IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 1, 2021, accepted February 15, 2021, date of publication February 22, 2021, date of current version March 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3060821

Analysis Based on Recent Deep Learning Approaches Applied in Real-Time Multi-Object Tracking: A Review

LESOLE KALAKE [1], WANGGEN WAN [1], (Senior Member, IEEE), AND LI HOU [2]
[1] School of Communications and Information Engineering, Institute of Smart City, Shanghai University, Shanghai 200444, China
[2] School of Information Engineering, Huangshan University, Huangshan 245041, China

Corresponding author: Lesole Kalake (tumelok1@shu.edu.cn)

This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant 18510760300, in part by the Anhui Natural Science Foundation under Grant 1908085MF178, in part by the China Postdoctoral Science Foundation under Grant 2020M681264, and in part by the Anhui Excellent Young Talents Support Program under Project gxyqZD2019069.

ABSTRACT The deep learning technique has proven to be effective in the classification and localization of objects on the image or ground plane over time. The strength of the technique's features has enabled researchers to analyze object trajectories across multiple cameras for online multi-object tracking (MOT) systems. In the past five years, these technical features have gained a reputation in handling several real-time multiple object tracking challenges. This contributed to the increasing number of proposed deep learning methods (DLMs) and networks seen by the computer vision community. The technique efficiently handled various challenges in real-time MOT systems and improved overall tracking performance. However, it experienced difficulties in the detection and tracking of objects in overcrowded scenes and motion variations and confused appearance variations. Therefore, in this paper, we summarize and analyze the 95 contributions made in the past five years on deep learning-based online MOT methods and networks that rank highest in the public benchmark. We review their expedition, performance, advantages, and challenges under different experimental setups and tracking conditions. We also further categorize these methods and networks into four main themes: Online MOT Based Detection Quality and Associations, Real-Time MOT with High-Speed Tracking and Low Computational Costs, Modeling Target Uncertainty in Online MOT, and Deep Convolutional Neural Network (DCNN), Affinity and Data Association. Finally, we discuss the ongoing challenges and directions for future research.

INDEX TERMS Deep learning, detection quality, high-speed tracking, multi-camera object tracking, real-time tracking.

I. INTRODUCTION

In the past five years, deep learning-based online multi-object tracking (MOT) paradigms have been inferior to sparse principal component analysis [1], [2]. The emergence and expansion of convolutional neural networks (CNNs) to DCNNs strengthened DLMs and tracking-by-detection (TBDs), thus contributing to discernible progress in online MOTs [3]–[7]. The DCNN features and neural layers were used to detect and track countless objects that move on the streets and public spaces [8], [9]. In contrast, the TBD is used to optimize the tracker's discriminative model, locate the target in the next frame based on detection results, and then generate and link object tracklets accordingly [3], [10]. This improved and strengthened the detection and tracking processes to address the challenges of online MOTs using multiple cameras. It also gradually expanded deep learning approaches in real-time MOTs based on the single-camera tracking technique. However, the approaches implemented with the single-camera tracking technique seemed more effective for offline MOT [11], [12] and harmed many algorithms due to the view angle. The view angle had limitations and could not provide multiple angles, hence making the single-camera technique's algorithms susceptible to velocity variations and vulnerable to misdetections, occlusions, and fragmentations [13] due to both camera and object

The associate editor coordinating the review of this manuscript and approving it for publication was Charith Abhayaratne.

movements [14], [15]. This ineffectively localized multiple objects, extracted features, created bounding box regression detections, generated tracklets, and contributed to inappropriate matching or mapping of the specific appearance information [6], [16], [17].

Currently, researchers [5], [11], [18], [19] have summarized only the multi-object tracking literature predicated on general visual tracking and detection techniques based on experimental studies rather than concentrating on deep learning methods based on online MOT. In the past five years, several proposed approaches have shown a significant performance enhancement in real-time MOT and were able to approximate human vision. They have impressively promoted tracking performance by reducing the misdetection rate with the integration of a tracking-by-detection paradigm [20]–[24]. This led to the emergence of various efficient and robust algorithms with minimum real-time tracking challenges and complications in video data processing [1], [5]. Therefore, it is important to summarize and analyze the existing DLMs and network-based online MOTs to pave the way for further studies. Hence, the present paper presents a systematic review of progress, challenges, and future research opportunities on DLM-based online multi-object tracking applications. It further compares and discusses how they enhanced the performance in online MOTs with various public datasets in various environmental setups. It then discusses the main functionalities and implementation strategies in detail.

This paper is organized as follows: Section I provides a brief background on online multiple object tracking (MOT) and problem formulations. Section II presents the methodology for gathering relevant works. Section III discusses the extensive literature by considering deep learning-based online multi-object tracking methods' advantages and persisting challenges. Section IV discusses the effectiveness of deep learning based on categorized themes: deep learning towards online multi-object tracking based on detection quality and associations online MOT-based detection quality and associations, real-time MOT with high-speed tracking and low computational costs, modeling target uncertainty in online MOT, convolutional neural networks (CNNs), and affinity and data associations. Section V concludes the study.

A. ONLINE MULTI-OBJECT TRACKING (MOT) PROBLEM FORMATION

Online multi-object tracking (MOT) is the variation of problem estimations based on the given input video sequence with several moving objects in frames [21]. It plays an essential role in video surveillance applications by locating moving objects in the video frames taken by either a single camera or multiple networked cameras. It forms the process of detecting, locating, associating, and tracking objects over a period by collecting the observations from the initial frame until the last-end frame. Then optimizes the sequential states by modeling the maximum posterior estimation from the conditional for all sequential states of all objects from the first frame to the last frame [25]. Wen *et al.* [26] capitalized on this theorem by creating CLEAR MOT evaluation metrics that have been implemented in neoteric work on deep learning-based real-time MOT methods, multi-camera tracking techniques (MCTs), and DCNNs with the tracking-by-detection (TBD) approach to track objects across multiple frames [19], [26]. These evaluation metrics enabled the standard calculations and presentation of multiple object tracking results on false positive (FP), false negative (FN), false alarm (FA), fragments of target trajectories (FM), multi-object tracking accuracy (MOTA), and multi-object tracking precision (MOTP) of public datasets created based on both single camera and multi-camera video capturing on different environmental scenes. Therefore, it was necessary for Wen *et al.* [26] to further benchmark and define the CLEAR MOT metric formulas for both MOTA and MOTP as follows:

