



Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)



A Comprehensive Survey on Multi-Object Tracking: From Classical to Deep Learning-Based Methods

Batool B. Aboud*, Hikmat Z. Neima, Maytham Alabbas

Department of Computer Science, College of Computer Science and Information Technology, University of Basrah, Basrah, Iraq.

Email: batool.baji@uobasrah.edu.iq, hikmat.taher@uobasrah.edu.iq, ma@uobasrah.edu.iq

ARTICLE INFO

Article history:

Received: 00 /mm/2026

Revised form: 00 /00/2026

Accepted: dd /mm/2026

Available online: 30/09/2026

Keywords:

Multi-Object Tracking (MOT)
Deep Learning
Computer Vision
Kalman Filter
Evaluation Metrics
Sensor Fusion

ABSTRACT

Computer Vision is one of the most prominent fields within the area of artificial intelligence, as it aims to provide computers with the same visual understanding as humans. Object detection and tracking play a critical role in the development of intelligent systems, as they allow for intelligent decision-making from the information gathered from vision systems. Multiple Object Tracking (MOT) is one of the most challenging yet important problems within the field of computer vision. MOT requires the system to detect and then keep track of multiple objects in a sequence of video frames. Some of the applications of MOT include autonomous vehicles, security systems, healthcare, and sports. The use of deep learning techniques, such as Convolutional Neural Networks (CNN), Transformers, and self-supervised learning, has improved MOT techniques. Additionally, the fusion of deep learning and sensor data with classical methods like the Kalman and Particle Filter has created methods with a good balance between accuracy and speed. Despite the improvements in these methods, there are still some challenges to the optimal performance of MOT algorithms. A survey of these methods, including classical, deep learning, and hybrid methods, as well as an overview of the datasets, metrics, applications, and current state-of-the-art methods and their evaluations, is presented in this paper.

<https://doi.org/10.29304/jqcm.2026.18.32838>

1. INTRODUCTION

Multi-object tracking (MOT) is a central problem in computer vision and can be widely used in intelligent transportation systems, surveillance systems, sports analysis, etc. This domain involves the problem of detecting and tracking multiple objects in a video sequence such that each object can be assigned a unique identifier across frames of the video so that smooth and accurate trajectories of the objects can be obtained as they move around. With the rapid development of technology, object tracking technology has become increasingly powerful and robust, especially with the development of deep learning methods, such as convolutional neural networks (CNNs), which make it possible to perform object tracking more accurately and efficiently in a variety of complex environments [1].

*Corresponding author: Batool B. Aboud

Email addresses: batool.baji@uobasrah.edu.iq

Communicated by 'sub etitor'

Although great success has already been achieved in tracking problems, MOT remains a challenging problem, including 3D MOT, occlusion, identity switching, and illumination variations (low light, weak frequency band change, and environmental conditions). To overcome these problems, multiple sensor-based data fusions, including cameras, radar, and LiDAR, are a viable approach for enhancing the object tracking performance. When their powers are combined, one may take advantage of the strength of the other, creating a more robust and accurate object-tracking monitor in a variety of environments [2][3]. Modern deep learning methods, including the most advanced algorithms (e.g., YOLOv8 and StrongSORT), have enhanced tracking systems in complex urban environments. Similarly, these advances have improved the accuracy of real-time tracking and can address issues such as partial object occlusion and rapidly changing motion. The availability of novel detection approaches and tracking learning algorithms has made pedestrian tracking systems stronger, more adaptable, more efficient, and suitable for many applications such as surveillance, autonomous driving, and crowd management [4].

In this paper, we review the recent advances in the field of MOT within computer vision, with a focus on both the widely adopted methods and emerging new methodologies for MOT. MOT aims to track multiple objects of interest in a sequence of 2D images or video frames. Due to the presence of various challenges, such as occlusion, illumination changes, scale variation, background changes, and crowded scenes, MOT remains a difficult problem to solve. This paper reviews the different approaches to MOT, including detection-based tracking, data association, and learning-based methods. Furthermore, MOT using appearance models, motion models, and temporal consistency is also discussed. Finally, the importance of MOT in a variety of different applications is also discussed, ranging from self-driving cars to surveillance, security systems, and sports analytics. Each of these areas of application requires high levels of accuracy and tracking capabilities in real time, even in challenging environments. Thus, MOT is an essential component of many modern intelligent vision systems. This survey paper aims to provide insights into the current state-of-the-art in MOT research, as well as in its development, for readers who wish to contribute to the field.

Section 1 introduces MOT and its applications in areas like self-driving cars and surveillance systems, as well as the challenges associated with these areas. Section 2 covers the fundamentals of MOT and its relationship to SOT. Section 3 outlines the challenges associated with MOT. In Section 4, the different techniques that have been employed to solve MOT are reviewed. Section 5 details the main datasets that have been used and published for MOT research. Section 6 introduces evaluation metrics for MOT systems. Section 7 explores the applications of MOT technology. Section 8 compares different MOT methods. Section 9 identifies current research gaps and future directions, and Section 10 concludes the paper.

2. Fundamentals of MOT

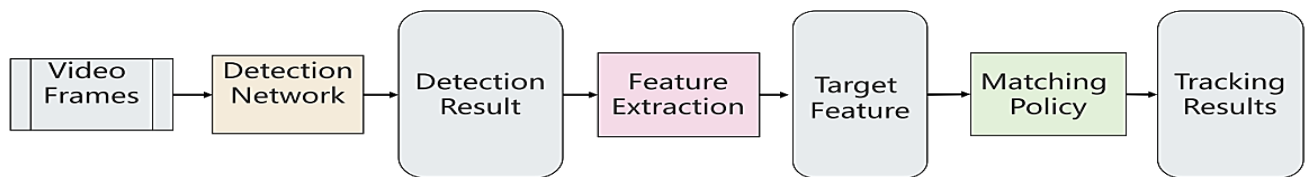
MOT involves the location and linking of multiple objects across video frames. It combines object detection methods, such as YOLO, with tracking frameworks, such as Kalman filters or deep learning models, to address challenges, such as occlusion and identity switching, enabling smooth object tracking.

2.1. Detection vs. Tracking

In detection, an object is localized in each frame of a video, and the location and dimensions of the object are identified using deep neural network methods, such as YOLO or Faster R-CNN [5]. On the other hand, it aims to maintain the identity of an object by connecting the already produced objects in different frames, which is usually implemented by the so-called "tracking-by-detection" method. It also uses techniques such as the Kalman filter or optical feature matching [6], [7]. The crucial difference is that detection aims to detect objects that appear in every frame, whereas tracking aims to preserve the continuity of objects through frames by maintaining the correspondence between objects [5]. There are some challenges in learning-based object detection (precision of locating objects in each frame), and the tracking challenges are occlusion and appearance change [1].

Frame-by-frame detection analyzes every frame independently and only detects targets on each frame, which can achieve real-time performance but with less accuracy, especially in complex scenes such as identity switches and occlusions [8][9]. Temporal tracking enhances accuracy by associating detections over time and considering the motion history of objects; however, it requires additional computation [10][11]. Bounding boxes are fundamental in detection and tracking and indicate the location of an object. In 3D MOT, LiDAR data can improve clutter performance because they represent spatial attributes [12]. Deep learning models have also been used for feature extraction (e.g., CNNs), which supports the identification of objects based on their appearance, and a motion prediction model (such as the Kalman Filter) for predicting the upcoming position of the objects [13][14][15]. Detection is challenging because of the prevalence of false positives (identifying non-existent objects) and false

negatives (not identifying real objects and object misses), in crowded settings [16][17]. However, tracking frequently encounters identity switches, in which the system mistakenly assigns different identifiers to objects owing to occlusions or appearance changes. Such problems are dealt with by methods such as the Kalman Filter,



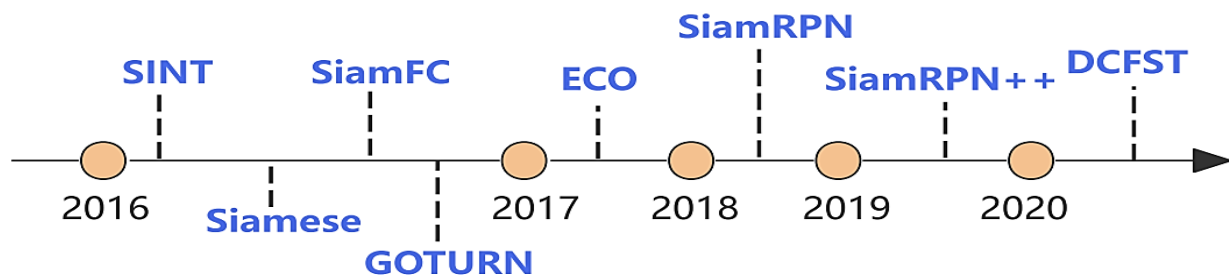
TrackFormer, to enhance the motion prediction and confidence association with the targets [8][18]. Figure 1 presents the main procedures of the MOT framework, which consists of four core components: object localization, feature extraction, data association, and track management [61].

