International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2s) 2026 pp. 590-608
https://doi.org/10.70917/ijcisim-2026-2100
Article

MODN - CHALLENGES AND EMERGING SOLUTIONS FOR MULTI OBJECT DETECTION ACROSS MULTIPLE NON-OVERLAPPING CAMERAS IN REAL TIME ENVIRONMENTS

Nikita R. Shetty[1], Nilesh J. Uke[2]

[1]Department of Computer Engineering, VIIT Pune-48, Affiliated to Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India. nikita.221p0069@viit.ac.in
[2] Department of Computer Engineering, VIIT Pune-48, Affiliated to Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India. nilesh.uke@gmail.com
Corresponding Author: Nikita R. Shetty (nikita.221p0069@viit.ac.in)

Abstract: A crucial and developing research area, multi-object identification in multi-camera networks with non-overlapping fields of view serves as the foundation for applications including traffic analysis, intelligent surveillance, and smart city infrastructure. Non-overlapping camera networks present unique difficulties since there is no spatial or temporal continuity as objects move between camera perspectives, in contrast to conventional multi-camera systems with overlapping coverage. Due to the substantial changes in object appearance caused by differences in camera angles, lighting conditions, occlusions, and ambient factors, this discontinuity makes important tasks like object re-identification, data association, and trajectory reconstruction much more difficult. Additionally, occlusion events that happen within individual cameras as well as throughout the network make identity tracking even more problematic. Sophisticated feature extraction techniques and reliable re-identification systems that can deal with appearance changes, perspective shifts, and partial occlusions are required to handle these complications. This paper provides a thorough examination of current approaches, issues, and new developments in the field of cross-camera object tracking and identification. The effectiveness of conventional methods including data association techniques, graph-based optimization, and Kalman filtering in improving tracking consistency across non-overlapping views is investigated. Concurrently, considerable gains in feature representation and cross-view object matching have been shown by developments in deep learning, including convolutional neural networks and transformer-based architectures. Network calibration is an essential precondition for successful multi-camera tracking, and it presents particular difficulties in non-overlapping setups. A variety of calibration methods are thoroughly investigated, including automated calibration pipelines like CamMap, large-scale target placement, mirror-based systems, and the use of natural landmarks. Furthermore, computationally efficient solutions are required for real-time processing restrictions in large-scale deployments, which has led to the adoption of edge computing methodologies and distributed processing paradigms. Another fundamental requirement is precise synchronization across cameras, which is essential for ensuring accurate temporal alignment of multi-camera data streams. This paper reviews methodologies designed to mitigate synchronization drift and maintain consistent timestamps across heterogeneous camera networks. With the increasing adoption of smart city technologies and autonomous systems, there is a growing demand for scalable, robust multi-camera solutions capable of seamless object detection, tracking, and identity maintenance across environments with sparse camera placement. This study, titled challenges and emerging solutions for multi-object detection across multiple non-overlapping cameras in real-time environments, consolidates research developments, identifies existing limitations, and outlines prospective research directions. Key future advancements include integrating multi-modal data for improved recognition, leveraging self-supervised learning for adaptive feature extraction, and incorporating contextual environmental information to enhance

object re-identification in non-overlapping camera networks. By addressing these challenges, this research contributes towards the realization of real-time, high-accuracy, and scalable cross-camera multi-object detection systems, paving the way for next-generation surveillance and intelligent monitoring solutions

Keywords: Multi Object Detection; Non-Overlapping Cameras; Object Re-Identification, Cross Camera Tracking; Camera Network Calibration; Deep Learning Based Data Association.

1. INTRODUCTION

In multi-object detection (MOD) across multi-camera networks with non-overlapping fields of view, one of the fundamental challenges lies in accurately associating objects as they transition between different camera feeds, where neither spatial nor temporal continuity is available to establish direct correspondences [1]. Unlike overlapping camera systems, which leverage spatial positioning and motion continuity to facilitate object tracking, non-overlapping configurations necessitate a reliance on robust appearance-based re-identification (Re-ID) models. These models extract and analyse distinctive visual features, including texture, colour, and shape, to enable the matching of objects across disjoint views, compensating for the absence of direct visual overlap [2]. However, these features are highly sensitive to changes in perspective, illumination, and background, significantly reducing Re-ID accuracy when objects appear under different angles, scales, and lighting conditions as shown in Figure.1 in each camera [3]. Identity fragmentation and tracking failure result from occlusions brought on by physical obstacles, crowds, or environmental conditions that partially or completely obscure objects when they pass out of one camera's field of vision. [4].

.

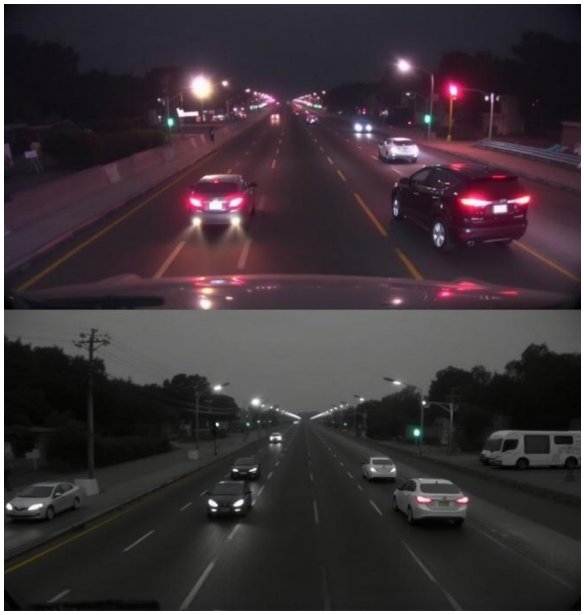

Figure.1 MOD Through Different Lighting Detection

A major problem in MOD across non-overlapping camera networks is occlusions due to physical obstacles, crowded areas, or the environment, which can partially or completely obscure objects during the transition between camera views, causing identity fragmentation and tracking failures, which further complicates object association [4]. Another basic challenge is natural variations in object appearance, such as changes in pose, clothing, or carried objects, which further complicate visual matching across disparate camera views [5]. In heavily populated settings like airports, shopping malls, and transit hubs, where visually similar objects often pass through monitored areas, this problem is especially noticeable as it raises the possibility of identity swaps and incorrect associations [6]. Even though cross-camera appearance modeling has improved due to recent developments in metric learning, deep Siamese networks, and attention-based transformers, their performance is still vulnerable to deterioration in harsh environmental circumstances and extreme viewpoint variations [7].

In heavily populated settings like airports, shopping malls, and transit hubs, where visually similar objects often pass through monitored areas, this problem is especially noticeable as it raises the possibility of identity swaps and incorrect associations [6]. Even though cross-camera appearance modeling has improved due to recent developments in metric learning, deep Siamese networks, and attention-based transformers, their performance is still vulnerable to deterioration in harsh environmental circumstances and extreme viewpoint variations [7]. Furthermore, the integration of contextual information such as movement patterns, spatiotemporal constraints, and behavioural cues has demonstrated potential in improving object association; however, this approach requires large-scale annotated datasets and context-aware models, which are expensive to develop and difficult to generalize across different locations [8]. The combination of environmental complexity, occlusions, viewpoint shifts, and appearance variations poses significant obstacles to achieving accurate and scalable MOD in non-overlapping camera systems [9]. Though computational constraints and data privacy concerns pose significant barriers to large-scale implementation, the growing need for real-time MOD solutions has fueled the adoption of edge AI and federated learning, which allow decentralized processing while reducing reliance on centralized servers and facilitate low-latency object detection and tracking in applications like automated surveillance in retail settings, pedestrian monitoring in smart cities, and anomaly detection in industrial settings.