$$MOTA = 1 - \frac{\sum_v \sum_t \left(FN_{v,t} + FP_{v,t} + IDS_{v,t}\right)}{\sum_v \sum_t GT_{v,t}} \tag{1}$$

where $FN_{v,t}$ and $FP_{v,t}$ denote false negatives and false positives, respectively. Then, $IDS_{v,t}$ represent identity switches of trajectories, and $GT_{v,t}$ is the number of ground truth objects at time index t of sequence v. Then, MOTP metrics as the average dissimilarities between true positives and ground truth:

$$MOTP = \frac{\sum_{i,t} d_i^t}{\sum_i c_t} \tag{2}$$

where c_t denotes the number of matches in frame t and d_i^t is the bounding box overlap per frame target with its assigned ground truth objects.

B. TRADITIONAL SINGLE-CAMERA MULTI-OBJECT TRACKING

The single-camera tracking (SCT) technique, as illustrated in Fig. 1, is a cost-inefficient traditional technical method used to detect multiple views of different objects. It enables the enhancement of trackers to track multiple objects in a video frame sequence based on the detection quality [27]. However, it provides a one-sided view and cannot provide multiple views due to its limitations in handling rotations, scaling, affinity distortions, quick movements, similarities, and occlusions [28], [29]. These limitations led to degraded overall detector performance, and Lee and Hong [30] incorporated separate detectors and classifiers for several different viewpoints to improve the detector performance.

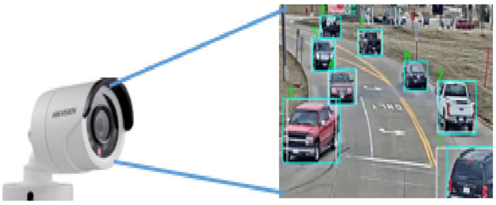


FIGURE 1. Single Camera Multi-Object Tracking Overview [19].

However, the combination could not produce satisfactory detection results due to difficulties in handling occlusions and misdetections on each detector and classifier [5]. Then Fajardo *et al.* [32], Azad and Misbahuddin. [31] further contributed to enhance the detector's performance by labeling the objects on the output of the maximal classifiers. They stretched and reinvigorated the algorithm by estimating the object distance and detection with tracking-by-detection in a deep convolutional neural network (DCNN). They utilized the network layers to extract features from the input video frame sequences with learnable filters and added biases from the parameters of each layer. Then, these filters and biases are represented by $w = \sum_{i=1}^{k} w_i$ and $\sum_{i=1}^{k} bi$, respectively. The generated feature map was represented by X_k and used to pass the results to the next layer as an element of σ repeatedly on each convolutional layer.

$$X_k^t = \sigma\left(W_k^{t-1} \cdot X^{t-1} + b_k^{t-1}\right) \tag{3}$$

The approach successfully overcame tracklet loss by handling multiple object new identities (IDs) and reassigning issues [14]. However, the rotation and one side view in the single-camera technique [33] contributed to the lack of robustness and difficulties in handling long occlusions. This resulted in high fragmentation, velocity changes, and appearance changes [24]. The challenges caused the splitting of camera object tracking into two tasks, i.e., (SCT) and inter-camera object tracking (ICT) [23]. Then, SCT is used to obtain multi object trajectories in a single camera view connected across multiple camera views through ICT [14]. Therefore, this laid a solid foundation for DLMs with MCT techniques based on online MOT [34], [35].

C. MULTI-CAMERA FOR MULTI-OBJECT TRACKING (MCT)

The technique has capitalized on the foundation laid with SCT approaches [35]. It uses ICT, as shown in Fig. 2, to capture the object across each camera on different angle views despite the velocity and appearance variations [14]. The object detection from different camera views is associated with trajectory objects and tracked independently on each camera view [36]. Then, the velocity and position of object features are computed by grouping trajectories into one cluster that enables the connection among the camera views to handle variations in motion, speed, and direction [37]. However, the multi-camera multi-object tracking technique needs to maintain the identity consistency of each target across multiple views and struggles when there is an object similarity appearance [38], [95]. In this case, the current most deep learning-based online MOT systems introduced the DCNN and tracking-by-detection (TBD) paradigm to solve the problem of associating a target with multiple potential views [39]. They are designed in an end-to-end deep neural network to learn the association between tracks and detections, statement updates, initialization, and termination of tracks [40]. They are further employed in the real-time tracking framework so that the associations between tracklets and detections are cascaded from high-confidence tracklets to low-confidence tracklets [41].

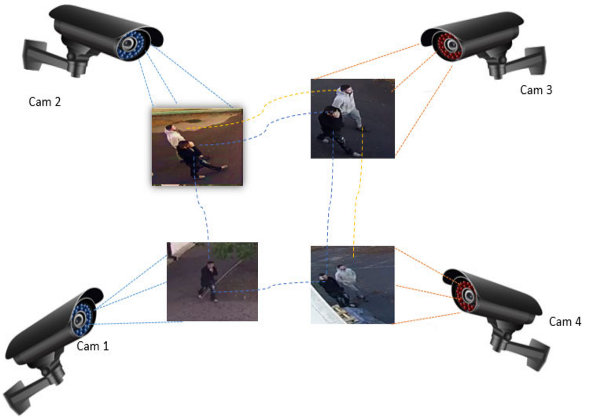


FIGURE 2. Multi-Camera for Multi-Object Multicamera for Multi-object Tracking (MCT) Overview.