Figure 1. The procedure of the MOT framework show detection and tracking processes

2.2. Single-Object Tracking (SOT) vs. MOT

SOT involves tracking a single object across video frames. This straightforward approach is ideal for applications in which only one target needs to be followed, such as in robotic systems tracking a specific object or in scenarios with isolated targets [19][20]. SOT is a competitive robotics field that aims to track a single target. Moreover, SOT has

been applied in scenarios involving isolated targets, such as sports analytics and video surveillance. The



performance problem with the SOT algorithm is that it cannot be scaled to simultaneously estimate multiple targets. Consequently, it suffers from target occlusions and high object speeds. It also faces difficulties when the targets overlap, and large changes in the appearance of the target occur [21] and [22], respectively. The algorithm is less complex because only one target is processed simultaneously. However, their processing requirements are relatively low because they track only one target. It primarily focuses on detecting the target and updating its location. The development of SOT methods is shown in Figure 2 [61].

In contrast, MOT involves tracking multiple targets simultaneously. The complexity lies in the need to correlate the data and determine which target in each frame corresponds to the target in the next frame [20] and [23]. Scalability is a challenge when dealing with multiple targets, particularly under time constraints in living systems. For example, in autonomous driving, MOT is essential for distinguishing between different vehicles on the road [24]. MOT algorithms typically use target detection methods for initial target detection, followed by data association techniques to track targets between frames. The problem is to match the targets across frames, which is very difficult in the presence of occlusions, overlapping, and sudden motion changes [25][26]. MOT requires a large amount of computational power because the data association and tracking of multiple targets are very complicated. In addition, occlusion and appearance similarity between targets are great challenges that result in identity switching and track loss. Figure 3 illustrates a hierarchical classification of MOT methodologies [61].

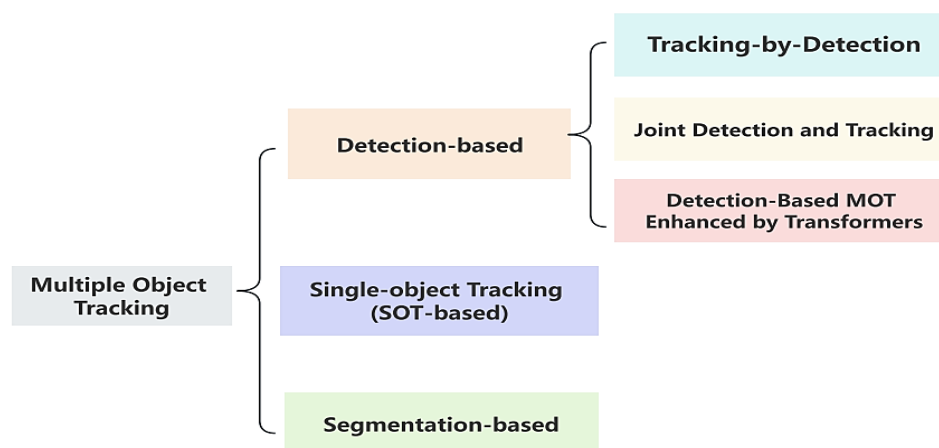


Figure 3. Hierarchical Classification of MOT Methodologies

3. Challenges in MOT

MOT in videos encounters many challenges, such as appearances, disappearances, movements, and changes in views, which may hinder trajectory associations over time. This challenge is much more demanding than a standard object detection problem. It requires objects to be tracked over long temporal sequences, with objects having a unique and consistent identity in the presence of scene clutter and dynamic movement.

3.1. Occlusion and Identity Switching

Occlusion and identity switching are among the most challenging problems in MOT, especially in highly crowded scenes, which can cause detection loss, path overlaps, and false positives [18] [27]. Object identities are also switched when objects are in proximity, which degrades tracking continuity [28]. Recurrent identification models are crucial for re-identification after occlusion when conventional algorithms, such as Kalman filters or Intersection over Union (IoU) matching, do not yield good results [29] [30]. Advanced systems further incorporate these models to increase the tracking accuracy in complex scenarios [28]. Such models help to associate lost tracks with the correct object after occlusion [31]. While they are computationally expensive, the innovative utilization of morphological features, along with linking algorithms such as the Hungarian algorithm, counteracts this problem and decreases the number of identity switches, thereby enhancing accuracy [21] [32].

3.2. Environmental Factors

Changes in lighting, particularly under low-light conditions, are among the most prominent challenges in tracking multiple objects. A decrease in image quality and an increase in noise cause performance degradation. To counter this, techniques such as adaptive frequency reduction are used to enhance low-frequency components and eliminate noise. In addition, glare from bright lights can cause objects to merge with the background. To improve tracking under these conditions, image-feature enhancement techniques and sensors, such as radar and cameras, are integrated. Moreover, it is more difficult to track under bad weather conditions, such as rain or fog, because of the additional noise and distortion in the signal and the loss of tracking position. In these scenarios, rain removal using a Siamese network was employed to alleviate the problem and enhance the accuracy of rain tracking [33]. Image detail is also lost by motion blur due to camera movement or high-speed object movement, and tracking may also degrade. To address this issue, several systems have employed deep learning models to enhance monitoring in fast-moving scenes [33][34][35].

3.3. Real-Time Processing and Computational Demands

Real-time processing in autonomous systems suffers because of the large amount of data that they must process, such as camera feeds and LiDAR, within the limited latency allowed. High latency may lead to slow reactions, which is a challenge in dynamic environments, particularly for vehicles that need to make decisions quickly. Reducing latency is paramount for ensuring safety and productivity [5]. Balancing accuracy and speed is a fundamental problem in computer vision models. Computational high-resolution models are very expensive (turnaround time can be long), and fast models lose accuracy. In [6] [7] it was proven that the deployment of hardware accelerators

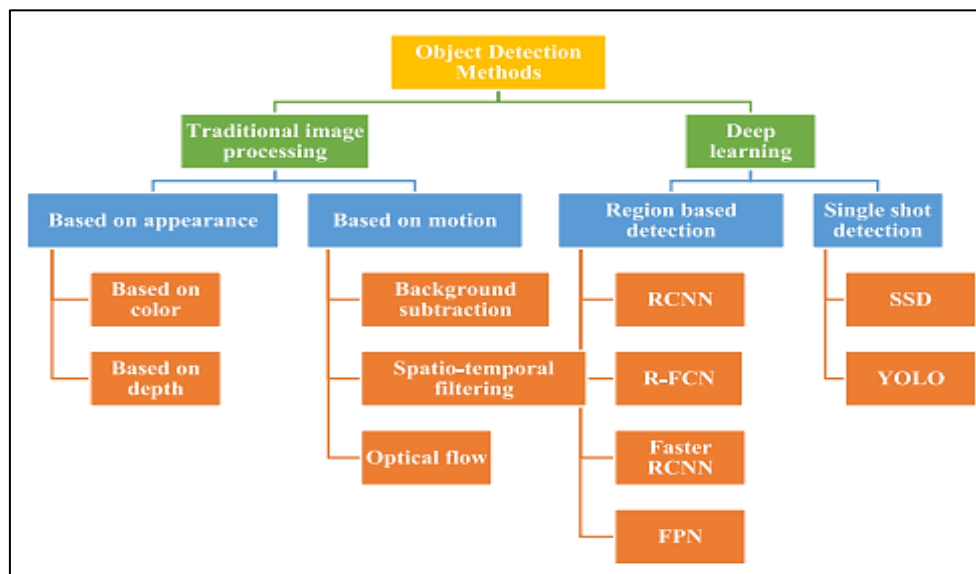
(e.g., GPUs) and lightweight architectures (e.g., Ghost Net) could efficiently reduce the computational cost while accelerating processing without loss in terms of accuracy.