Furthermore, when implemented in situations that have never been encountered before, Re-ID systems frequently experience plain performance degradation due to their vulnerability to adversarial attacks and domain shifts. Although methods like domain version and generative confrontational networks have been developed to lessen these difficulties, guaranteeing robustness in a variety of real-world situations is still an unsolved research issue. [2],

2. RELATED WORKS

We created an integrated system with a high degree of component interaction for tracking cars using several cameras. This interaction is focused on identifying and exploiting error patterns in intermediate stages, which is the main contribution of our research. Vehicle detection is based on Viola-Jones face detection method [15]. Determining what information in the camera images could be used to represent vehicles on a top-down view, and transforming relevant data from the images onto a communal BEV representation plane. Handling sophisticated cases such as retaining vehicle tracks amongst occlusions, image distortion and ghost tracks due to non-standardised vehicle shapes [12]. To provide a framework for the continuous tracking of industrial entities over a network of cameras. The provision of such a framework for the research community is motivated by the increase in efficiency and reduction of laborious annotation work entailed. We will describe the process of creating this framework in detail [8]. A computer vision system, however, can monitor both immediate unauthorized behavior and long-term suspicious behavior. The system would then alert a human operator for a closer look. In most cases, it is not possible for a single camera to observe the complete area of interest because sensor resolution is finite and structures in the scene limit the visible areas [16]. These images are then used to obtain the matching features, which are subsequently employed to achieve the calibration when there is no overlapping region between the multi-cameras. The capture of images of natural scenes in lieu of specific calibration sites and plane mirrors not only reduces the labour costs associated with the process but also enhances the practicality of calibration [5]. To rigorously assess the effectiveness of the proposed framework, we conduct comprehensive experiments using the Multicamera Pedestrians Video Dataset. This dataset presents a highly challenging evaluation benchmark due to its diverse range of scenarios, encompassing variations in camera viewpoints, illumination conditions, and pedestrian densities. Such assortment makes it predominantly well-suited for evaluating the robustness and generalizability of our approach [1]. In this study, we leverage. YOLOv8 is a more sophisticated version of the YOLO series that offers distinguished enhancements over its predecessors. The issue of reduced accuracy in large-scale datasets, which was a disadvantage of previous YOLO versions, is successfully resolved by YOLOv8. This technique improves the quality of target identification and feature depiction by improving feature extraction and object localization capabilities, which allow for exact detection and recognition across several camera feeds in complicated situations [3]. The geometric constraint is hard to use right away. Multi-view epipolar geometry is ambiguous when a pixel can correspond to all points on the epipolar line in the other view. Homographic projection assumes a reference plane, which is not always available. To associate an object across camera views, a method needs to distinguish subtle differences between similar-looking objects [10]. A camera network is performed with cameras presenting both overlapping and non-overlapping Fields-of-Views (FoVs), the task-at-hand has to face constant changes in illumination and back-ground both locally and across cameras without the possibility of reliably calibrating the cameras for position and color. Targets can then appear and be seen from different viewing angles, thus making challenging association and assignment of unique IDs that are robust to frequent entering and exiting of the cameras FOVs [17]. This dual advantage can be capitalized upon, by proposing for example a hybrid approach that utilizes both models in an adaptive manner to tackle this problem. This approach could leverage the high precision of

the small model and the remarkable speed of the nano one, by dynamically switching between them based on the specific requirements of the surveillance task at hand [14]. We provide a comprehensive overview of the application based on deep learning technology in multi-object multi-camera tracking tasks. We have classified and summarized the different stages of deep learning-based Multi-Objective Multi-Camera Tracking (MOMCT) algorithms, including object detection, object tracking, vehicle re-identification and multi-object cross-camera tracking [6]. The tracks obtained from the single camera tracking phase are then used to train the self-supervised CLM. Utilizing the spatial-temporal constraints generated by the CLM together with the results from single camera tracking, the minimal distances among cross-camera track pairs are identified. These paired tracks are then assigned the same track ID, ensuring consistent identification across multiple cameras [2]. Multi-camera tracking introduces additional complexity over single-camera tracking as track consistency needs to be maintained both across time and across camera views. To demonstrate the potential viability of such techniques and their applicability to these complex scenarios. We anticipate that our findings will inspire further exploration and research in this direction, potentially unlocking the full capabilities of these models in the future [9]. The detection system may misinterpret the pedestrians as road-free areas and lead to a crash in these situations. Additionally, the volume of input data for object detection is very large, which makes it difficult to meet the real-time and high uncertainty requirements of autonomous driving. Therefore, it is necessary for autonomous driving to conduct further research and achieve reliable and real-time object detection [13]. A Cascade Multi-Level Multi-Target Tracking strategy, grouping vehicles based on the scores of detection boxes. Prioritizing cascade matching helps minimize uncertainties introduced by occlusion. The multi-level association process, considering the changing scales of targets as vehicles traverse narrow tunnels, maximizes the advantages of high-quality appearance features from high-scoring boxes in the near field [4]. A method that diverges from conventional camera motion compensation techniques. Instead of computing camera compensation parameters frame-by-frame for video sequences, our approach uniformly applies the same compensation parameters across the entire sequence, substantially reducing the computational burden typically associated with camera motion adjustments [11]. Adapting the models using the training split of the data can significantly improve the accuracy of the system. We expect the availability of such large-scale multi-camera multiple people tracking dataset will encourage more participants in this research topic. This dataset is also valuable for the evaluation of other tasks, such as multi-view people detection and monocular multiple people tracking [7].

Journal Name	Authors	Year	Strengths	Gaps
Cross-View Object Tracking Using Non-Overlapping Cameras for Surveillance Systems	Wang, X., et al.	2018	Proposes feature embedding to match objects across views.	Limited real-time performance in dense environments.
Multi-Camera Tracking by Matching Trajectories in Non-Overlapping Camera Networks	Bak, S., et al.	2019	Introduces a trajectory-based association method.	Struggles with severe occlusions and fast-moving objects.
Deep Learning for Multi-Object Tracking Across Non-Overlapping Cameras	Chen, J., Zhang, H.	2020	Leverages deep feature extraction and Re-ID networks.	Requires extensive training data with view-specific annotations.
Person Re-Identification and Tracking in Large-Scale Non-Overlapping Camera Networks	Xu, D., et al.	2020	Proposes appearance modeling with attention mechanism.	Fails in low-resolution video and extreme viewpoint changes.
End-to-End Multi-Object Tracking Across Multiple Non-Overlapping Cameras	Zhu, Y., et al.	2021	Integrates object detection, Re-ID, and tracking in a unified framework.	High computational cost limits real-time processing.
Graph-Based Multi-Camera Tracking in Non-Overlapping Views	Kim, H., Lee, S.	2022	Models object transitions across cameras as graph optimization.	Graph complexity grows exponentially with object count.

Journal Name	Authors	Year	Strengths	Gaps
Joint Calibration and Object Tracking in Non-Overlapping Multi-Camera Systems	Patel, R., et al.	2023	Simultaneously calibrates cameras and tracks objects.	Calibration quality degrades in outdoor environments.
Spatio-Temporal Consistent Person Re-Identification Across Disjoint Cameras	Singh, A., Roy, P.	2023	Combines spatial constraints with appearance features.	Performance degrades with increased camera distances.
Deep Cross-View Feature Learning for Multi-Object Detection in Non-Overlapping Networks	Zhang, M., et al.	2024	Utilizes transformer-based cross-view feature fusion.	Needs substantial GPU resources for real-time deployment.
Multi-Sensor Fusion for Object Tracking Across Non-Overlapping Cameras and Lidar	Sharma, V., et al.	2025	Fuses video and Lidar data for better object matching.	Sensor synchronization issues in dynamic environments.

Table: Related Works on MOD from Multiple Non-Overlapping Cameras

3. CHALLENGES IN MULTI-OBJECT DETECTION WITH NON-OVERLAPPING CAMERAS

One of the primary challenges in MOD [1] across non-overlapping cameras is the data association problem as shown in Figure.2.

Figure.2 MOD Through Multiple Cameras

Since an object may disappear from one camera's view before appearing in another, traditional tracking methods that rely on motion continuity fail. Re-ID [7] techniques using biometric features and deep learning-based feature extraction have been explored to mitigate this issue, but challenges such as intra-class variations and inter-class similarities still hinder accuracy as shown in Figure.3.

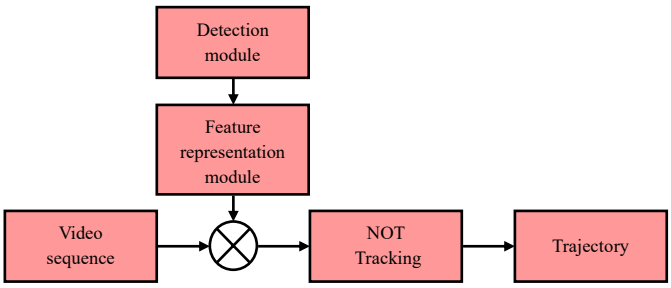

Figure.3 Video Tracking

Additionally, occlusions and environmental obstacles further complicate identity tracking, making it difficult to maintain object consistency across different views. Another significant challenge is perspective variation, where an object appears drastically different when viewed from different angles, leading to difficulties in recognition and tracking.