II. METHODOLOGY

We performed two systematic electronic searches in Google Scholar and Web of Science according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [42]. An extensive database search was conducted via expression with most essential terms such as "Multi-Object Tracking", "Real-time Multi-Object Tracking", "Deep Learning Object Tracking", "Online Multi-Object Tracking", "High-Speed Tracking", "Deep Convolution Neural Network", and "Target Detection and Tracking" over the last 5 years, from 2015-2020. The final search in these databases was performed on the 25th of July 2020 and was restricted to peer-reviewed documents, such as journals and conference papers. Then, 80 duplicates within the retrieved articles in the databases were removed.

As depicted in Fig. 3, we initialized the search expression with the diverse coalescence of key terms such as "Multi-Object Tracking, Target Detection, Tracking, and Real-time Multi-Object Tracking " that were used on Web of Science and Google Scholar and this returned 5000 articles. We further intensified the search expression by adding "Online Multi-Object Tracking" and 1,500 articles with duplicates were returned. We further restricted and reinvigorated both the search expressions and filter by adding the "Deep Convolutional Neural Network (DCNN), High-Speed Tracking and publications' range period (2015-2020) and screened the outcomes (180 articles) to eliminate duplicates while ensuring authenticity and competency. Therefore, the study reviewed 95 peer-reviewed papers published within the past five years and supported the DLM-based online MOT, tracking-by-detection, and DCNN.

III. EVALUATIONS OF DEEP LEARNING METHODS BASED ONLINE MOT

In this section, we explore the DLM-based MOT framework and proven records. The deep learning framework effectively

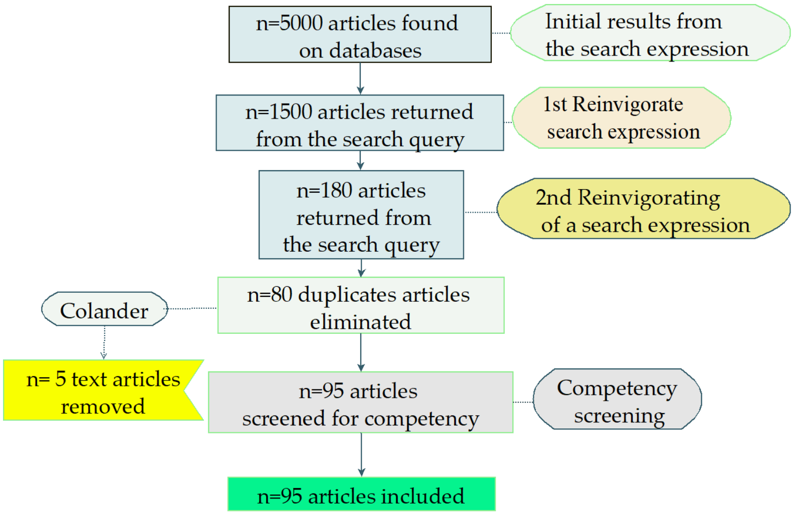

n=5000 articles found on databases
Initial results from the search expression
n=1500 articles returned from the search query
1st Reinvigorate search expression
n=180 articles returned from the search query
2nd Reinvigorating of a search expression
n=80 duplicates articles eliminated
Colander
n= 5 text articles removed
n=95 articles screened for competency
Competency screening
n=95 articles included

FIGURE 3. The approach used to extract articles predicated on the diverse coalescence of keys on a search expression.

improved the tracking performance from various tracking predictions and data associations [43], [44]. It automates the capacity learning of appearance features via DCNN to promote discrimination and robustness for occlusion handling in online tracking optimization strategies [45], [46]. Therefore, this has made DLMs more resourceful in promoting the accuracy of motion prediction and the performance of bipartite matching between tracklets and detection [1], [10], [47]. Thus, in Fig. 4, we categorized the approaches into four main themes based on the capabilities and objectives in dealing with various challenges in real-time MOT.

A. DEEP LEARNING TOWARDS ONLINE MULTI-OBJECT TRACKING BASED ON DETECTIONS QUALITY AND ASSOCIATIONS

The detection quality is significant to improve the tracker's capabilities in handling the object's appearance similarities, generating and associating the tracklets, reducing false object detections, calculating, grouping the similarity trajectories, and drifting [74], [75]. The TBD and most advanced DLMs primarily rely on the quality of detections to generate and associate the tracklets effectively, as depicted in Fig. 5 [3], [5], [6].

In this section, we used Table 1 and showed an overview of the DLMs that are integrated with CNN to increase the detection quality rate by breaking input video into frames [18], [47]. We further analyzed the anterior work on the deep learning technique towards real-time MOT.

Xiang *et al.* [47] proposed a multiple online object tracking decision-making strategy using template tracking, optical flow, and data association to strengthen a tracking-by-detection technique by handling the target dynamics and association history. They modeled the objects' similarity function by combining the different cues, appearance, location, and motion. This triplet loss-based CNN learned the distance metric between trackers and detections and used the long short-term memory (LSTM) prediction module to terminate the object in the next frame. However, the algorithm relied heavily on optic flow with template matching and resulted in poor detections and target data association history. It considered only motion features and omitted appearance features; hence, it experienced low detection quality (25.3%) and tracking results (30.3%) on the real-time MOT dataset (MOT2015). To overcome these shortcomings in detection quality, Milan *et al.* [16] extended the RNN and introduced a joint tracking and segmentation approach to estimate the state of the tracked object by strengthening the detector response. The network treats the states of objects, current observations, their matching matrix, and existence probabilities as inputs. Then it outputs the predicted states, updated results, and terminates the object based on new existence probabilities. The proposed algorithm further computes the matching matrix and groups the designed LSTM-based networks to model the matching process between one object's state and current observations. It then uses low-level image information and super-pixels to specify a target as background. This has enabled them to capitalize on the advantages of both high-level spatial information and low-level motion cues to create a unified graphical model for multi-object tracking and motion segmentation. It strengthened the algorithm to measure the object distance, size, location, and velocity through the implementation of the conditional random field (CRF) model. The strategy used a super-pixel procedure to assign labels to all pixels belonging to the semantic object on the video sequence. It assigned unique IDs to each detection at a super-pixel level in the input video. However, the approach improved the detection quality rate (76.0%) and struggled to accurately associate the target history in crowded scenes. Consequently, better tracking results (65.3%) were recorded in a real-time MOT evaluation of a dataset (PETS2015).