3.4. Object Appearance and Behavioral Variability

The pose and scaling changes that can happen to the subjects, as well as the changes in their clothes, illumination, and the actions that they perform, can all pose challenges to the system. Moreover, these problems are worsened in the presence of multiple targets, especially in patients with MOT. Changes in lighting within the captured images can impact the appearance of those subjects within the images, making it challenging to maintain their identity between different frames of the video [24][9]. Additionally, movement issues such as fast movements of the objects or targets to be tracked can cause significant problems in single-object recognition systems. These movement problems include missing tracks of the objects, erratic trajectories of the target, and identity switches for the objects to be tracked [24][19]. However, by incorporating motion prediction models like Kalman filters and attention models into the recognition system, the precision of tracking the objects increases [2].

3.5. Data Association and Scalability

In crowded scenes, occlusions and deformable objects are the primary influences of trajectory fragmentation and identity switching [36]. Objects can overlap or be similar in the extent to which they are difficult to distinguish. This makes real-time track continuity challenging and demands advanced data association techniques, such as cross-modality fusion across multiple sensors[37]. Data association is computationally expensive and involves objects. The Hungarian algorithm performs this task but does not scale in highly populated environments owing to its fundamental combinatorial nature. In this case, additional mechanisms, such as adaptive matching and trajectory segmentation, can be developed to maintain tracking performance when differentiating between active and lost trajectories [10].



4. MOT Methods

Object tracking has transitioned from classical methods with handcrafted features to deep-learning-based networks that can effectively and automatically extract features and representations. In this regard, many hybrids have been proposed, from classical models and faster motion prediction to deep models and robust vision recognition comparisons, to find a balance between accuracy and speed. Figure 4 shows the classification of object detection methods [62].

Figure 4. Classification of object detection methods.

4.1. Traditional Methods

Multi-target tracking is based on traditional tracking by employing mathematical models with assumptions

regarding the movement and properties of objects. It uses linear models to make sense of signals and motion, such as the Kalman filter, distance, and similarity-based correlation. Although simple, it achieves a quick and robust response under limited computation and forms a fundamental part of deep- learning-based tracking.

4.1.1. Kalman Filter and Variants

The Kalman filter is an algorithm that estimates the state of a dynamic linear system. It works by predicting the state of the system after the next step, given its current state, and then correcting this prediction using fresh measurements [38]. The Kalman filter is optimal in systems affected by Gaussian noise. In contrast, the Extended Kalman Filter (EKF) is a variation of the Kalman filter designed for non-linear systems. It linearizes the system by performing a first-order Taylor expansion around the current state estimate of the non-linear system. This allows the use of the EKF in non-linear system models, which often results when considering the problem of tracking moving objects, where the dynamics of the system or measurement models are nonlinear. It is popular in disciplines such as robotics, navigation, and sensor fusion, where nonlinearity plays a key role [11][8]. The UKF, which estimates the mean and variance using sigma points, is preferred to the EKF for non-linear systems. This advantage makes the UKF suitable for applications such as autonomous vehicles and the multi-sensor fusion of LiDAR, radar, and cameras in a dynamic environment. The UKF is the best option for highly non-linear systems, because it does not use any approximations, unlike the EKF [8].

4.1.2. Particle Filter

PF is used to estimate the states of three-stage non-linear non-Gaussian state estimation systems: prediction, weight updating, and resampling. During the prediction, the particles are propagated through the system model to forecast the state. The weights are then updated by the likelihood of the observations, and resampling is used to handle the 'particle degeneracy' issue that occurs after several iterations, where the particles lose most of their information value [1],[12]. Because particle filters can deal with non-Gaussian noise, they are well-suited for complicated systems. Nonetheless, they suffer from degeneracy, such that the number of useful particles quickly decreases, leading to a loss of particle diversity. In addition, the computational complexity is relatively high because it must run over many particles, making it unsuitable for real-time applications [39] [14].

4.1.3. Limitations of Classical Approaches

Classical object-tracking methods, such as KF, are inconvenient in non-linear and crowded scenarios. The Kalman filter performs very well for a linear system; however, a linear approximation must be performed for a non-linear system, and a loss of accuracy appears in the case of a large amount of noise. Such issues are particularly debilitating in multi-object or real-time settings [11][38]. Current traditional methods are also ineffective at handling changes in appearance during tracking. They focus on predicting the motion trajectory and spatial shape of the object while ignoring its appearance. Such methods are not efficient when confronted with challenges like partial occlusions and substantial appearance transformations of the objects of interest [39]

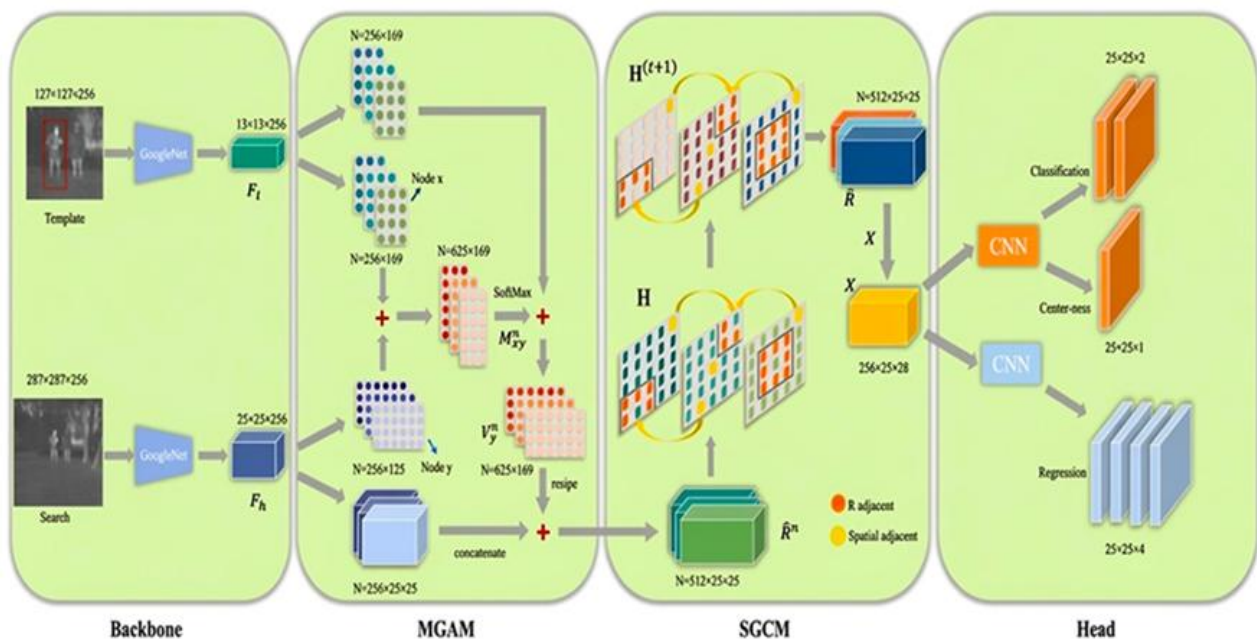
4.2. Deep Learning-Based Methods

Deep feature tracking techniques use deep neural networks to learn a series of features that can be extracted from video footage. These methods provide high accuracy in complex scenes. However, deep learning models are computationally expensive and require a lot of training data to train these models, that limit their applications in settings with constrained computational and data resources.