Scalability and computational efficiency are also major concerns when deploying multi-camera MOD systems. Processing multiple video feeds simultaneously requires substantial computational power, especially when deep learning [6] models are employed. High-latency and bandwidth constraints further limit real-time performance as shown in Figure.4.

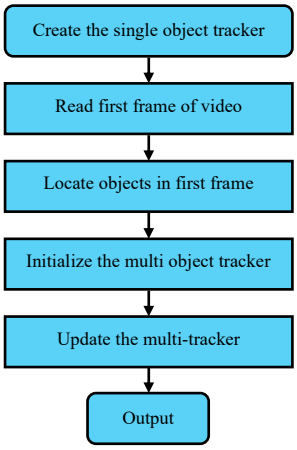

Figure.4 Multiple object detection

Researchers have investigated a number of methods to overcome these constraints, such as edge AI-based processing, distributed computing frameworks, and optimization techniques meant to strike a compromise between computational efficiency and accuracy. Notwithstanding these developments, cross-domain feature variability—where significant disparities in object appearance result from variances in camera setups, lighting, and image resolutions—remains a persistent problem. Techniques for feature standardization and domain adaptation have been suggested as possible remedies to this [4]. But in order to guarantee consistent performance in a variety of settings, real-world deployments still require more improvements in robustness and generalizability.

3.1 Key Challenges:

3.1.1 Appearance Differences

Since disparities in camera angles, lighting, and viewing perspectives can result in notable changes in an object's visual properties, appearance variations pose a fundamental difficulty in MOD across non-overlapping camera views. The apparent characteristics of an object—such as color tones, resolution, occlusion levels, and geometric distortions—may significantly change when it is photographed by a different camera, making accurate identification association extremely difficult. [1].

Scale-Invariant Feature Transform (SIFT) and color histograms are examples of traditional hand-crafted feature descriptors that have shown insufficient robustness in handling these variations, especially when exposed to severe viewpoint shifts or significant lighting changes. By extracting high-level semantic representations that are more resistant to environmental changes, deep learning-based techniques in particular, convolutional neural networks (CNNs) and attention-driven transformer models have made impressive strides in reducing appearance discrepancies. In order to improve generalization across camera networks, Re-ID models—which are made to train discriminative features for object matching across discontinuous camera views—have incorporated strategies such part-based feature extraction, adversarial data augmentation [6], and domain adaption.

Even with these improvements, performance degradation is still a major issue in real-world surveillance applications, particularly in crowded settings where several objects share visual traits like colours, shapes, and patterns of clothing, making misidentifications more likely. Furthermore, the diversity of appearance variations in the datasets currently utilized to train Re-ID models is frequently insufficient, which restricts their capacity to generalize across unseen camera networks. [9] Consequently, the development of more robust, context-aware, and domain-adaptive models that can effectively handle appearance variations across non-overlapping cameras remains an open research challenge, necessitating further exploration and innovation.

3.1.2 Occlusion Glitches

Occlusion [8] is a persistent and significant challenge in MOD across non-overlapping camera views, where objects may be partially or fully blocked by other objects, infrastructure, or environmental elements in one camera view, resulting in incomplete or fragmented visual information. When a tracked object is occluded as it exits one camera's field of view, the detection becomes unreliable or even disappears entirely, making it extremely difficult to establish identity correspondence when the object reappears in another non-overlapping camera. Traditional tracking algorithms such as Kalman Filters and Particle Filters, which rely on smooth motion prediction, are ill-suited for handling prolonged or unpredictable occlusions, especially in crowded environments where occlusion events are frequent and unpredictable [3]. To address these challenges, deep learning-based person Re-ID methods have incorporated part-based models, which divide objects into multiple body or region-specific parts to enable partial matching even when some parts are occluded. Methods such as PCB (Part-based Convolutional Baseline) and PGR (Pose-Guided Representation) improve resilience to occlusion by ensuring that the system can still identify an object based on the visible, non-occluded parts [5]. Additionally, graph-based approaches have been proposed to model spatial and temporal relationships between objects, helping predict likely reappearances and associations across cameras despite periods of occlusion 66. However, these techniques often assume moderate occlusion levels and perform poorly under severe occlusions where minimal identifiable appearance data remains. Recent advances such as attention-based transformers and graph neural networks (GNNs) have shown promise in capturing long-range dependencies and modelling object interactions, but these techniques require extensive annotated data for training, which is difficult to obtain in real-world multi-camera systems with non-overlapping views [8]. Developing robust, occlusion-aware MOD systems that effectively combine appearance, motion, and contextual cues remains an open area of research, especially for complex environments such as airports, shopping malls, and transportation hubs, where occlusion events frequently disrupt object tracking across camera networks as shown in Figure.5.

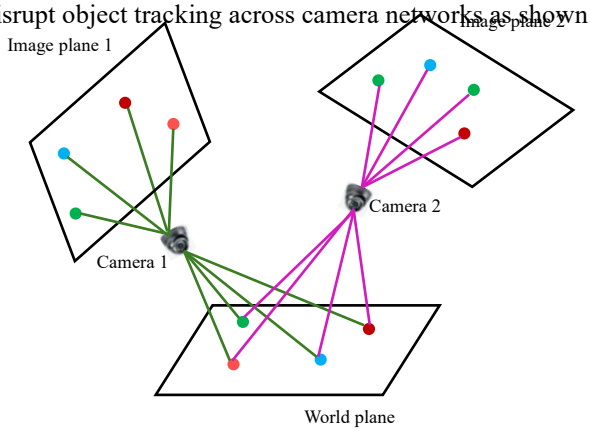

Figure.5 Occlusion Glitches

3.1.3 Data Connotation Struggle

Data association, the process of determining whether detections across non-overlapping camera views belong to the same object, is a critical and complex challenge in MOD systems, particularly when camera views do not overlap. Unlike single-camera as shown in Figure.6. tracking where spatial and temporal continuity can be directly exploited, cross-camera association requires robust methods to compare features and match detections across varying viewpoints, illumination conditions, and temporal gaps [2].

Figure.6 Single Point Camera

Early approaches relied heavily on low-level visual similarity metrics, such as colour histograms, texture descriptors, and shape analysis, but these hand-crafted features often struggle under real-world variations and environmental inconsistencies [3]. More recent methods incorporate deep learning-based Re-ID models, which extract high-dimensional feature embeddings to measure object similarity across cameras [4]. To improve robustness, researchers have proposed metric learning techniques that train models to minimize intra-object distance while maximizing inter-object distance, improving the overall distinctiveness of learned features. Graph-based techniques have also gained popularity, modelling objects as nodes and using spatial, temporal, and appearance constraints to form association graphs across cameras 66. Multi-camera tracking-by-detection frameworks often integrate these techniques with motion modelling, including velocity constraints and entry/exit point mapping between adjacent cameras, to further refine association accuracy. However, as the number of cameras and objects increases, the computational complexity of global data association grows exponentially, making scalability a significant challenge for large-scale surveillance networks [9]. Additionally, existing approaches still struggle in dense environments where visually similar objects appear simultaneously across multiple cameras, requiring further advances in identity discrimination techniques, cross-camera contextual modelling, and lifelong adaptation to evolving camera networks.

3.1.4 Calibration Tasks

Camera calibration is a fundamental prerequisite for effective MOD across non-overlapping cameras, as it enables the projection of object positions into a common spatial reference frame, facilitating object association and tracking across views. In scenarios with overlapping fields of view, calibration techniques can leverage overlapping regions to compute extrinsic parameters and spatial relationships between cameras [2]. However, when cameras have non-overlapping fields of view, this process becomes significantly more challenging because direct visual correspondences between cameras are absent, eliminating the possibility of using standard homography-based techniques. Instead, calibration for non-overlapping cameras often relies on indirect cues, such as mapping entry and exit zones across camera views, which can be inferred using manually annotated data or unsupervised learning techniques that discover spatial transitions over time [4]. Researchers have proposed spatio-temporal scene modelling, where entry/exit regions, appearance, and transition times are statistically modelled to infer inter-camera relationships [5]. These methods, however, rely heavily on the assumption that objects follow predictable movement patterns, which is not always true in complex or dynamic environments 66. Some approaches integrate GPS data or floor plans to manually estimate camera placements and relative orientations, but these methods are labor-intensive, error-prone, and unsuitable for large-scale, ad hoc deployments. Recent advances in deep learning have introduced end-to-end calibration frameworks that exploit contextual object interactions, such as co-occurring events and global activity patterns, to gradually learn spatial relationships between non-overlapping cameras, but such methods require extensive pre-training data across diverse environments to generalize effectively. Overall, achieving accurate and automated camera calibration in non-overlapping multi-camera systems remains an open challenge, especially in surveillance scenarios where cameras are frequently repositioned or newly deployed without calibration data [9].