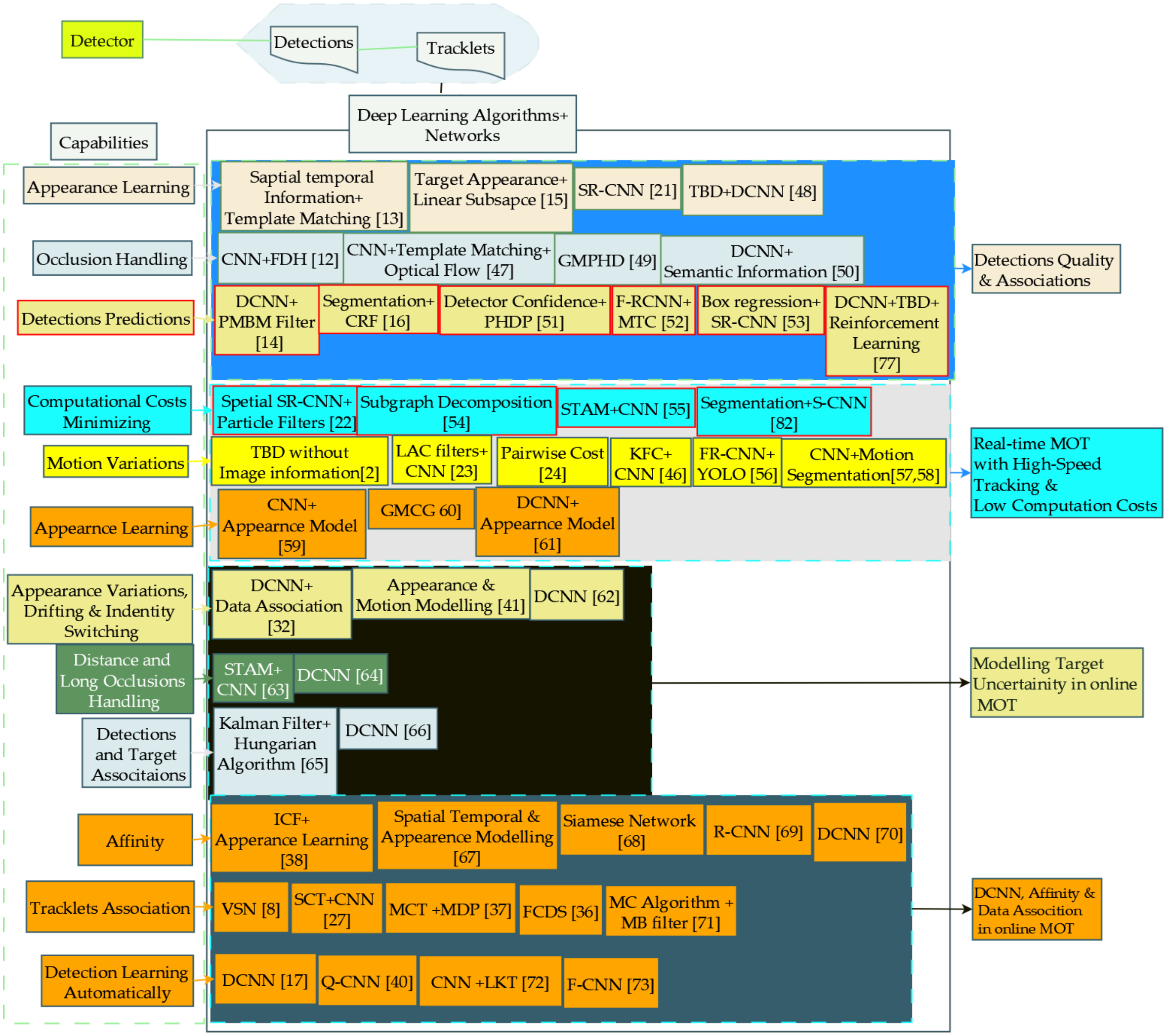

Detector
Detections
Tracklets
Deep Learning Algorithms+ Networks
Capabilities
Appearance Learning
Saptial temporal Information+ Template Matching [13]
Target Appearance+ Linear Subsapce [15]
SR-CNN [21]
TBD+DCNN [48]
Occlusion Handling
CNN+FDH [12]
CNN+Template Matching+ Optical Flow [47]
GMPHD [49]
DCNN+ Semantic Information [50]
Detections Quality & Associations
Detections Predictions
DCNN+ PMBM Filter [14]
Segmentation+ CRF [16]
Detector Confidence+ PHDP [51]
F-RCNN+ MTC [52]
Box regression+ SR-CNN [53]
DCNN+TBD+ Reinforcement Learning [77]
Computational Costs Minimizing
Spetial SR-CNN+ Particle Filters [22]
Subgraph Decomposition [54]
STAM+CNN [55]
Segmentation+S-CNN [82]
Motion Variations
TBD without Image information[2]
LAC filters+ CNN [23]
Pairwise Cost [24]
KFC+ CNN [46]
FR-CNN+ YOLO [56]
CNN+Motion Segmentation[57,58]
Real-time MOT with High-Speed Tracking & Low Computation Costs
Apperance Learning
CNN+ Apperance Model [59]
GMCG 60]
DCNN+ Appearance Model [61]
Appearance Variations, Drifting & Indentity Switching
DCNN+ Data Association [32]
Appearance & Motion Modelling [41]
DCNN [62]
Distance and Long Occlusions Handling
STAM+ CNN [63]
DCNN [64]
Modelling Target Uncertainity in online MOT
Detections and Target Associtaions
Kalman Filter+ Hungarian Algorithm [65]
DCNN [66]
Affinity
ICF+ Apperance Learning [38]
Spatial Temporal & Appearence Modelling [67]
Siamese Network [68]
R-CNN [69]
DCNN [70]
Tracklets Association
VSN [8]
SCT+CNN [27]
MCT +MDP [37]
FCDS [36]
MC Algorithm + MB filter [71]
DCNN, Affinity & Data Association in online MOT
Detection Learning Automatically
DCNN [17]
Q-CNN [40]
CNN +LKT [72]
F-CNN [73]

FIGURE 4. A framework of the DLM based MOTs investigated and categorized papers.