4.2.1. Transition from Handcrafted to Learned Features

The transition of features from handcrafted to learned has been made possible with the use of convolutional neural networks. Convolutional neural networks can automatically learn the features needed to recognize objects from training data without the need for manual feature engineering [23]. These networks can effectively deal with similar or occluded objects in complex scenes [[23 Many trackers that can perform MOT use Siamese networks to learn similar scores for the same objects from consecutive video frames. Additionally, the use of metric learning in these trackers improves the tracking of objects that are linked over time by the tracker [25][36]. Furthermore, self-supervised methods utilize information from the adjacent frames in the video to learn representations of the objects in the scene, thus eliminating the need for costly human annotations for training these models and improving their scalability [36].

4.2.2. Modern Architectures



Recently, transformer-based models have become the most adopted models for performing MOT. Models such as TrackFormer and MOTT use a self-attention mechanism that explores the relationship between all the objects in all the frames in the video, thus making them capable of tracking complex scenes with occlusions and interactions among the objects of interest [15][16]. TrackFormer uses a lightweight attention mechanism, and MOTT enhances computational efficiency based on a green learning model that achieves state-of-the-art performance without any degradation in speed [16].

Graph Neural Networks (GNNs) have also been applied to MOT, such as modeling the spatiotemporal relationships of objects. For instance, the CTGMOT model fuses CNNs and GNNs to extract appearance and motion information simultaneously and has proven its efficiency in tracking athletes in team sports in challenging scenarios [17][30]. The combination of 2D, 3D, and RGB-T (thermal) tracking improves the model accuracy by incorporating multidimensional data and results in more stable tracking in complex environments. These integrated data maintain good performance in coping with scenarios such as occlusion or visibility changes [4][18][40]. Figure 5 presents the overall architecture of the multiple graph attention module [39].

Figure 5. The overall architecture of the multiple graph attention module.

In Table 1, a comparison of recent MOT models in terms of architecture, primary mechanism, and core objective is presented.

Table 1. Comparison of recent MOT models

Model	Type / Main Components	Primary Mechanism	Core Goal / Advantage
TrackFormer	Pure Transformer (Encoder–Decoder).	Uses <i>Self-Attention</i> with tracking and object queries for temporal association.	End-to-end tracking; robust to occlusion and dense scenes.
MOTT	Lightweight Transformer (<i>Green Learning</i>).	Employs CSWin Encoder and Deformable DETR Decoder to reduce computation.	High accuracy with low computational cost.
CTGMOT	Hybrid CNN + Transformer + GNN.	Combines local, global, and relational Features for tracking.	Improved accuracy and stability in complex motion

 scenes.

4.2.3. Self-Supervised and Unsupervised Learning

Unsupervised and machine learning are important methods for MOT to reduce dependence on expensive and time-consuming labelled data [41]. They leverage the model's predictions to find object-frame associations without direct supervision. In the case of interference and visual artifacts, such tracking is both robust and efficient [42]. In SS-MOT, Contrastive learning provides a solution for training a model that is independent of defining target trajectories at the instance level. This approach learns discriminative object representations by enforcing the similarity between the same object over frames and maximizing the similarity between different objects [36]. This allows MOT models to automatically achieve object association (i.e., without manually labelled data).

4.3. Hybrid Methods

Hybrid approaches in MOT Hybrid approaches in MOT represent a step forward in bridging classical and modern deep learning-based methods by integrating their merits to achieve higher performance and computational efficiency.

4.3.1. Integration of Classical and Deep Learning

Deep-Sort is a popular hybrid model that uses a classical Kalman filter to predict object motion and a CNN to detect it; therefore, it can effectively handle difficult videos that contain fast-moving and highly occluded objects [42][43]. Adaptive processing techniques, including ignoring unimportant frames or processing intelligent prediction processes, such as slow motion using a linear filter and fast motion using a quadratic approximation, enable a balance between computational speed and accuracy [42][44]. Their robustness in cluttered scenes can be further enhanced by combining them with deep-learning-based object detection and classic algorithm-based motion trajectory optimization, which ensures robustness when dealing with many simultaneously moving objects [45].

4.3.2. Advantages and Trade-offs

The integration of the classical method with deep learning-based tracking achieves higher accuracy, particularly in highly dynamic or complex scenes, including occlusions and rapid object movements [46][5]. These methods enhance the adaptation capabilities of the system, with which tracking is accomplished under severe variations robustly and accurately. In contrast, hybrid methods attempt to minimize the computational cost by "selective processing." Frame cancellation via object speed or motion prediction is also utilized to save on computation, only if the tracking performance is not adversely impacted, such that the system can be run more efficiently on the existing hardware [1] and [12], respectively. However, there are a few shortcomings of the hybrid method, such as the complexity of integration and the requirement for a large amount of training data. Combining various techniques, such as CNNs, Kalman filters, and traditional tracking techniques, requires careful consideration and balance. Furthermore, many of these techniques require significant training data [47][18].

5. Datasets and Benchmarks

Datasets and benchmarks play critical roles in the development and evaluation of MOT methods. Various datasets have been created to allow for the testing of MOT methods in a variety of settings, to fuel competition between algorithms, and to enable the development of further MOT methods as a result.

5.1. Standard Datasets

MOT challenges are another important topic within the field of computer vision datasets (e.g., MOT15, MOT16, MOT17, and MOT20). Each of these datasets contains a variety of challenging environments with realistic situations within those environments, situations that include challenges like high levels of crowd density, complex motion within the environment, occlusion of individuals in video clips from those environments, and congestion of individuals within specific areas of those environments [6]. These datasets are provided with videos annotated with the accurate position of objects in consecutive frames and can be used for evaluating tracking algorithms on benchmark metrics, such as MOTA and IDF1 [42] [10]. In the context of autonomous driving, the KITTI dataset is a prominent source of 3D object tracking [18][5] and sensory data, including images, radar, and lidar [8], making general environment perception systems possible. The BDD100K dataset also includes a variety of scenes under varied climate and environmental conditions [9], and is important for advancing the abilities of computer vision systems (e.g., in the domain of autonomous driving applications). In addition, the UA-DETRAC dataset is dedicated to vehicle tracking in crowded urban traffic scenes [24] and has been adopted to monitor performance and enhance

vehicle-tracking algorithms [20].

Name	Domain	Sensors	Key Challenges	Size	Primary Use Case
MOT15	MOT	Various	Occlusions,	11 sequences	MOT in crowds

5.2. Specialized and Emerging Datasets

Several specialized or emerging datasets have recently been proposed to benchmark MOT algorithms under challenging and unconventional conditions. These datasets are used to implement specific scenes or features that are biased towards traditional tracking frameworks and further contour the development of target tracking. One of them is the SportsMOT dataset, which is assigned to sports players and sports object performance analysis in sports videos. This data set consists of labelled trajectories from three sports categories: basketball, volleyball, and soccer [30] [21] [48]. The other is the DanceTrack dataset of dancing sequences, which contains fast-moving and visually similar objects, complicating tracking in challenging dynamic environments [27]. This dataset was designed to evaluate the ability of the algorithms to handle rapid overlaps and object merging [41]. Other datasets address these environmental challenges, whereas LMOT focuses on its ability to track objects under low-light conditions, which is a challenge for traditional visual recognition systems [9]. Like the work of the authors in RainyTrack, this allows the synthesis of artificial rain effects in video sequences for the development and evaluation of vision algorithms under rainy weather conditions [33]. By contrast, UniRTL integrates RGB, thermal, and low-light conditions and is a valuable benchmark for evaluating tracking performance in challenging situations [19]. Finally, Refer-KITTI is a tailored extension of the KITTI suite that includes language-driven object tracking, in which natural language is used to instruct tracking algorithms in a challenging environment. This dataset characterizes a new trend of combining natural

language reasoning with object tracking to develop more interactive and intelligent tracking algorithms [28]. Table 2 provides

a comparative analysis of the datasets in terms of several aspects, including challenges and data size.