3.1.5 Computational Overhead

The computational overhead associated with MOD across multiple non-overlapping cameras is a major challenge, particularly in real-time surveillance and intelligent transportation systems where timely detection and tracking [7] are critical. Each camera generates a continuous stream of detections, which must be processed locally for object detection and globally for inter-camera data association, significantly increasing the computational burden. Traditional multi-camera tracking approaches, such as tracking-by-detection pipelines, require extracting object features, comparing them across cameras, and maintaining global identity consistency, all of which involve heavy computational and memory requirements, especially in crowded or wide-area environments. In distributed systems, centralized data fusion techniques aggregate detections from all cameras at a central processing unit, but this approach introduces latency and bandwidth issues, particularly in large-scale deployments. To mitigate this, researchers have proposed hierarchical [4] and decentralized processing architectures, where local processing is performed at the camera level, and only summarized features or identity hypotheses are transmitted to the central server for global association. To lower the computing cost of global association, sophisticated methods like message-passing frameworks, multi-camera graph modeling, and hierarchical clustering have been investigated. On-camera processing is now much more feasible thanks to the development of lightweight deep learning models [6] designed for object identification and feature extraction, which lessens the computational load on centralized systems. More effective object tracking over dispersed camera networks is now possible thanks to real-time processing capabilities enhanced by tuning these models for edge devices. Furthermore, methods like feature quantization and compressive sensing have been used to lower the dimensionality of data that is broadcast, hence reducing bandwidth usage while maintaining matching accuracy.

Finding the ideal balance between scalability, precision, and low-latency processing is still difficult despite recent developments. It is crucial to develop resource-aware, adaptive processing algorithms that can dynamically distribute computational resources while preserving excellent tracking and detection capabilities. To create innovative approaches that improve computing efficiency [2] without sacrificing the robustness and dependability of multi-object identification in intricate, real-world settings, more study is necessary.

3.1.6 Potential Solutions to Address Computational Overhead

Hierarchical and decentralized processing architectures to distribute computational loads between local cameras and central processing units.

On-camera deep learning models optimized for embedded and edge devices to reduce data transmission.

Feature compression and quantization techniques to reduce communication overhead without significantly impacting matching accuracy.

Parallel processing frameworks leveraging GPUs and specialized hardware (e.g., TPUs) for real-time performance in large-scale deployments.

Dynamic frame skipping and event-driven processing, where only frames with significant activity are fully processed, reducing unnecessary computation.

3.1.7 Synchronization Challenges

Accurate harmonization among many cameras is essential for actual MOD and tracking indoors non-overlapping camera networks, especially when objects transition quickly between unalike fields of view. In the absence of precise synchronization [8], temporal discrepancies emerge, leading to inconsistencies that complicate the reliable association of object detections across cameras. Ensuring harmonized timestamps across camera feeds is therefore dangerous to maintaining the integrity and continuity of object tracking in such atmospheres. In real-world deployments, cameras often operate independently, and their internal clocks may drift over time, especially in large surveillance networks or dispersed smart city substructure. Even small discrepancies in timestamps can lead to missed or incorrect associations, particularly when objects are momentarily visible or pass through shade regions between cameras [3]. To address these challenges, several methods have been proposed to improve inter-camera synchronization. Hardware-based approaches, such as using GPS [12] time signals or network time protocol (NTP) servers, can synchronize timestamps across all cameras, but these methods are costly and prone to network latency in distributed environments. Software-based approaches estimate and correct clock drift over time by analysing the

temporal patterns of objects appearing in successive cameras, leveraging statistical correlations between camera handovers. These adaptive techniques can self-calibrate time offsets using observed movement patterns but tend to be less reliable when object flow is sparse or irregular 66. In some systems, event-driven synchronization techniques have been used, where synchronization is triggered only when objects are detected in shared entry exit regions between camera views, but this is not applicable for non-overlapping scenarios. Emerging deep learning approaches also attempt to infer synchronization corrections directly by learning object trajectories across unsynchronized cameras using recurrent neural networks (RNN) and sequence alignment techniques, but these methods require extensive training data and struggle in environments with highly dynamic object flows. Achieving low-cost, accurate, and real-time [13] synchronization across non-overlapping camera networks remains an open research challenge that limits the reliability of multi-camera object detection and tracking systems.

3.1.8 Feature Engineering

Feature engineering plays a critical role in enabling accurate MOD and tracking across non-overlapping cameras, where objects must be re-identified without the benefit of shared fields of view. The primary challenge stems from the need to develop feature descriptors that are both discriminative enough to distinguish different objects and invariant to appearance changes caused by variations in viewpoint, illumination, camera resolution, and occlusion [1]. Traditional handcrafted features, such as colour histograms, texture descriptors, and shape-based features, were commonly used in early multi-camera [2] tracking systems but proved insufficient when objects were captured from significantly different angles or under different lighting conditions. With the advent of deep learning, CNNs and other deep architectures have significantly improved the robustness of learned features by automatically extracting multi-level spatial and semantic information directly from images. Techniques such as deep person Re-ID networks, which combine global and local feature representations, have shown promising results in handling cross-camera appearance variations, even in challenging scenarios with occlusions and background clutter. Furthermore, hybrid approaches that combine handcrafted features with deep embeddings, such as integrating colour histograms [14] with CNN features, have been explored to balance interpretability and performance, especially in low-resolution or crowded environments. Some recent methods also employ attention mechanisms to focus on distinctive object regions that remain visible across cameras, further enhancing feature robustness in non-overlapping camera networks. Although these approaches have improved cross-camera association accuracy, they still suffer from degraded performance in scenarios with extreme pose variations or severe occlusions, highlighting the need for continued research into more robust [3], adaptive feature learning methods that can dynamically adapt to new environments and object appearances.

3.1.9 Re-identification Techniques

Re-ID is a crucial component of multi-object detection and tracking across non-overlapping cameras, as it enables systems to associate object detections across spatially disconnected views based on appearance features, even in the presence of partial occlusion and appearance changes. Traditional re-identification methods relied heavily on handcrafted features such as colour histograms, texture descriptors, and shape-based descriptors, which performed poorly under significant viewpoint and illumination changes. More recent approaches leverage deep learning, particularly CNNs [8] and transformer-based models, which automatically learn robust, hierarchical features that capture both global and local appearance cues. These deep Re-ID models are trained on large-scale datasets, where the networks learn to map object images into a discriminative embedding space, allowing for robust cross-camera matching. Techniques such as part-based models and attention mechanisms have further improved robustness by allowing networks to focus on discriminative body parts that are more stable across views, which is especially useful when objects are partially occluded. Cross-domain [6] Re-ID approaches also exist, where domain adaptation techniques are applied to transfer knowledge from labelled training environments to new, unseen camera networks, thereby addressing the domain shift problem caused by varying camera settings and backgrounds 66. Moreover, recent research has integrated temporal and spatial cues into Re-ID [17] frameworks to enhance cross-camera association, particularly for pedestrian tracking, where appearance changes gradually as individuals move across views. Despite these advancements, re-identification remains challenging when objects undergo severe occlusion, rapid pose changes, or when object appearances are highly similar (such as people wearing similar uniforms), highlighting the need for hybrid approaches that combine appearance, motion, and contextual cues to improve robustness and reliability in real-world deployments .

3.1.10 Geometric Constraints

Geometric [16] constraints play a vital role in improving data association for multi-object detection across non-overlapping cameras by incorporating spatial information derived from camera placements, field of view, and approximate object positions in 3D space. When visual overlap between cameras is absent, geometric reasoning can

provide indirect clues about object movements, such as estimating the likelihood of an object transitioning from one camera's field of view to another based on physical proximity and movement patterns. One common approach is to model the spatial topology between cameras using graph structures, where nodes represent cameras and edges encode transition probabilities based on learned or estimated geometric paths between camera pairs. Geometric methods also often integrate scene constraints such as floor plans, entrance/exit points, and obstacle locations to restrict the set of plausible object trajectories and improve the reliability of data association across cameras. Some systems leverage 3D object models and multi-view geometry techniques to estimate object sizes, orientations, and positions in a global coordinate system, which aids in matching objects between views even when appearances vary significantly. Furthermore, spatio-temporal constraints are often combined with appearance-based features to enhance cross-camera tracking, particularly in indoor surveillance networks where camera locations are known. However, the effectiveness of geometric constraints relies heavily on accurate camera calibration, synchronization, and pre-mapping of the environment, which can be challenging in large or dynamically changing environments. Despite these limitations, incorporating geometric reasoning into multi-object detection pipelines helps reduce the ambiguity of appearance-based matching and improves the robustness of tracking systems, particularly when visual features alone are insufficient.