However, this result is not very convincing because the training samples were insufficient to learn an optimized model at once.

To manage these problems, Sanchez-Matilla *et al.* [51] used multiple detectors with high-end and low-end confidence values to improve tracking performance. They used weak (low confidence score) detections to support an existing track when robust detections were missing. Then, a perspective-dependent sampling mechanism is introduced to create newborn particles depending on their distance from the camera. They further used the probability hypothesis density particle (PHDP) framework to collect outputs from detectors. However, their approach failed to discriminate against a target in close range, resulted in low detection quality (14.0%) in real-time MOT evaluations of public datasets (MOT2015 and MOT2016), and could not improve the tracking accuracy (38.8%).

Kutschbach *et al.* [49] presented an application of the Gaussian mixture probability hypothesis density (GMPHD) filter for multi-object tracking in video data. They extended both the kernelized correlation filters and GMPHD to use the fast scale space tracking (FSST) scheme and two separated models for estimating target translation and scaling. The algorithm extracted the HOG feature from a region of 2.5 times its size and weighted it by a cosine window to highlight the target in the center and to avoid boundary issues. To associate detections with the tracks, the birth covariance was set to a significant value in every possible direction. However,

TABLE 1. Overview of deep learning methods used for online MOT based on detections quality and associations.

Algorithm	Training Data Size	Testing Data Size	Year	Technique	Description
Ray and Chakraborty [12]	VOT-2015: 22 videos ,(12568 Frames, (02:58 minutes))	VOT-2015:12 Videos, (6557 Frames, (01:36 minutes))	2017	Foreground Detection and History(FDH)	Use foreground detection and recent history of dissimilarities frames to strengthen the detector in variable background
Wei *et al.* [13]	PET-09:1 video(S1.L1), (795 Frames, (01:54 minutes))	PET-09:1 video (S2.L2), (436 Frames, (01:02:minutes))	2017	Spatiotemporal Information+ Template	Combine spatial and temporal models to determine the trajectory confidence in each frame.
Scheidegger *et al.* [14]	KITTI: 10 videos, (3235 Frames , (03:22 minutes))	KITTI: 7 videos, (1500 Frames, (01:47 minutes))	2018	DCNN+PMBM Filter	Generate trajectories of the detected object in a world coordinate through DCNN. Then, feed detections from sequential images into a PMBM filter
Fagot-Bouquet *et al.* [15]	MOT2015: 1 video (S1.L1), (795 Frames, (01:54 minutes))	MOT2015:1 video (S2.L2) ,(436 Frames, (01:02:minutes))	2016	Target Appearance+ Linear Subspace	Use a sparse linear combination dictionary of elements to handle object selection criterion error.
Milan *et al.* [16]	PETS2010: 6 videos, (1200 Frames , (02:26 minutes))	PET2010: 5 videos, (1000 Frames, (01:04 minutes))	2015	Segmentation+ CRF	Use joint tracking and segmentation approach to improve the detector's response, and CRF to define the object characteristics.
Ning *et al.* [21]	OTB-30: 22 videos, (11274 Frames, (06:48 minutes.))	OTB-30: 8 videos ,(4479 Frames, (01:18 minutes))	2017	SR-CNN	Use SR-CNN to learn and examine the historical locations and visual features of the past frame.
Xiang *et al.* [47]	MOT2015: 1 video (S1.L1),(795 Frames, (01:54 minutes))	MOT2015:1 video (S2.L2),(436 Frames, (01:02:minutes))	2015	Template Matching and Optical Flow	Uses template tracking, optical flow, and data association.
Sun *et al.* [48]	MOT2017: 7 videos, (5316 Frames, (01:97 minutes))	MOT2017: 3 videos, (2275 Frames, (01:21 minutes))	2019	TBD+DCNN	Combine the appearance modeling, affinities, and network. Then, compute a reliable trajectory.
Kutschbach *et al.* [49]	UA-DETRAC: 2 videos, (274 Frames, (00:27 minutes))	UA-DETRAC: 1 video, (379 Frames, 00:37 minutes))	2017	GMPHD	Extend the kernelized correlation filters into GMPHD to improve the tracking accuracy.
Zhao *et al.* [50]	KITTI: 6 videos, (1767 Frames, (01:33 minutes)	KITTI: 15 videos, (664 Frames, (01:00 minutes)	2018	DCNN	Use the target-specific semantic information inherited from the detector to improve accuracy.
Sanchez-Matilla *et al.* [51]	MOT2015: 1 video (S1.L1) ,(795 Frames, (01:54 minutes))	MOT2015:1 video (S2.L2) (436 Frames, (01:02:minutes))	2016	Detector Confidence Levels +PHDP	Use the PHDP framework to collect outputs from multiple detectors.
Zhang *et al.* [52]	DukeMTMC: 4 videos, (1743 Frames,(02:59 minutes))	DukeMTMC 5 videos,(1567 Frames, (35:00 minutes))	2017	MCT+ Fast R-CNN	Propose multi-camera multi-object tracking by hierarchical clustering onto Fast R-CNN.
Li *et al.* [53]	OTB-30: 22 videos, (11274 Frames, (06:48 minutes.))	OTB-30: 8 videos ,(4479 Frames,(01:18 minutes))	2018	Template Branch+ Boxes Regression+ SRPN	Use template branch and bounding box regressions into SRPN to improve the detection speed.
Ren *et al.* [77]	MOT2016:7 videos, (5316 Frames ,(03:58 minutes)	MOT2016: 7 videos, (5919 Frames, 04:13 minutes))	2018	DCNN+ TBD+ Reinforcement Learning	Use reinforced learning and unified deep networks for object detection and prediction.

their extended GMPHD on the dataset (UA-DETRAC) could not handle the large birth covariance of the first track and resulted in preventing the initialization of new tracks. This led to delayed track extraction and lower tracking performance (14.5%) with a moderate detection quality rate (63.4%).