Table 2: Datasets Comparison

MOT16	MOT	Cameras (RGB)	stability	(550 frames)	
		Various Cameras (RGB)	Heavy crowding, occlusion	14 sequences (750 frames)	Benchmark for crowded scenes
MOT17	MOT	Various Cameras (RGB)	Detector performance, occlusion	14 sequences (1,000 frames)	Detector performance in crowds
MOT20	MOT	Various Cameras (RGB)	High density, occlusion	8 sequences (1,000 frames)	Tracking in dense crowds
KITTI	Autonomous Driving	Stereo RGB, LiDAR	3D tracking, traffic motion	21 sequences (15,000 frames)	Vehicle navigation, sensor fusion
BDD100K	Autonomous Driving	RGB Cameras	Diverse weather, data size	100,000 frames	Robust tracking in diverse environments
UA-DETRAC	Vehicle Tracking	RGB Cameras	Vehicle tracking, occlusion	1,500 videos	Traffic analysis, vehicle tracking
SportsMOT	Sports Analytics	RGB Cameras	Fast motion, occlusion	1,700 sequences (5,000 frames)	Sports performance, tracking
UniRTL	RGB-Thermal Benchmark	RGB-Thermal Sensors	Low-light, fusion	10,000 frames	All-weather Tracking
Refer-KITTI	Language-Guided	RGB Cameras, Text	Language integration	50,000 frames	Language-guided tracking
RainyTrack	Synthetic Augmentation	RGB Cameras (Rain)	Rain, visibility	50,000 frames	Robust tracking in rain
DanceTrack	Fast Motion	RGB Cameras	Appearance similarity, motion	100 sequences (200,000 frames)	Tracking fast-moving objects

6. Evaluation Metrics

The performance of the tracking algorithms was evaluated using MOT evaluation metrics. These metrics indicate the extent to which the algorithm contributes to the accuracy of tracking, persistence, and location. In the following section, we introduce the MOT evaluation metrics that we apply in this study, along with their meanings.

6.1. Multi-Object Tracking Accuracy (MOTA)

MOTA is a measure of the accuracy of an MOT system on a video [3]. It can be determined by Equation 1 [61]:

$$MOTA = 1 - \frac{(FN + FP + IDSW)}{GT} \quad (1)$$

where FN is the number of false negatives (missed objects), FP is the number of false positives (incorrectly detected objects), and IDSW is the number of identity switches. This is an important metric for tracking quality, and a higher MOTA result indicates better tracking performance.

MOTA represents the global accuracy of a tracker, which is affected by error rates, such as identity switches and false positives. This metric can also be sensitive to overlaps or occlusions of objects. MOTA also depends on the accuracy of the ground truth (GT); if the ground truth is noisy, the performance might not be accurate. It has been used in conjunction with other metrics, such as MOTP and HOTA, to further analyze its performance [10][29].

6.2. Multi-Object Tracking Precision (MOTP)

MOTP assesses the accuracy of object tracking. It is related to the assessment of positioning accuracy. The performance of this metric is based on the IoU between the predicted and actual positions of the objects [31][6]. It is closely related to the MOTA metric. While the former assesses the total number of errors in the tracking, the MOTP metric assesses only the localization accuracy of the objects being tracked. This parameter is important for environments with much interference or for fast-moving objects. The MOTP metric is commonly defined by Equation 2 [13].

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \quad (2)$$

Where the terms, $d_{t,i}$ and c_t are the bounding box overlap with its assigned ground truth object and the count of matches at frame (time) t , respectively.

6.3. Identity-Based Metrics

Another set of metrics related to the identity of the objects being tracked includes the IDF1, ID, and IDSW metrics. The IDF1 is a metric that balances the precision and recall of the object identities being tracked. The IDSW metric determines the number of identity switches that occur due to clutter or misdetection within the video [32]. These metrics are important to include in the assessment of the algorithm, especially for dynamic crowds [50]. Metrics that measure precision, recall, and the F1-score of the system are also utilized to assess the tracking quality. The precision of the system reflects the accuracy of the system in tracking the objects. The recall metric ensures that there are no objects missed by the system. The F1-score is a combined metric of precision and recall that can be used to reflect the overall performance of the system [5], [35]. The IDF1 metric is defined by Equation 3 [32].

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \quad (3)$$

Where IDFP and IDFN are the number of false positive ID and false negative ID, respectively. IDTP is the number of true positive ID.

6.4. Higher-Order Metrics

Another metric that considers the accuracy of the objects being tracked is the Higher Order Tracking Accuracy (HOTA), which factors in accuracy as well as the consistency of the system over time in its tracking of those objects [48]. Additionally, the CLEAR metric assesses the number of identity switches made by the system in cases of volatile object movements [43]. These two metrics allow for a more thorough analysis of the system and its performance in the tracking tasks. The HOTA metric can be used to assess the accuracy of the system in tracking the objects. CLEAR can be used to assess the system's ability to track complex and fluctuating object movements. These two metrics have been utilized in numerous tracking studies to improve the performance of those systems, especially in complex scenarios.

6.5. Speed and Efficiency Metrics

A set of metrics for evaluating object tracking systems is in the area of speed and efficiency. Metrics like frames per second help to indicate the speed of processing of the video between the various components of the system. Higher frames per second are associated with higher accuracy in real-time systems, such as autonomous vehicles [5]. Additionally, system efficiency is related to the size of the model, as well as the amount of memory that it utilizes to function. Generally, a larger model will provide higher accuracy in identifying the objects in the video stream, but at the cost of increased memory usage of the system. System efficiency is also correlated with the inference time of the system. In this regard, the trade-off between the accuracy of detection and memory utilization becomes important, as in many systems that employ only the integration of time histories to improve the accuracy of detection and tracking while maintaining low memory usage [51].

7. Applications of MOT Systems

MOT systems are the fundamental cornerstones for converting academic ideas into applicable and effective solutions. These systems are widely used in critical applications, including autonomous driving, state-of-the-art security systems, and sports analysis, and they offer a high level of precision and efficiency in motion recognition and behavior analysis.

7.1. Autonomous Vehicles

Sensor fusion is one of the most fundamental technologies used by AVs to accurately recognize the environment. Sensor fusion combines data from multiple sensors, including cameras, radar, and LiDAR, to provide a more comprehensive and robust perception of the environment. These sensors have distinct functionalities: cameras capture high-resolution visual information from the environment, radar measures distance and speed information, and LiDAR provides accurate 3D spatial measurements [5][24]. Through the integration of these data streams, autonomous vehicles (AVs) can overcome the limitations of individual sensors, such as reduced performance in

inclement weather and low-light conditions [52]. Real-time decision-making is also crucial for autonomous systems. AVs make collision-free, safe, and smart driving decisions by processing the fused sensor data. This consists of complex algorithms that enable rapid evaluation of the current situation, prediction of possible dangers, and optimal decision-making [8] [53].

7.2. Surveillance and Security

Crowd monitoring systems based on algorithms, such as DeepSORT, can track individuals and identify abnormal behaviors, such as rows or abrupt gatherings, in public areas using the output from several cameras. Such systems analyze videos in real time based on motion and appearance features to distinguish between people; thus, they minimize interference and improve the precision of abnormal event detection [54]. Multi-camera tracking algorithms are crucial for monitoring overcrowded scenes in which people move across different cameras. Past objects are associated with detected objects using algorithms that utilize the similarity between the objects and the image, distance, or IoU. The AS-Track algorithm has slightly better performance based on spatial and appearance features and thus provides better tracking accuracy in scenes with occlusions [55].

7.3. Sports Analytics

In sports, players are tracked using MOT techniques to calculate their performance measures. Algorithms, such as OpenPose [48] and DeepSORT [48][56], can take advantage of both motion and appearance features for robust real-time tracking in high-speed team sports with heavy occlusions, such as basketball. These procedures allow player positions (and hence game-related behaviors) to be analyzed and performance monitored [21]. The key points of athletes in motion can also be tracked in real-time using OpenPose/DeepSORT, even when moving fast and somewhat irregularly. Together, they reinforce the tactical side and allow for extremely detailed analyses of the players' actions, positions, and interactions within the game [2] and [30], respectively.