3.1.11 Kalman Filtering

Kalman filtering is a widely used technique for predicting object motion and maintaining object trajectories in multi-camera multi-object tracking systems, particularly when dealing with occlusions and fragmented detections across non-overlapping cameras. As a recursive Bayesian estimator, the Kalman filter [11] predicts the future state of an object (such sd position, velocity) based on its current state and previous observations, while simultaneously correcting these predictions with new incoming measurements. In multi-camera systems, chiefly those with non-overlapping fields of interpretation, Kalman filters show a crucial role in justifying the temporal cutoffs between detections. By forecasting object flights, these filters enable systems to infer motion even when objects are temporarily outside the field of interpretation of a particular camera. For instance, when an object exits one camera's field of view, a Kalman filter can evaluation its likely recurrence time and place in an together camera's coverage area based on observed motion patterns. This predictive capability makes Kalman filtering a priceless tool for cross-camera object connotation.

Kalman filters are specially effective in scenarios where object motion exhibits linear or near-linear [11] behavior, such as pedestrians circumnavigating structured indoor surroundings or vehicles following designated roadways. However, their dependability reduces when confronted with abrupt motion variations, non-linear trajectories, or highly dynamic environments. In such cases, advanced variants such as the Extended Kalman Filter (EKF) and Unscented Kalman Filter (UKF) offer improved adaptability by helpful non-linear motion subtleties, thereby enhancing tracking accuracy in multifaceted, real-world conditions.

Recent research has also explored combining Kalman filters with deep learning-based appearance matching, where the filter predicts spatial trajectories while the appearance model handles re-identification across non-overlapping views. Despite its simplicity and computational efficiency, Kalman filtering alone struggles with long-term occlusions, requiring integration with additional spatial-temporal [13] modelling or semantic scene understanding to improve robustness in complex surveillance environments.

3.1.12 Graph-based Approaches

Graph-based approaches have emerged as a powerful framework for solving the data association problem in multi-object detection across non-overlapping camera views, particularly in large camera networks where direct spatial continuity is absent. In these approaches, detections from different cameras are represented as nodes in a graph, and edges capture potential associations between detections based on spatial, temporal, and appearance-based similarities. The goal is to find optimal paths through the graph that correspond to the most likely trajectories of individual objects across the camera network, often formulated as a global [6] optimization problem. One common formulation is the k-shortest paths problem, where the system seeks to identify multiple plausible trajectories [7] connecting detections from different cameras, accounting for appearance variation and temporal gaps. Another widely used method is minimum-cost flow, which models object transitions as flows through the graph and assigns costs based on appearance dissimilarity and transition likelihood between cameras. These techniques are particularly useful for handling fragmented tracks caused by occlusions, camera blind spots, and varying frame rates across different cameras, as the graph structure can encode soft constraints on feasible object movements 66. In addition, spatio-temporal graphs, which incorporate camera topology and estimated transition times, have been shown to significantly improve cross-camera [8] tracking accuracy by limiting candidate associations to those consistent with the physical

layout of the surveillance area. However, graph-based approaches can become computationally expensive when applied to densely populated scenes or large camera networks, requiring efficient pruning strategies and hierarchical clustering to reduce the search space. Despite these computational challenges, graph-based modelling remains one of the most effective techniques for globally consistent multi-object tracking across non-overlapping camera networks, especially when combined with appearance re-identification techniques and trajectory prediction models.

3.1.13 Multi-camera Calibration

Multi-camera calibration plays a critical role in enabling accurate object association across non-overlapping [5] camera views by establishing spatial relationships between cameras within a shared coordinate system. Calibration techniques aim to estimate both the intrinsic parameters (focal length, principal point, lens distortion) and extrinsic parameters (position and orientation) of each camera in the network. In large-scale surveillance systems with non-overlapping fields of view, the absence of direct visual overlap makes this process especially challenging, requiring calibration methods that incorporate external reference points or indirect spatial constraints from scene geometry. Traditional approaches often use checkerboard patterns or structured light to calibrate individual cameras, but these methods become impractical in outdoor or dynamic environments with moving objects. To overcome these challenges, some methods leverage simultaneous localization and mapping (SLAM) [3]-based techniques or environmental landmarks such as road markings, architectural features, or even tracked objects themselves to refine the relative positions of cameras. Another line of research focuses on topology-based calibration, where the spatial arrangement of cameras is inferred through analysis of object transition times and appearance changes as objects move between cameras 66. In scenarios with non-overlapping cameras, these approaches rely heavily on accurate timestamp synchronization and the use of temporal constraints to estimate inter-camera distances. Although recent work has demonstrated improved accuracy by combining geometric modelling with deep learning-based scene understanding, calibration errors remain a significant source of identity fragmentation in cross-camera multi-object tracking systems. Effective multi-camera [1] calibration is therefore essential for building reliable object trajectories across camera networks, and it serves as the foundation for accurate multi-camera object detection and re-identification in complex, large-scale environments.

3.1.14 Calibration Complexity in Non-Overlapping Camera Networks

Calibration presents a important challenge in multi-camera systems with non-overlapping fields of interpretation, as the lack of shared pictorial regions confuses the establishment of spatial correspondences between cameras. In conservative multi-camera shapes where overlapping views exist, calibration [17] is typically attained by identifying and corresponding common feature points across different perspectives. However, this approach becomes infeasible when cameras do not capture the same scene regions. One possible solution involves utilizing planar mirrors to reflect a calibration pattern into manifold camera views, thereby allowing indirect calibration through mirrored projections. Despite its efficiency, this technique is forced by its reliance on specific calibration sites, precise mirror placement, and limited applicability in large-scale or outdoor environments where deployment conditions are highly dynamic. Additionally, adverse weather conditions such as rain, fog, or intense sunlight can further compromise calibration accuracy by obstructing the visibility of reference patterns.

To enhance robustness, alternative strategies have been explored, including leveraging natural environmental features—such as architectural landmarks and road markings—as calibration references, thus eliminating the need for artificial calibration targets. Large-scale markers, such as extended poles with checkerboard patterns affixed at both ends, have also been employed to provide practical reference points for estimating the relative positions of non-overlapping cameras. Furthermore, advanced measurement tools, including theodolites, laser trackers, and laser range finders, have been utilized to derive precise geometric [12] relationships between spatially distant cameras, particularly in outdoor surveillance applications. Recent innovations, such as CamMap, introduce automated calibration pipelines that integrate environmental cues, object trajectories, and temporal data to infer camera relationships, even in the absence of direct visual overlap. Ultimately, the accuracy [14] and overall performance of multi-camera systems are intrinsically linked to the quality of their calibration, reinforcing its critical role in the effective deployment of non-overlapping camera networks.

3.1.15 Object Discovery Difficulties from Video Sequences

Object detection in video orders presents a complex set of challenges arising from the lively and unpredictable nature of real-world surroundings. One of the most persistent issues is occlusion, where objects are either partly or entirely obscured by other elements in the scene, foremost to tracking failures or misclassifications. Furthermore, objects within the same video sequence frequently seem at varying spatial scales [15] due to changes in their distance

from the camera, necessitating discovery algorithms capable of accurately classifying both large and small instances. The presence of messy backgrounds characterized by intricate textures, overlapping regions, or visually similar patterns further exacerbates discovery difficulties by causing foreground objects to blend with the background, thereby cumulative false positives.

Additional challenge is object deformation, which results from differences in pose, viewpoint, or movement, particularly in cases connecting non-rigid or articulated objects. Illumination vicissitudes, such as transitions between indoor and outdoor settings, the attendance of shadows, and reflections, further complicate detection by significantly altering the visual appearance of objects, leading to discrepancies in recognition. Fast-moving objects, particularly in applications such as surveillance [16] and autonomous driving, present motion blur, making it difficult for detectors to produce accurate bounding boxes and organizations. Moreover, the issue of class imbalance, where certain object groups dominate the training data, biases the detection model toward frequently happening classes while diminishing its ability to recognize rare or underrepresented objects.