To improve the detection module ability to detect and extract the tracks in a timely manner Zhao *et al.* [50] proposed a compressed DCNN feature-based correlation filter and used semantic information [78] inherited from the detector. The approach integrated the two modules for the online MOT approach and enhanced the ability to reidentify (ReID) the tracked object once it is lost. It also generated proposals for small objects in deep layers with semantic information and hence reduced the false detection rate. However, the approach failed to crop the target's region of interest (ROI) in the detection stage and left the small object proposal generated in the shallow feature layers. This resulted in low computational complexity, a high misdetection rate that caused a low tracking accuracy (32.7%), and a moderate detection quality rate (57.2%) in a real-time MOT evaluation of a public dataset (KITTI). Then, Scheidegger *et al.* [14] proposed using DCNN and a Poisson Multi-Bernoulli Mixture (PMBM) filter to produce trajectories of the detected object in a world coordinate system. Their approach used a deep neural network to detect and estimate the distance of objects from a single input image. It fed the detections from the sequential images into a Poisson multi-Bernoulli mixture (PMBM) filter. Then, the existing single-short multi-box detector (SDD) was incorporated to strengthen the detection of small objects on deeper layers rather than shallow layers. This played a significant role in building a multi-scale object detector that effectively detected small objects with fewer false negatives on datasets (KITTI). Consequently, it improved the tracking accuracy (80.0%) and detection quality rate (91.0%) with a brawny estimation function that effectively calculated distances between objects.

The non-static surface gives the impression that motion in the background pixels affects the detection quality [3], [12], [29]. Hence, Ray and Chakraborty [12] used foreground detection and recent dissimilarity frames to strengthen the detector and track associations in a variable background. The approach separates background and foreground information and then removes flickering background or noise by analyzing the pseudo-motion-compensated on the current frame and preceding frames. It uses the estimation function to predict the state of an object and the Kalman filter to track the object. This solved the associations' issue under occlusions by refining the target region. Although the proposed approach increased the target detection and association across the video frames, it failed to track and differentiate small objects [79] with similarities under complex scenes on a real-time MOT dataset (VOT2016 and MOT2016) and achieved moderate tracking accuracy (51.2%) with a high detection quality rate (86.0%).

Extending the work on the detector and unique identifications of the target, Zhang *et al.* [52] introduced multi-camera multi-object tracking by hierarchical clustering into a Fast R-CNN framework. The approach implemented the hierarchical clustering algorithm to merge trajectories and proposed solving the object similarities via the appearance feature extraction process. However, it was too slow to track objects in a real-time evaluation of a dataset (DukeMTMC) due to hierarchal paths and hence failed to handle appearance variations with more frequent object identity changes. This led to a moderate performance in both tracking accuracy (54.1%) and detection quality rate (55.0%). Then, Li *et al.* [53] incorporated template branch and bounding box regression into a Siamese regional proposed network (SRPN) to improve the detection speed. The distance between tracklet pairs is learned via the extended Siamese network. The extended network extracted features for each detection in tracklets and transferred these features to bidirectional gated recurrent unit (GRU) networks. Then, the algorithm generated the tracklets and split them into short sub-tracklets according to the local distance between bidirectional GRU outputs. The sub-tracklets are reconnected to the long trajectories using similarities between temporal pooling global features. This helped jettison the outliners via a cosine window and a scale range. Consequently, it improved the detection quality rate (83.0%) with weak associations that led to a moderate performance on tracking accuracy (49.6%) in real-time MOT datasets (OTB2015).

Sun *et al.* [48] suggest tackling object appearance and data association issues with tracking-by-detection via DCNN. The approach combined appearance modeling, affinities, and networks to compute reliable trajectories and object associations on the current frame based on detections from multiple previous frames. The target appearances and affinities in a pair of video frames were jointly learned in an end-to-end fashion. This enabled the softmax layer of the network to separately look forward and backward in time for unidentifiable objects in the frame pairs. It also contributed to handling the appearance and disappearance of multiple objects between video frames. However, the approach's overall network did not make assumptions for the input frame pairs to appear consecutively in a video. Although this promoted robustness against object occlusions, it could not cope very well with the data association of object fashion in real-time MOT evaluations of datasets (MOT15, MOT17, and UA-DETRAC) with similarities in the frames that were at close locations in the scenes. Consequently, it degraded the detection quality rate (41.1%) with a better tracking performance (52.4%). To address this problem, Ren *et al.* [77] proposed a deep prediction-decision network in a collaborative deep reinforcement learning (C-DRL) method that simultaneously detected and predicted objects under a unified network via deep reinforcement learning. To solve target associations and location problems, the approach considered each object as an agent and tracked it via the prediction network. It further sought the optimal tracked results by exploiting the collaborative interactions of different agents and environments via the decision network. The network learned the object movement, size, speed, and direction [80] and predicted the next step of the target on the frame very well on real-time public datasets (MOT2015 and MOT2016). However, it becomes sensitive to appearance features and fragmentation, especially under long occlusions and heavy interactions. Hence, in some videos with high sampling rates, the approach experienced a high number of object losses for relatively more frames due to occlusion that led to a high number in IDS. This degraded both the tracking performance (47.3%) and detection quality rate (30.4%).

Wei *et al.* [13] developed the learning framework to address the issues in tracking and misdetections under heavy object interactions. They used temporal-spatial information to determine the trajectory confidence in each frame. The approach divides this process into a local and global association to associate the trajectories with high confidence with the detection result of the current frame to the local association and the one with low confidence with the detection results of the current frame that are not matched to the global association. This combination of spatial and temporal models of a public dataset (PET2009) enhanced the tracklet association and midsection in real-time object tracking. Compared to Ren *et al.* [77], the proposed algorithm improved the tracking accuracy by 9% with a low detection quality rate (17.0%).

