J. Wang, and Z. Liu, "Mmptrack: Large-scale densely annotated multi-camera multiple people tracking benchmark," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4849–4858. DOI: 10.1109/WACV56688.2023.00484.

[53] D. Davila, D. Du, B. Lewis, C. Funk, J. Van Pelt, R. Collins, K. Corona, M. Brown, S. McCloskey, A. Hoogs, and B. Clipp, "Mevid: Multi-view extended videos with identities for video person re-identification," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1634–1643. DOI: 10.1109/WACV56688.2023.00168.

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

[55] M. Kristan, A. Leonardis, J. Matas, *et al.*, "The tenth visual object tracking vot2022 challenge results," in *Computer Vision – ECCV Workshops*, Springer Nature Switzerland, 2022, pp. 431–460. DOI: 10.1007/978-3-031-25085-9_25.

[56] 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 Workshops*, 2023, pp. 1796–1818.

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

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

[59] W. Li, Y. Xiong, S. Yang, S. Deng, and W. Xia, *SMOT: single-shot multi object tracking*, 2020. DOI: 10.48550/arXiv.2010.16031.

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

[61] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Computer Vision – ECCV*, Springer International Publishing, 2020, pp. 107–122. DOI: 10.1007/978-3-030-58621-8_7.

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

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

[64] 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, 2021. DOI: 10.1007/s11263-021-01513-4.

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

[66] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," vol. 20, no. 1, pp. 61–80, 2009. DOI: 10.1109/TNN.2008.2005605.

[67] 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, 2017. DOI: 10.1109/TCSVT.2016.2589619.

[68] 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. DOI: 10.48550/arXiv.2211.09663.

[69] K. G. Quach, P. Nguyen, H. Le, T.-D. Truong, C. N. Duong, M.-T. Tran, and K. Luu, "Dyglip: A dynamic graph model with link prediction for accurate multi-camera multiple object tracking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 779–13 788. DOI: 10.1109/CVPR46437.2021.01357.

[70] 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. DOI: 10.48550/arXiv.2308.13229.

[71] J. Komorowski, G. Kurzejamski, and G. Sarwas, "Footandball: Integrated player and ball detector," in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 5, SciTePress, 2020, pp. 47–56. DOI: 10.5220/0008916000470056.

[72] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," vol. 3, pp. 115–143, 2002. DOI: 10.1162/153244303768966139.

[73] K. Kurach, A. Raichuk, P. Stanczyk, M. Zajac, O. Bachem, L. Espeholt, C. Riquelme, D. Vincent, M. Michalski, O. Bousquet, and S. Gelly, *Google research football: A novel reinforcement learning environment*, 2019. DOI: 10.48550/arXiv.1907.11180.

[74] S. Wang, H. Sheng, Y. Zhang, D. Yang, J. Shen, and R. Chen, "Blockchain-empowered distributed multi-camera multi-target tracking in edge computing," pp. 1–10, 2023. DOI: 10.1109/TII.2023.3261890.

[75] M. Nikodem, M. Słabicki, T. Surmacz, P. Mrówka, and C. Dołęga, "Multi-camera vehicle tracking using edge computing and low-power communication," vol. 20, no. 11, 2020. DOI: 10.3390/s20113334.

[76] E. Zemene and M. Pelillo, *Interactive image segmentation using constrained dominant sets*, 2016. DOI: 10.48550/arXiv.1608.00641.

[77] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, 2017. DOI: 10.48550/arXiv.1706.03762.

[78] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, *End-to-end object detection with transformers*, 2020. DOI: 10.48550/arXiv.2005.12872.

[79] F. Zeng, B. Dong, T. Wang, C. Chen, X. Zhang, and Y. Wei, *MOTR: end-to-end multiple-object tracking with transformer*, 2021. DOI: 10.48550/arXiv.2105.03247.

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

[81] P. Sun, Y. Jiang, R. Zhang, E. Xie, J. Cao, X. Hu, T. Kong, Z. Yuan, C. Wang, and P. Luo, *Transtrack: Multiple-object tracking with transformer*, 2020. DOI: 10.48550/arXiv.2012.15460.

[82] 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. DOI: 10.48550/arXiv.2306.17000.

[83] 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*, vol. 11209, 2019, pp. 379–396. DOI: 10.1007/978-3-030-01228-1_23.

[84] Y. Hou and L. Zheng, *Multiview detection with shadow transformer (and view-coherent data augmentation)*, 2021. DOI: 10.48550/arXiv.2108.05888.

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

[86] 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 *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 2022, pp. 489–499. DOI: 10.1109/WACVW54805.2022.00055.

[87] 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. DOI: 10.48550/arXiv.2310.13350.

[88] T. N. Kipf and M. Welling, *Variational graph auto-encoders*, 2016. DOI: 10.48550/arXiv.1611.07308.

[89] H. Hsu, Y. Wang, and J. Hwang, "Traffic-aware multi-camera tracking of vehicles based on reid and camera link model," Association for Computing Machinery, 2020, pp. 964–972. DOI: 10.1145/3394171.3413863.

[90] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," vol. 24, no. 5, pp. 603–619, 2002. DOI: 10.1109/34.1000236.

[91] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8856–8865. DOI: 10.1109/CVPR52688.2022.00866.

[92] X. Zhou, V. Koltun, and P. Krähenbühl, *Tracking objects as points*, 2020. DOI: 10.48550/arXiv.2004.01177.

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

[94] 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 *International Conference on Consumer Electronics - Taiwan*, 2023, pp. 629–630. DOI: 10.1109/ICCE-Taiwan58799.2023.10227006.

[95] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, *Observation-centric sort: Rethinking sort for robust multi-object tracking*, 2023. DOI: 10.48550/arXiv.2203.14360.

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

[97] H.-W. Huang, C.-Y. Yang, Z. Jiang, P.-K. Kim, K. Lee, K. Kim, S. Ramkumar, C. Mullapudi, I.-S. Jang, C.-I. Huang, and J.-N. Hwang, *Enhancing multi-camera people tracking with anchor-guided clustering and spatio-temporal consistency id re-assignment*, 2023. DOI: 10.48550/arXiv.2304.09471.

[98] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, *Bot-sort: Robust associations multi-pedestrian tracking*, 2022. DOI: 10.48550/arXiv.2206.14651.

[99] Z. Yang, Y. Wei, and Y. Yang, *Associating objects with transformers for video object segmentation*, 2021. DOI: 10.48550/arXiv.2106.02638.

[100] Z. Yang and Y. Yang, *Decoupling features in hierarchical propagation for video object segmentation*, 2022. DOI: 10.48550/arXiv.2210.09782.

[101] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3714–3720. DOI: 10.1109/CVPR.2013.476.

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

[103] 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).

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