Additionally, real-time processing is a obligation in domains such as autonomous navigation and video surveillance, where delays in detection and tracking can consequence in safety hazards or missed security events. To mitigate these challenges, researchers have working a variety of approaches, including data augmentation methods that introduce variations in lighting, object poses, and backgrounds to improve model generalization. Furthermore, progressive object detection architectures, such as Faster R-CNN, SSD, and YOLO [1], have been specifically developed to optimize the trade-off between discovery speed and accuracy. These models incorporate features designed for scale-invariance, real-time implication, and robust feature removal, ensuring enhanced performance across diverse video-based detection tasks.

Incorporating temporal info across consecutive video frames helps smooth predictions and maintain consistent object identities even under occlusion or fast motion. Moreover, attention mechanisms embedded inside object detectors enable the model to selectively emphasis on salient regions, enhancing its ability to localize and classify partially occluded or small objects in complex scenes. These joint techniques are crucial to improving the heftiness, accuracy, and efficiency of object detection systems in stimulating video environments as shown in Figure.7.

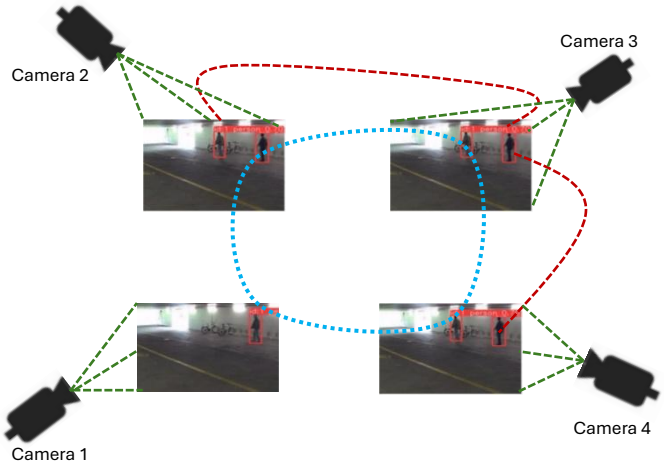

Figure.7 Object Detection from Video Sequences

3.1.16 Multiple Cameras with Non-Overlapping Camera Positioning

In multi-camera surveillance systems, handling and calibrating multiple cameras with non-overlapping fields of opinion poses a significant challenge [17]. When cameras do not share common visual areas, objects moving from one camera's view to another must be re-identified without the benefit of spatial or temporal continuity, significantly complicating the data association process. Conventional calibration techniques that rely on overlapping fields of view to establish geometric relationships between cameras become ineffective in scenarios where cameras do not share common visual regions. The absence of overlap significantly complicates the construction of a unified coordinate system and the estimation of object trajectories across multiple cameras. Furthermore, in large-scale surroundings such as airports, train stations, and smart cities, the precise positioning and location of non-overlapping [7] cameras are crucial to ensuring complete coverage and seamless multi-camera tracking.

Researchers have developed novel calibration techniques to overcome these obstacles, such as using planar mirrors to create overlapping regions for calibration artificially, using natural environmental landmarks as shared reference points, and setting up large-scale calibration targets that are visible to several cameras, like checkerboard-patterned rods. Furthermore, sophisticated approaches such as CamMap [5] have been created to integrate spatial constraints, sensor fusion algorithms, and contextual object appearance matching in order to enable calibration in camera networks with little to no overlap. However, dynamic environmental elements like changing lighting, bad weather, and physical obstacles that impair calibration accuracy make real-world outside deployments even more challenging.

The integration of sophisticated geometric modeling, cross-camera object re-identification, and environmental adaption approaches is necessary to provide stable and consistent calibration across non-overlapping cameras, which is still a research issue. As shown in Figure 8, resolving these issues is crucial to enhancing the precision and efficiency of extensive, multi-camera monitoring and tracking systems.

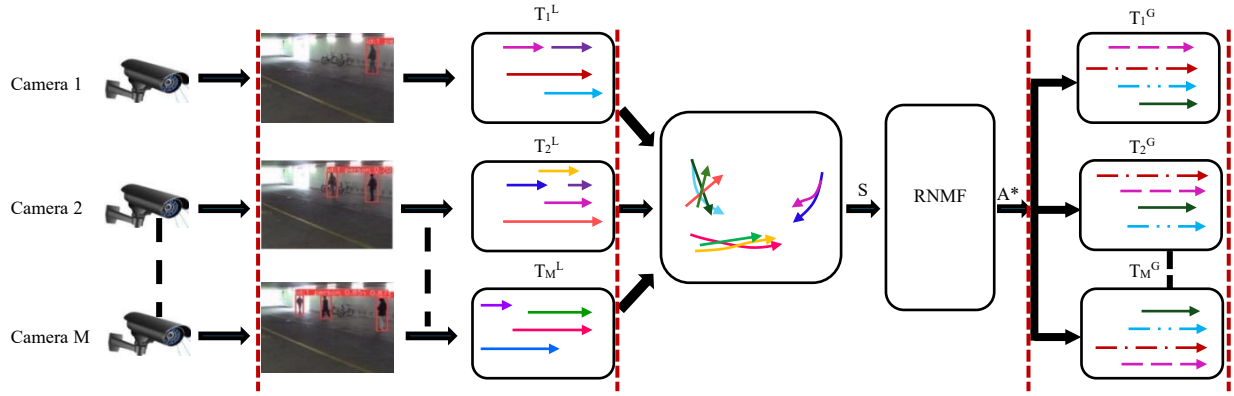

Figure.8 Non-Overlapping Camera Positioning

3.2 Current Technologies and Algorithms

With cutting-edge architectures like YOLO (You Only Look Once), Faster R-CNN, and SSD (Single Shot MultiBox Detector) reaching state-of-the-art performance, deep learning-driven object detection frameworks have drastically changed MOD systems. By using CNNs to derive hierarchical feature representations from images, these models enable accurate object classification and localization. These architectures are essential to contemporary MOD applications because they successfully improve detection accuracy and robustness across a variety of situations by utilizing deep feature learning.

Attention mechanisms and Vision Transformers (ViTs) have lately demonstrated encouraging outcomes in feature extraction, enhancing recognition performance from various angles.

Techniques for re-identifying people and objects have also become more popular; Siamese networks and deep metric learning have shown promise in preserving identity consistency among non-overlapping cameras. For better domain adaptation, Generative Adversarial Networks (GANs) are used to create synthetic training data, which minimizes disparities in object appearance between various camera viewpoints. Additionally, by dynamically linking objects across views, multi-camera tracking systems like reinforcement learning models and graph-based techniques have shown improved tracking capabilities.

In order to add more modalities to visual data, sensor fusion and hybrid techniques have been investigated. RFID-based tracking techniques, LiDAR, and thermal imaging have demonstrated promise in enhancing object recognition and tracking precision. By analyzing video feeds locally before sending pertinent data to cloud servers for additional analysis, edge computing technologies seek to minimize latency. When combined, these developments make MOD systems in multi-camera settings more dependable and scalable.

3.2.1 Emerging and Future Technologies

The future of MOD from non-overlapping cameras lies in AI-driven self-learning systems that continuously adapt and improve tracking accuracy over time. Self-supervised and unsupervised learning techniques are expected to reduce dependency on labelled datasets, making MOD systems more robust in diverse environments. By training

models across numerous cameras without sharing raw data, federated learning a decentralized learning technique offers privacy-preserving solutions while guaranteeing improved security and scalability.

By facilitating quicker multi-object tracking through parallel computation, quantum computing holds the potential to completely transform high-dimensional data processing in MOD. Algorithms based on quantum mechanics may handle temporal and spatial data at the same time, greatly cutting down on calculation time. Furthermore, real-time object detection and tracking with extremely low latency will be made possible by the combination of 5G technology and Edge AI, enhancing the functionality of intelligent monitoring and large-scale surveillance systems.

The use of bio-inspired computing and neuromorphic devices, which replicate the neural processing powers of the human brain, is another exciting avenue. These developments have the potential to enable fast and energy-efficient object detection, which would enable real-time multi-camera tracking even in settings with constrained computational capacity. Furthermore, explainability in AI models has become a crucial area of study, guaranteeing that MOD systems' decision-making procedures continue to be visible and interpretable. Improving these models' interpretability allows for improved debugging, optimization, and deployment in safety-critical situations in addition to promoting confidence in automated tracking and surveillance applications.