Ning *et al.* [21] introduced a spatially recurrent convolutional neural network (SR-CNN) by extending the spatial and temporal work to learn visual features of the past frame by examining the historical locations. The approach tried to learn from historical visual semantics, detections, and tracklets by enabling automatic learning onto a tracker. It incorporated LSTM and enforced an end-to-end spatial-temporal regression with a single evaluation to enhance efficiency and effectiveness [81] by spatially glimpsing on various regions and regressing on the heat maps. However, the approach could not accurately link the tracklets and had hardly reidentified (RID) objects under prolonged occlusions. Consequently, it resulted in poor tracking accuracy (43.0%) and detection quality rate (17.0%) in a real-time MOT evaluation with a public dataset (OTB-30). Then, to locate and handle misdetection on a similar object, Fagot-Bouquet *et al.* [15] formulated a multi-frame data association process based on a sliding window and minimized energy sparsity that represented all detections. The technique implemented the TBD paradigm based on the sliding window and estimated trajectories for best associating object detections. However, when the number of frames increased on the sliding window, the appearance model suffered. Consequently, this led to a failure for the proposed approach to associate detection effectively. It also had trouble handling object tracking under crowded scenes on datasets (2DMOT2015 and MOT2016) and resulted in a deteriorated overall tracking performance.

B. DEEP LEARNING METHODS IN REAL-TIME MOTS WITH HIGH-SPEED TRACKING AND LOW COMPUTATIONAL COST

The slow algorithm tends to lose track of many tracked objects with speed variations. The object's speed assumptions seemed to be a common issue that mostly led to unsatisfactory overall tracking performance. Therefore, in this section, the deep learning methods in real-time MOTs with high-speed tracking and low computational cost presented in Table 2 are analyzed.

Zamir *et al.* [60] proposed an algorithm to solve data association issues via generalized minimum clique graphs by finding the detections that correspond to one particular object in different video frames. The approach has expanded the node definition for clustering by grouping the nodes of an input graph into disjoint clusters. It searches for a subgraph set of nodes that requires the minimum cost for the complete graph to be produced. This required the authors to introduce the hypothetical nodes technique that handled the exit or entry problem and long-term occlusion occurrences during the tracking process. However, the assumptions made on object velocity over a short period caused the algorithm to struggle in modeling the motion of one person over the long run without knowing the destination structure of the scene and especially when people were heavily interacting. To construct a hypothesis tree for multiple hypothesis association and tracking, Kim *et al.* [61] extended the MHT framework with appearance features using a multi-output regularized least square method. Their approach exploited high-order appearance information by incorporating long-term appearances via appearance feature extractions and deep neural networks. The appearance features are dimensionally reduced from deep dimensional features to handle appearance and motion variations during tracking. These features boosted the approach in handling exit-entry issues and long occlusions with high computational costs [44] and a high tracking-speed rate (0.9 seconds/frame) in the real-time dataset (TUD Campus) but failed to model the motion of an object for an extended period. Then, Tang *et al.* [54] introduced subgraph decomposition to improve the motion model for multiple object tracking through a finite set of hypothesis detections. Their subgraph multi-cut model had the property of jointly addressing the spatial issue (within-frame) and temporal (across-frame) associations. This gave the advantage of using the minimum cost subgraph multi-cut to link and cluster plausible detections jointly across space and time. Although the proposed approach enhanced performance on both tracking speed (0.86 seconds/frame) and tracking accuracy (80.9%) on the public benchmark dataset (TUD campus), it could not efficiently track objects with motion variations and hence incurred high computational costs. Ruchay *et al.* [23] proposed an algorithm to track targets based on local adaptive correlation filters and enabled object tracking with motion variations in high-speed scenes. The adaptive procedure is applied for a typical scene background and multiple composite filters. The impulse responses of optimum correlation filters are used to synthesize composite filters for distortion invariant object tracking. This is employed via a prediction scheme that uses composite correlation filters to track multiple objects with invariance to poses, occlusion, clutter, and illumination variations. However, it has consequently

TABLE 2. Overview of deep learning methods in real-time MOTS with high-speed tracking and low computational cost.

Algorithm	Training Data Size	Testing Data Size	Year	Technique	Description
Bochinski *et al.* [2]	MOT2016: 7 videos, (5316 Frames, (03:58 minutes)	MOT2016: 7 videos, (5919 Frames, (04:13 minutes)	2017	TBD	Use the TBD paradigm to track targets at high speed without using the image information.
Weng and Kitani [22]	KITTI: 6 videos, (1767 Frames, (01:33 minutes)	KITTI: 15 videos, (664 Frames, (01:00 minutes)	2019	CNN	Combine two filters with CNN to reduce the computational costs and system complexity.
Ruchay *et al.* [23]	MOT2015: 4 videos, (1165 Frames, (02:06 minutes))	MOT2015: 3 videos, (868 Frames, (01:13 minutes))	2017	Locally Adaptive Correlation (LAC) Filters	Use LAC filters to track speedy objects
Sharma *et al.* [24]	KITTI: 6 videos, (1767 Frames, (01:33 minutes)	KITTI: 15 videos, (664 Frames, (01:00 minutes)	2018	Pairwise cost	Use bounding-box image pixels and geometric shapes to track speedy objects.
Shin *et al.* [46]	VOT-2015: 22 videos ,(12568 Frames, (02:58 minutes))	VOT-2015: 12 videos, (6557 Frames, (01:36 minutes))	2020	KFC	Incorporate three functional modules onto KFC based on the tracking method to increase processing speed
Tang *et al.* [54]	TUD: 1 video (Campus), (71 Frames, (00:03 minutes))	TUD: 1 video (Crossing), (201 Frames, (00:08 minutes))	2015	Subgraph Decomposition	Use subgraph multi-cut to link and cluster plausible detections jointly across space.
Redmon *et al.* [56]	MOT2015: 4 videos, (2165 Frames, (02:36 minutes))	MOT2015: 3 videos, (868 Frames, (01:13 minutes))	2016	Faster R-CNN+YOLO	Use Faster-CNN and YOLO to track targets at high speed.
Wang *et al.* [59]	MOT2016:7 videos, (5316 Frames, (03:58 minutes)	MOT2016:7 videos, (5919 Frames, (04:13 minutes)	2019	CCN+ Appearance Model	Combine temporal and appearance features to form a unified framework.
Chu *et al.* [55]	MOT2016:7 videos, (5316 Frames, (03:58 minutes)	MOT2016:7 videos, (5910 Frames, (04:13 minutes)	2017	STAM+CNN	Use STAM and CNN to solve computational efficiency.
Keuper *et al.* [57] and Chen and Ren [58]	MOT2016:7 videos, (5316 Frames,(03:58 minutes)	MOT2016:7 videos, (5919 Frames,(04:13 minutes)	2018	Motion segmentation	Combine bottom-up motion segmentation with top-down.
Zamir *et al.* [60]	TUD: 3 videos, (1175 Frames,(02:09 minutes))	TUD: 4 videos, (2175 Frames,(02:39 minutes))	2015	Generalized Minimum Clique Graphs(GMCG)	Use GMCG to find detections that correspond to one particular object in frames.
Kim *et al.* [61]	MOT 2017: 2 videos, (1650 Frames,(01:00 minutes)	MOT2017: 3 videos, (2275 Frames, (01:21 minutes))	2018	DCNN	Use DCNN to extend appearance modeling.
Voigtlaender *et al.* [82]	MOTS/MOT2019: 4 videos,(2862 Frames,(02:13 minutes))	MOTS/MOT2019: 4 videos,(3044 Frames, (02:45 minutes))	2018	Segmentation+ S-CNN+ Faster R-CNN tracker	Use Segmentation and S-CNN to minimize computational cost, and Faster R-CNN to increase tracking speed.