7.4. Healthcare and Industrial Safety

Computer vision with artificial intelligence (AI) is indispensable for industrial and healthcare safety. In an industrial setting, a range of automated personal protective equipment (PPE) monitoring systems (e.g., mask and helmet equipment) has been more advanced than manual management in the process of protection control, particularly in crowded areas. The use of detection algorithms and built-in tracking facilities has already reduced accidents caused by not wearing PPE on worksites, including cramped spaces [57][58]. Accurate and efficient guidance of surgical instruments is required to improve surgical accuracy and patient safety. However, conventional solutions still have limitations in terms of accuracy and efficiency. To address this, deep learning-based solutions (such as LSTM networks) have been proposed to track instruments quickly and accurately, or computer vision systems with virtual markers have been adopted to detect and estimate the motion of instruments. Such improvements ensure the correct degree of accuracy in surgical tasks and allow for a real-time balance between speed and accuracy [59]

7.5. Retail and Customer Behavior Analysis

Computer vision-based surveillance systems (e.g., YOLOv5 and DeepSORT) are transforming in-store customer behavior analysis by extracting insights into customer mobility and product interactions [60]. Heatmaps are used to discover crowded and infrequent areas that can advance medium strategies for shop layout and product arrangement [61]. Observing interactions (item viewing and handling) also allows the measurement of dwell time and engagement for personalized marketing and inventory control [62]. Compliance with data security laws, such as the GDPR, continues to be a critical element in ensuring the success of these systems.

8. Comparative Analysis

The last few years have seen a widespread development of motion tracking as a result of improvements in artificial intelligence and image processing algorithms. The fast and reliable tracking systems play an important role in a variety of applications. It is necessary to evaluate these methods for comparison in their performance. Such a performance analysis will provide the basic understanding of each type of approach to MOT, allowing those evaluating the technology to find the most suitable approach for their specific applications.

8.1. Speed and Efficiency

Table 3 shows the impact of the architecture of these motion tracking systems on their processing performance and cost. Traditional tracking systems, such as SORT, fail to achieve high accuracy in complex scenarios but can process high speeds with little computational effort. Pure deep learning approaches, such as FairMOT and MOTR, achieve

higher accuracy but sacrifice some of the system's processing speed due to the significant computational calculations that it performs every frame. A mix of deep learning and classic tracking methods, such as ByteTrack and DeepSORT, offers a middle ground between the two extremes. They are the most feasible option for implementation in various applications.

Table 3. Comparison of MOT Method Categories Based on Processing Speed

Method Category	Representative Methods	Primary Speed Factor
Traditional Methods	SORT, Kalman Filter, Particle Filter	Linear equations and simple features with minimal computations per frame
Hybrid Methods	ByteTrack, DeepSORT, BoT-SORT, StrongSORT	Less frequent deep detection with lightweight classical tracking between frames and skipping frames.
Pure Deep Learning	MOTR, TrackFormer, TransTrack, FairMOT	Intensive computations (CNN/Transformer/GNN) per frame or across temporal sequences

8.2. Accuracy Comparison

Accuracy is the most important performance metric for motion tracking systems. As shown in Table 4, there is a comparison of the accuracy of various cutting-edge MOT algorithms using the MOT17 dataset. Three metrics are used to calculate accuracy: MOTA, HOTA, and IDF1. The MOTA metric takes measurements of the overall tracking performance of the system, considering false positives, false negatives, and identity switches. HOTA provides a more balanced view of the system's detection and association quality. The IDF1 score considers how well the system performs its identifications over time. As can be seen in the table, transformer and graph-based models such as CTGMOT and MOTRV2 reach over 92% MOTA scores, indicating higher MOTA scores compared to DeepSORT and other older methods on the MOT17 dataset.

Table 4. Comparison of MOT Methods

Method	MOTA	IDF1	HOTA
DeepSORT [20]	69.1	59.2	52.1
FairMOT [42]	73.7	72.3	59.3
CTGMOT [30]	98.1	75.2	74.6
StrongSORT [35]	77.8	75.2	61.3
QDtrack [40]	68.7	66.3	53.9
MOTRV2 [10]	91.9	71.7	69.9
ByteTrack [40]	80.3	77.3	63.1
TransTrack [47]	92.6	71.5	68.9

8.3. Applicability by Domain

The applications of MOT and the technologies employed in each discipline vary. Hybrid approaches are used in autonomous cars, combining sensors like cameras and LiDAR with advanced detection and tracking algorithms to enable accurate, real-time tracking. Security surveillance systems rely on DeepSORT and multi-camera re-identification and tracking technology to ensure continuity across diverse situations. Graphs, transducers, and positional estimation are used in sports analysis to monitor players and assess performance. To support surgical and kinetic applications, the healthcare industry concentrates on privacy-preserving tracking technologies. To comprehend client mobility while abiding by data protection laws, tracking technologies are utilized in the retail industry in conjunction with behavioral analysis and heat maps. Table 5 presents the application domains of MOT systems along with their associated techniques and technologies.

Table 5. Application Domains of MOT Systems with Associated Techniques and Technologies

Ref	Domain	Used Techniques	Tools and Technologies
[5] [6] [8] [26] [38] [49] [50] [51] [55]	Autonomous Vehicles	Hybrid (Deep Learning + Kalman), Sensor Fusion	YOLOv8, StrongSORT, Radar, LiDAR
[13] [29] [32] [48] [52] [56]	Security Surveillance	DeepSORT, YOLO, Re-identification, multi-Camera	FairMOT, LMOT Dataset, Adaptive Kalman Filter
[23] [30] [47] [54]	Sports Analytics	GNNs, Transformers, Pose Estimation	CTGMOT, OpenPose, SportsMOT
[2] [57]	Healthcare	YOLOv5 + StrongSORT, RFID, Pose Estimation	AI with privacy-awareness, Surgical tracking
[58] [59]	Retail	YOLOv5 + DeepSORT, LSTM, Heatmaps, Behavioral Analytics, GDPR Compliance	YOLOv5, DeepSORT, LSTM, and Advanced behavioral tracking algorithms

9. Research Gaps and Future Directions

9.1. Research Gaps

Object tracking systems face several challenges that affect their overall accuracy. These methods have low robustness to adverse environmental conditions, such as darkness and rain, resulting in degraded image quality and poor tracking accuracy. Interestingly, deep learning methods rely heavily on labelled data, which may not be sufficient when the new environment is partially unknown. In addition, a few evaluation measures are suitable for evaluating the overall performance of a system in real-world scenarios. There are also significant ethical issues to attend to, such as privacy concerns and bias in the model, and ensuring compliance with laws such as the GDPR.

9.2. Future Directions

MOT algorithms have been transformed into suitable edge devices using lightweight model-based approaches to achieve accurate results with minimal power consumption. Increasingly, machine learning and low-data acquisition are employed to develop models that learn quickly without the need for large, labelled datasets. Natural Language Processing systems that accept user input of requests, such as "follow the red person," are integrated and provide the ability for interactive commands, increasing interaction with the system. Emphasis is being placed on XAI (Explainable Artificial Intelligence) to develop more transparent and reliable models. Event cameras and Eurovision can improve the speed and accuracy of monitoring in dynamic environments. Other unified frameworks that integrate detection, segmentation, and tracking into a single architecture and improve spatial-temporal consistency while preserving simplicity have also been proposed.

10. Conclusion

Deep learning has significantly improved motion tracking technology. The use of various sensors, such as LiDAR, radar, and cameras, has also improved the accuracy of tracking systems. However, there are still various challenges in the field of MOT that need to be overcome. Furthermore, there are some critical ethical and privacy challenges that need to be overcome in both technical and legal methods. Finally, the use of MOT systems in various domains such as self-driving cars, security, and healthcare will create substantial improvements in future applications of these systems.

References

- [1] N. Zhou, L. Lau, R. Bai, and T. Moore, "A genetic optimization resampling-based particle filtering algorithm for indoor target tracking," *Remote Sensing*, vol. 13, no. 1, pp. 1–22, Jan. 2021, doi: 10.3390/rs13010132.
- [2] L. Liu, Y. Dai, and Z. Liu, "Real-time pose estimation and motion tracking for motion performance using deep learning models," *Journal of Intelligent Systems*, vol. 33, no. 1, Jan. 2024, doi: 10.1515/jisys-2023-0288.
- [3] J. Shen and H. Yang, "Multi-object Tracking Model Based on Detection Tracking Paradigm in Panoramic Scenes," *Applied Sciences*, vol. 14, no. 10, May 2024, doi: 10.3390/app14104146.
- [4] H. Liu, "Research Advanced in 2D Multi-object Tracking based on Deep Learning," *Highlights in Science, Engineering and Technology*, vol. 72, 2024, doi: 10.54097/h7jpeh64.
- [5] W. Farhat, O. Ben Rhaïem, H. Faïedh, and C. Souani, "Pedestrian detection and tracking using an enhanced YOLOv9 model for automotive vehicles," *Measurement*, vol. 254, Oct. 2025, doi: 10.1016/j.measurement.2025.118009.