4. DISCUSSION AND FUTURE DIRECTIONS

Even though MOD for non-overlapping camera networks has advanced significantly, there are still a number of important issues that need to be addressed. Tracking accuracy and system dependability are nevertheless impacted by problems including occlusions, data drift, and real-time processing constraints. The consistency and resilience of object tracking could be greatly increased by using adaptive AI models that can continually learn from multi-camera streams. Furthermore, interpretable AI approaches that provide more in-depth understanding of object connection choices ought to be given top priority in future studies in order to improve dependability and trust in practical applications.

The combination of neuromorphic computing, edge AI, and reinforcement learning is anticipated to propel the next generation of MOD systems, allowing for more effective and flexible tracking solutions. Additionally, developments in hybrid tracking methods and multi-modal sensor fusion will enhance scalability and accuracy. In order to handle the challenges of large-scale surveillance and monitoring applications, researchers should also investigate innovative techniques for dynamic perspective modification to guarantee smooth and reliable tracking across vast non-overlapping camera networks.

4.1 MOD from Multiple Cameras with Non-Overlapping Performance Views

The give below figures shows the tests in MOD in multi-cameras with non-overlapping fields of view considerations

.

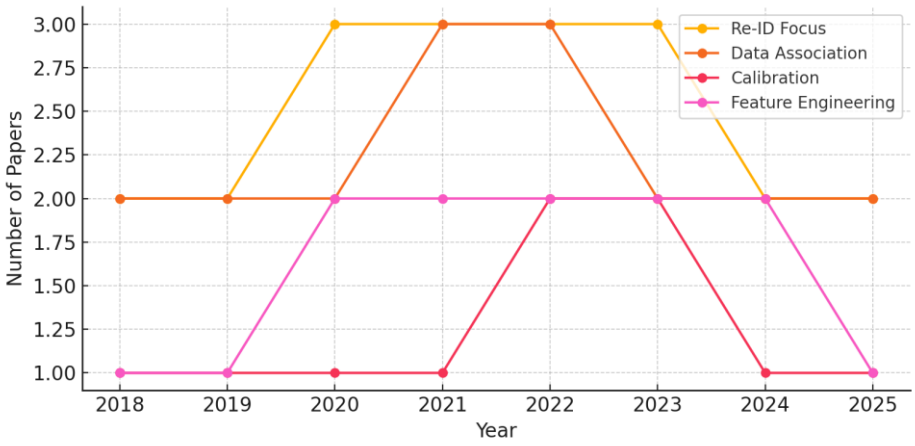

Figure.9 Research Focus Trends Over Time

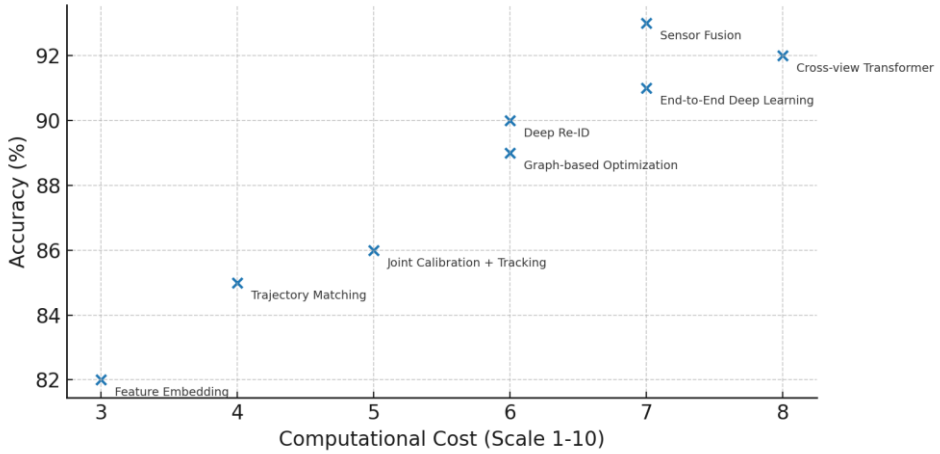

Figure.10 Accuracy vs Computational Cost

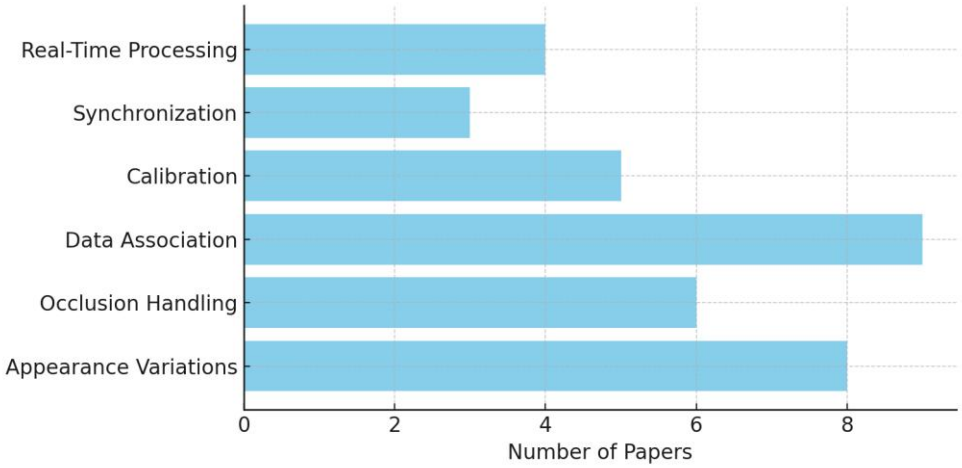

Figure.11 Distribution of Key Challenges

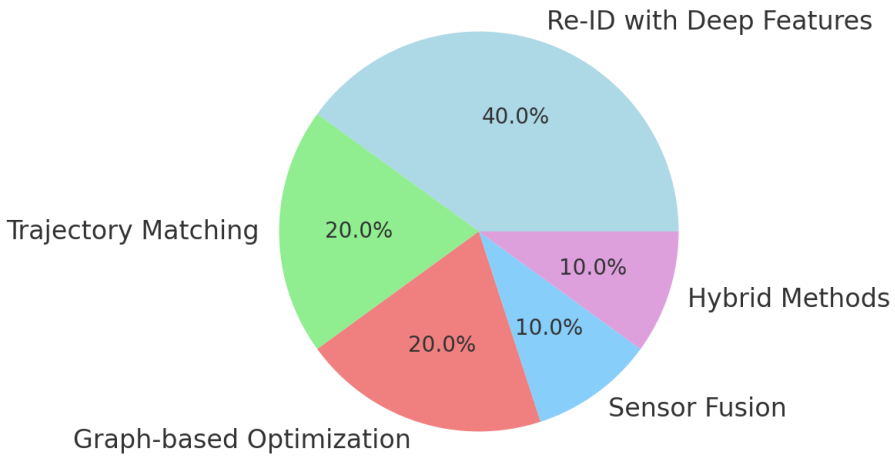

Figure.12 Techniques Used for Multi-Object Association

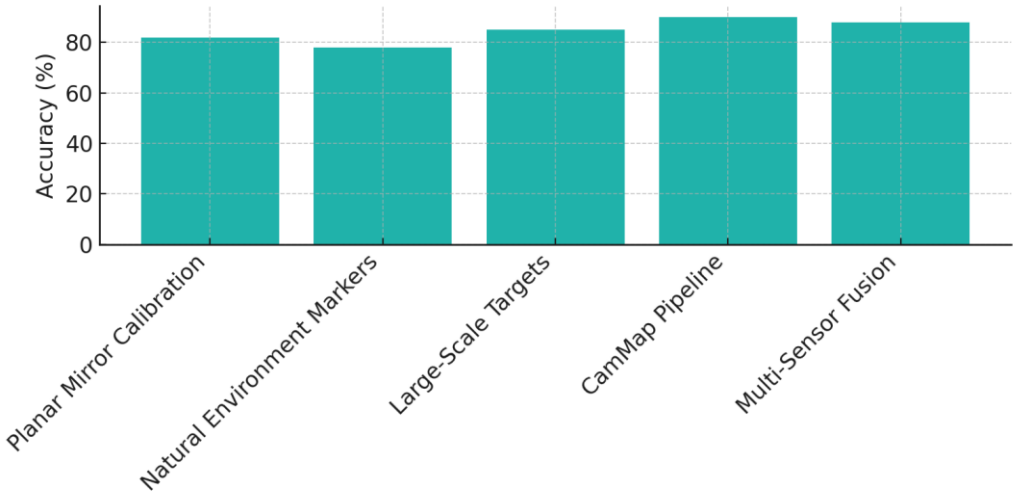

Figure.13 Calibration Method Performance

5. CONCLUSION

Compared to conventional object recognition and tracking frameworks, multi-object identification in multi-camera systems with non-overlapping fields of view poses a challenging set of obstacles. Maintaining stable object identities across diverse viewpoints is made much more difficult by the lack of spatial consistency between non-overlapping camera perspectives. Variations in object appearance brought on by different camera angles, ambient lighting, backdrop complexity, and partial and complete occlusions make this problem worse. Even though deep learning-based object detection models like YOLO, Faster R-CNN, and SSD have shown remarkable performance in single-camera configurations, significant progress in feature extraction, object re-identification, and spatiotemporal data association is required before they can be applied to non-overlapping multi-camera systems.

Critical issues such as differences in visual appearance, occlusions, inconsistent synchronization, and complicated camera calibration were methodically examined in this review. While some advancements are provided by current techniques like re-identification models, graph-based approaches, Kalman filtering, and geometric modeling, they frequently show drawbacks in extremely dynamic and unrestricted real-world settings. Although a number of calibration methods, such as the use of planar mirrors, large-scale calibration targets, natural environmental landmarks, and automated frameworks like CamMap, have shown promise, it is still difficult to achieve accurate calibration across large camera networks. Furthermore, temporal synchronization between several cameras is essential since timestamp misalignment can significantly impair the accuracy of object connections across various viewpoints.The computational load related to object matching and real-time data processing across extensive non-overlapping camera networks is another urgent issue. This difficulty is especially noticeable in applications like autonomous car networks, wide-ranging security monitoring systems, and smart city surveillance. More developments in energy-efficient edge computing, adaptive AI-driven approaches, and reliable cross-camera feature learning techniques are needed to handle these complexities and improve scalability, accuracy, and real-time performance.

Cross-camera tracking accuracy in non-overlapping setups has been shown to be much improved by deep learning techniques, especially those that use deep feature extraction, attention mechanisms, and self-supervised learning. In this setting, advanced Re-ID models that learn invariant features across many perspectives are becoming more and more significant. These methods can enhance long-term tracking and identity retention when combined with graph-based optimization strategies that simulate relationships between detections across cameras. However, when confronted with obstacles like extensive deployments, interactions between dense objects, and erratic environmental changes, current approaches remain constrained. The smooth integration of appearance modeling, object re-identification, temporal synchronization, camera calibration, and real-time processing in non-overlapping camera networks is currently not fully addressed by any one framework. Even while progress has been achieved in each of these areas, it is still difficult to create an end-to-end solution that can successfully integrate these essential elements.

Future Needs

Targeted research in a number of crucial areas is needed to enhance multi-object detection in non-overlapping camera networks. An significant area of development is the fusion of multi-modal sensor data, which includes video feeds as well as data from LiDAR, radar, infrared, and depth sensors. This kind of fusion can enhance object representation, decrease occlusion-induced ambiguities, and increase the accuracy of cross-camera association. Additionally, cross-camera re-identification employing self-supervised and semi-supervised learning approaches may reduce the need for massive annotated datasets, which are sometimes impossible to gather in real-world, large-scale surveillance scenarios.

Another interesting approach is to use contextual information, such scene semantics, spatial constraints, and object interactions, to enhance re-identification accuracy and make more reliable trajectory estimation possible. Graph neural networks (GNNs) and transformer-based designs may provide better relationship reasoning between detected things across different viewpoints. Furthermore, scalability through distributed computing frameworks and edge AI deployments will be crucial for large-scale systems in order to prevent excessive processing overhead and maintain real-time inference capabilities.

The development of adaptive calibration methods that can dynamically adjust to environmental changes, such as changing lighting, seasonal shifts, or camera disruptions, without requiring human intervention, would also significantly increase long-term operating stability. Strong temporal synchronization methods—perhaps combining blockchain-based timestamp verification or federated time-stamping systems—may also be necessary to reduce temporal discrepancies across non-overlapping cameras. Ultimately, addressing these problems will pave the way for the creation of intelligent, adaptable, and reliable multi-object tracking and recognition systems. Large-scale, real-

time deployments for a range of applications, including intelligent transportation systems, smart cities, autonomous robotic navigation, and comprehensive monitoring, will be made possible by these advancements.

References:

1. Nirali Anand Pandya and Narendrasinh C. Chauhan, "Multi-Camera Person Tracking: Integrating YOLOv8 with ByteTrack", SSRG International Journal of Electrical and Electronics Engineering, Volume 11, Issue 10, October 2024.
2. Yuqiang Lin, Sam Lockyer and Nic Zhang, "City-Scale Multi-Camera Vehicle Tracking System with Improved Self-Supervised Camera Link Model".
3. Wennan Wu and Jizhou Lai, "Multi Camera Localization Handover Based on YOLO Object Detection Algorithm in Complex Environments", IEEE Access, 23 January 2024.
4. Hongkai Zhang, Ruidi Fang, Suqiang Li, Qiqi Miao, Xinggang Fan, Jie Hu and Sixian Chan, "Multi-Camera Multi-Vehicle Tracking Guided by Highway Overlapping FoVs", 9 May 2024.
5. Changshuai Dai, Ting Han, Yang Luo, Mengyi Wang, Guorong Cai, Jinhe Su, Zheng Gong and Niansheng Liu, "NMC3D: Non-Overlapping Multi-Camera Calibration Based on Sparse 3D Map", 13 August 2024.
6. Lunlin Fei and Bing Han, "Multi-Object Multi-Camera Tracking Based on Deep Learning for Intelligent Transportation: A Review", 10 April 2023.
7. Xiaotian Han, Quanzeng You, Chunyu Wang, Zhizheng Zhang, Peng Chu, Houdong Hu, Jiang Wang and Zicheng Liu, "MMPTRACK: Large-scale Densely Annotated Multi-camera Multiple People Tracking Benchmark", 30 Nov 2021.
8. Jerome Rutinowski, Hazem Youssef, Sven Franke, Irfan Fachrudin Priyanta, Frederik Polachowski, Moritz Roidl and Christopher Reining, "Semi-automated computer vision-based tracking of multiple industrial entities: a framework and dataset creation approach", EURASIP Journal on Image and Video Processing, 2024.
9. Alexandru Niculescu-Mizil, Deep Patel and Iain Melvin, "MCTR: Multi Camera Tracking Transformer", 11 Sep 2024.
10. Zhongang Cai, Junzhe Zhang, Daxuan Ren, Cunjun Yu, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo and Chen Change Loy, "MessyTable: Instance Association in Multiple Camera Views".
11. Kefu Yi, Kai Luo, Xiaolei Luo, Jiangui Huang, Hao Wu, Rongdong Hu and Wei Hao, "UCMCTrack: Multi-Object Tracking with Uniform Camera Motion Compensation", The Thirty-Eighth AAAI Conference on Artificial Intelligence, 2024.
12. Kang Shan, Matteo Penlington, Sebastian Gunner, Konstantinos Koufos, Mehrdad Dianati, Andrew Fairgrieve and Ian Kirwan, "Experimental Study of Multi-Camera Infrastructure Perception for V2X-Assisted Automated Driving in Highway Merging".
13. Haibin Liu , Chao Wu and Huanjie Wang, "Real time object detection using LiDAR and camera fusion for autonomous driving", 2023.
14. Ayoub El-Alami, Younes Nadir and Khalifa Mansouri, "A review of object detection approaches for traffic surveillance systems", International Journal of Electrical and Computer Engineering (IJECE), Vol. 14, No. 5, October 2024.
15. Jorge Nino Castaneda, Vedran Jelaca, Andres Frias, Reyes Rios Cabrera and Tinne Tuytelaars, "Non-Overlapping Multi-Camera Detection and Tracking of Vehicles in Tunnel Surveillance".
16. Pier Luigi Mazzeo, Paolo Spagnolo, and Tiziana D'Orazio, "Object Tracking by Non-overlapping Distributed Camera Network", 2009.
17. Senquan Yang, Fan Ding, Pu Li and Songxi Hu, "Distributed multi-camera multi-target association for real-time tracking", 2022...