- [6] X. Xiao and X. Feng, "Multi-object Pedestrian Tracking Using Improved YOLOv8 and OC-SORT," *Sensors*, vol. 23, no. 20, p. 8439, Oct. 2023, doi: 10.3390/s23208439.
- [7] Y. Liu, B. Li, D. Liu, and Q. Duan, "Adaptive spatial aggregation and viewpoint alignment for three-dimensional online multiple fish tracking," *Computers and Electronics in Agriculture*, vol. 236, Sep. 2025, doi: 10.1016/j.compag.2025.110408.
- [8] K. Salazar and H. V. Li, "Multi-object Tracking Sensor Fusion: Radar and LiDAR Integration with Kalman Filtering Techniques: Beyond Sight: Radar-LiDAR Fusion for Robust Multi-Object Tracking," Master of Science thesis, Department of Electrical Engineering, Uppsala University, Uppsala, Sweden, 2024.
- [9] X. Wang, K. Ma, Q. Liu, Y. Zou, and Y. Fu, "Multi-object tracking in the dark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 382–392.
- [10] L. Sun, B. Li, D. Gao, and B. Fan, "Adaptive multi-object tracking algorithm based on split trajectory," *The Journal of Supercomputing*, vol. 80, no. 15, pp. 22287–22314, Oct. 2024, doi: 10.1007/s11227-024-06285-5.
- [11] E. E. Gohari, R. Ramezani, and A. Baraani, "Adaptive probabilistic multi-model framework for robust 3D multi-object tracking under uncertainty," *Expert Systems with Applications*, vol. 272, May 2025, doi: 10.1016/j.eswa.2025.126719.
- [12] G. S. Walia, A. Kumar, A. Saxena, K. Sharma, and K. Singh, "Robust object tracking with crowd search optimized multi-cue particle filter," *Pattern Analysis and Applications*, vol. 23, no. 3, pp. 1127–1144, Aug. 2020, doi: 10.1007/s10044-019-00847-7.
- [13] A. Ratre, "GMM-based Imbalanced Fractional Whale Particle Filter for Multiple Object Tracking in Surveillance Videos," *International Journal of Computer Network and Information Security*, vol. 17, no. 2, pp. 34–50, Apr. 2025, doi: 10.5815/ijenis.2025.02.03.
- [14] M. A. Awal, M. A. R. Refat, F. Naznin, and M. Z. Islam, "A particle filter based visual object tracking: A systematic review of current trends and research challenges," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, pp. 1291–1300, 2023, doi: 10.14569/IJACSA.2023.01411131.
- [15] X. Bian, W. Zhao, L. Tang, H. Zhao, and X. Mei, "An Enhanced Particle Filtering Method Leveraging Particle Swarm Optimization for Simultaneous Localization and Mapping in Mobile Robots Navigating Unknown Environments," *Applied Sciences*, vol. 14, no. 20, Oct. 2024, doi: 10.3390/app14209426.
- [16] S. Chen and X. Ren, "Multi-object tracking algorithm based on improved TrackFormer," in *Association for Computing Machinery (ACM) International Conference Proceeding Series*, Nov. 2023, pp. 194–202, doi: 10.1145/3652628.3652661.
- [17] S. Wu, A. Hadachi, C. Lu, and D. Vivet, "MOTT: A new model for multi-object tracking based on green learning paradigm," *AI Open*, vol. 4, pp. 145–153, Jan. 2023, doi: 10.1016/j.aiopen.2023.09.002.
- [18] G. Zhao, F. Meng, C. Yang, H. Wei, D. Zhang, and Z. Zheng, "A review of object tracking based on deep learning," *Neurocomputing*, vol. 651, p. 130988, Oct. 2025, doi: 10.1016/j.neucom.2025.130988.
- [19] R. Zeng and M. Ramezani, "A dynamic-confidence 3D multi-object tracking method based on spatio-temporal association," *Measurement*, vol. 256, Dec. 2025, doi: 10.1016/j.measurement.2025.117864.
- [20] L. Zhang et al., "UniRTL: A universal RGBT and low-light benchmark for object tracking," *Pattern Recognition*, vol. 158, Feb. 2025, doi: 10.1016/j.patcog.2024.110984.
- [21] A. Prasannakumar and D. Mishra, "Deep Efficient Data Association for Multi-object Tracking: Augmented with SSIM-Based Ambiguity Elimination," *Journal of Imaging*, vol. 10, no. 7, Jul. 2024, doi: 10.3390/jimaging10070171.
- [22] J. Cao, Y. Peng, H. Wei, L. Mo, L. Fan, and L. Wang, "KPTR: Key point transformer for LiDAR-based 3D object detection," *Measurement*, vol. 242, Jan. 2025, doi: 10.1016/j.measurement.2024.115820.
- [23] H.-W. Huang et al., "Iterative scale-up expansion and deep feature association for multi-object tracking in sports," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 163–172.
- [24] S. P. K. Reddy, J. Harikiran, and B. S. Chandana, "Deep CNN Based Multi-object Detection and Tracking in Video Frames with Mean Distributed Feature Set," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 723–734, doi: 10.1016/j.procs.2024.04.069.
- [25] M. Li, M. Liu, W. Zhang, W. Guo, E. Chen, and C. Zhang, "A Robust Multi-Camera Vehicle Tracking Algorithm in Highway Scenarios Using Deep Learning," *Applied Sciences*, vol. 14, no. 16, Aug. 2024, doi: 10.3390/app14167071.
- [26] L. Cheng, A. Sengupta, and S. Cao, "Deep Learning-Based Robust Multi-object Tracking via Fusion of mmWave Radar and Camera Sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 11, pp. 17218–17233, 2024, doi: 10.1109/TITS.2024.3421339.
- [27] M. Li, Y. Zhang, Y. Jia, and Y. Yang, "Advancing multi-object tracking through occlusion-awareness and trajectory optimization," *Knowledge-Based Systems*, vol. 310, Feb. 2025, doi: 10.1016/j.knosys.2024.112930.
- [28] D. Wu, W. Han, T. Wang, X. Dong, X. Zhang, and J. Shen, "Referring Multi-object Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14633–14642, doi: 10.1109/CVPR52729.2023.01406.
- [29] C. Yang, A. Xu, Y. Li, and X. Wu, "Multi-pedestrian tracking based on iterative filtering and observation discrimination," *Signal, Image and Video Processing*, vol. 19, no. 11, Nov. 2025, doi: 10.1007/s11760-025-04504-x.
- [30] W. Cao, X. Wang, X. Liu, and Y. Xu, "A deep learning framework for multi-object tracking in team sports videos," *IET Computer Vision*, vol. 18, no. 5, pp. 574–590, Aug. 2024, doi: 10.1049/cvi2.12266.
- [31] F. Ngeni, J. Mwakalonge, and S. Siuhi, "Solving traffic data occlusion problems in computer vision algorithms using DeepSORT and quantum computing," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 11, no. 1, pp. 1–15, Feb. 2024, doi: 10.1016/j.jtte.2023.05.006.
- [32] Y. Li, L. Zhan, L. Wu, H. Liu, J. Min, and X. Wang, "Re-identification assistance and multi-stage association for pedestrian multi-object tracking," *Scientific Reports*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-07276-z.
- [33] J. Liu, "RainyTrack: Enhancing Object Tracking in Adverse Weather Conditions with Siamese Networks," in *Association for Computing Machinery (ACM) International Conference Proceeding Series*, Jan. 2024, pp. 308–315, doi: 10.1145/3647649.3647698.
- [34] H. Pang, R. Ma, J. Su, C. Liu, Y. Gao, and Q. Jin, "Blinding and blurring the multi-object tracker with adversarial perturbations," *Neural Networks*, vol. 176, Aug. 2024, doi: 10.1016/j.neunet.2024.106331.
- [35] K. Ren, C. Hu, H. Xi, Y. Li, J. Fan, and L. Liu, "SemTG-Track: Multi-modal fine-grained semantic-unit temporal guidance for multi-object tracking," *Expert Systems with Applications*, vol. 289, Sep. 2025, doi: 10.1016/j.eswa.2025.128359.
- [36] X. Feng, Y. Liu, H. Yang, X. Jiao, and Z. Liu, "Self-supervised multi-object tracking based on metric learning," *Complex and Intelligent Systems*, vol. 10, no. 5, pp. 7077–7088, Oct. 2024, doi: 10.1007/s40747-024-01475-3.
- [37] X. Wang, Z. Sun, A. Chehri, G. Jeon, and Y. Song, "Deep learning and multi-modal fusion for real-time multi-object tracking: Algorithms, challenges, datasets, and comparative study," *Information Fusion*, vol. 105, May 2024, doi: 10.1016/j.inffus.2024.102247.
- [38] M. Á. Naya, E. Sanjurjo, A. J. Rodríguez, and J. Cuadrado, "Kalman filters based on multi-body models: linking simulation and real world. A comprehensive review," *Multibody System Dynamics*, vol. 58, no. 3–4, pp. 241–280, Aug. 2023, doi: 10.1007/s11044-023-09893-w.
- [39] P. Gao, S. M. Li, F. Wang, H. Fujita, H. Aljuaid, and R. Y. Yuan, "Learning multi-level graph attentional representation for thermal infrared object tracking," *Engineering Applications of Artificial Intelligence*, vol. 155, Sep. 2025, doi: 10.1016/j.engappai.2025.110957.

-
- [40] Y. Wang, Y. Qing, K. Huang, C. Dang, and Z. Wu, "Preformer MOT: A transformer-based approach for multi-object tracking with global trajectory prediction," *Fundamental Research*, 2025, doi: 10.1016/j.fmre.2024.06.015.
 - [41] R. K. Saha, R. Singhal, and M. Nambiar, "CAMOT: Content Aware Multi-object Tracking," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Oct. 2023, doi: 10.1145/3639856.3639866.
 - [42] X. Guo, Y. Zheng, and D. Wang, "PMTrack: Multi-object Tracking with Motion-Aware," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2024, pp. 3091-3106.
 - [43] R. N. Razak and H. N. Abdullah, "Improving multi-object detection and tracking with deep learning, DeepSORT, and frame cancellation techniques," *Open Engineering*, vol. 14, no. 1, Jan. 2024, doi: 10.1515/eng-2024-0056.
 - [44] M. Sukkar, M. Shukla, D. Kumar, V. C. Gerogiannis, A. Kanavos, and B. Acharya, "Enhancing Pedestrian Tracking in Autonomous Vehicles by Using Advanced Deep Learning Techniques," *Information (Switzerland)*, vol. 15, no. 2, Feb. 2024, doi: 10.3390/info15020104.
 - [45] I. Daramouskas, D. Meimetis, V. Patrinooulou, V. Lappas, V. Kostopoulos, and V. Kapoulas, "Camera-Based Local and Global Target Detection, Tracking, and Localization Techniques for UAVs," *Machines*, vol. 11, no. 2, Feb. 2023, doi: 10.3390/machines11020315.
 - [46] S. Meng, D. Shao, J. Guo, and S. Gao, "Tracking without label: Unsupervised multiple object tracking via contrastive similarity learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 16264-16273.
 - [47] Y. Cui, C. Zeng, X. Zhao, Y. Yang, G. Wu, and L. Wang, "SportsMOT: A large multi-object tracking dataset in multiple sports scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 9921-9931.
 - [48] Y. Zhu, Z. Liu, K. Li, Y. Li, X. Qi, and N. Ding, "Research on Multi-Objective Pedestrian Tracking Algorithm Based on Full-Size Feature Fusion," *IEEE Access*, vol. 13, pp. 39107-39116, 2025, doi: 10.1109/ACCESS.2025.3546818.
 - [49] Z. Meng, X. Xia, R. Xu, W. Liu, and J. Ma, "HYDRO-3D: Hybrid Object Detection and Tracking for Cooperative Perception Using 3D LiDAR," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 8, pp. 4069-4080, Aug. 2023, doi: 10.1109/TIV.2023.3282567.
 - [50] N. Senel, K. Kefferpütz, K. Doycheva, and G. Elger, "Multi-Sensor Data Fusion for Real-Time Multi-object Tracking," *Processes*, vol. 11, no. 2, Feb. 2023, doi: 10.3390/pr11020501.
 - [51] K. Vinoth and P. Sasikumar, "Multi-sensor fusion and segmentation for autonomous vehicle multi-object tracking using deep Q networks," *Scientific Reports*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-82356-0.
 - [52] S. Ren, "Multi-camera association tracking algorithm for pedestrian target based on difference image," *Systems and Soft Computing*, vol. 7, Dec. 2025, doi: 10.1016/j.sasc.2025.200282.
 - [53] Y. Guo, Y. Li, S. Wang, K. Sun, M. Liu, and Z. Wang, "Pedestrian multi-object tracking combining appearance and spatial characteristics," *Expert Systems with Applications*, vol. 272, May 2025, doi: 10.1016/j.eswa.2025.126772.
 - [54] Q. Hu, A. Scott, C. Yeung, and K. Fujii, "Basketball-SORT: an association method for complex multi-object occlusion problems in basketball multi-object tracking," *Multimedia Tools and Applications*, vol. 83, no. 38, pp. 86281-86297, Oct. 2024, doi: 10.1007/s11042-024-20360-2.
 - [55] F. Li, Y. Chen, M. Hu, M. Luo, and G. Wang, "Helmet-Wearing Tracking Detection Based on StrongSORT," *Sensors*, vol. 23, no. 3, Feb. 2023, doi: 10.3390/s23031682.
 - [56] V. K. Gupta, P. Rathore, P. Prajapati, and R. Shukla, "Automated PPE Detection Using YOLOv8 for Real-Time Workplace Safety Monitoring," *International Journal of Innovative Science and Research Technology*, pp. 2228-2233, May 2025, doi: 10.38124/ijisrt/25apr1699.
 - [57] S. Lu et al., "Surgical instrument posture estimation and tracking based on LSTM," *ICT Express*, vol. 10, no. 3, pp. 465-471, Jun. 2024, doi: 10.1016/j.ict.2024.01.002.
 - [58] M. Shili, S. Jayasingh, and S. Hammedi, "Advanced Customer Behavior Tracking and Heatmap Analysis with YOLOv5 and DeepSORT in Retail Environment," *Electronics*, vol. 13, no. 23, Dec. 2024, doi: 10.3390/electronics13234730.
 - [59] T. D. Nguyen et al., "A retail store customer behavior analysis system: Design and implementation," in *Artificial Intelligence Applications and Innovations (AIAI 2024)*, Cham: Springer, 2024, pp. 305-318, doi: 10.1007/978-3-031-63223-5_23.
 - [60] R. Irampaye, J. Chen, X. Zhu, and Y. Zhou, "Visual multi-object tracking by interaction network," in *Proceedings of the 2024 7th International Symposium on Autonomous Systems (ISAS)*, Institute of Electrical and Electronics Engineers (IEEE), 2024, doi: 10.1109/ISAS61044.2024.10552490.
 - [61] S. Li, H. Ren, Y. Cao, et al., "A Review of Multi-Object Tracking in Recent Times," *IET Computer Vision*, vol. 18, no. 1, pp. 1-25, Feb. 2024, doi: 10.1049/cvi2.12231.
 - [62] A. Ramachandran and A. K. Sangaiah, "A review on object detection in unmanned aerial vehicle surveillance," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 215-228, 2021, doi: 10.1016/j.ijcce.2021.11.