improved the speed tracking and led to high-speed tracking (0.56 seconds/frame) with difficulties in handling illumination variations and prolonged occlusions. It also contributed to a deteriorated tracking performance (53.3%) with moderate computational costs.

To increase tracking accuracy while preserving processing speed, Shin *et al.* [46] incorporated three functional modules, including tracking failure detection, re-tracking using multiple search windows, motion vector analysis, and a preferable search window onto a kernelized filter (KFC)-based tracking method. The technique uses detection failure to analyze the peak and average of neighboring correlation values. It further re-tracks the target using tracking failure and calculates a motion vector of the target by selecting the preferred search window during tracking failure detection. Although the proposed approach registered a high rate of both tracking accuracy (70%) and tracking speed (1.9 seconds/frame), its retracking process required an additional computational load for multiple search windows on a public dataset (Visual Tracker Benchmark). This led to a high rate of motion direction prediction and target losses during the tracking process in crowded scenes.

Sharma *et al.* [24] proposed an approach that minimizes the computational costs by estimating motions on rough frames based on odometry and implemented features on the background. It adopted the tracking-by-detection paradigm and took input from the monocular video frame sequence. It further used the target information beyond the bounding box image pixels to estimate the 3D shape and posing. The approach illustrated pairwise costs disambiguating across track viewpoint variations and relative target movements but suffered from IDS and fragmentations due to motion variations in a real-time MOT evaluation of a public benchmark dataset (KITTI). Therefore, it resulted in moderate computational costs and high tracking speed (1.6 seconds/frame) with satisfactory tracking accuracy (84.2%).

To model object motion, Keuper *et al.* [57] and Chen and Ren [58] proposed a motion segmentation technique that combined bottom-up motion segmentation with top-down multiple object tracking. It grouped point trajectories through clustering of bounding boxes to improve tracking accuracy on small dense objects. It then used a supervised CNN to minimize the computational cost and a Faster R-CNN tracker to obtain detections in a video sequence without knowing the object's category and interest. To enhance the tracking accuracy, the detections from the Faster R-CNN detector were trained in real time using the MOT 2016 public dataset. However, the technique enabled the approach to track objects at high speeds (1.8 seconds/frame), but it experienced moderate computational costs [34], complexity, difficulties in handling motion variation and annotated objects that are caused by over-segmentation in real-time MOT evaluations of public datasets (MOT2016 and MOT2017). This led to a deteriorated overall tracking performance (47.1%).

To extend the task of online MOT on segmentation tracking with the creation of dense pixel-level annotations and semi-automatic annotation procedures. Voigtlaender *et al.* [82] proposed a new baseline method that jointly addressed object detection and segmentation with a single convolutional neural network (SCNN). The approach implements the TrackR-CNN tracker as a baseline to address all aspects of multi-object tracking and segmentation (MOTS) duties in real-time MOT evaluations of public datasets (MOT2016, KITTI, and MOT19). It further extends TrackR-CNN to Mask R-CNN [11] with 3D convolution layers to incorporate temporal information and tracklet associations over time. Then, the TrackR-CNN masked-based detections together with association features are used as input to a tracking algorithm to decide which detections to select and link with bounding boxes. Although this led to a highly satisfactory tracking speed (0.5 seconds/frame), it has also contributed to the algorithm's failure to handle the segmentation of speedy objects. Hence, the overall tracking performance (47.1%) experienced deteriorations. Bochinsk *et al.* [2] suggested a tracking-by-detection paradigm to track targets with high speed without using image information. The approach also incorporated a simple IOU tracker to track targets by associating detections with the highest intersection over union (IOU) to the last detections in the previous frame. It rooted out the short tracks to improve the algorithm's sensitivity towards false positives. This contributed to a high tracking speed (0.4 seconds/frame), but more detections on the tracker have caused many mispredictions of detections in real-time evaluations of the DETRAC and MOT16 datasets. Therefore, the approach resulted in a high number of target losses and achieved an unsatisfactory tracking performance (25.3%) with high computational resources.

Real-time object tracking with speed tracking is a crucial technology for visual analysis, object detection, and motion variation handling [2]. Redmon *et al.* [56] and Ren *et al.* [76] proposed the regional proposal network for object proposals and shared the regional classification through convolutional layers and Fast R-CNN. The technique used Fast R-CNN to produce local proposals with optimized classification and bounding box regression tasks. It enhanced the processing speed by using CNN fully connected layers for region proposals without handcrafted features. However, the approach was complex with noisy detections and suffered from overfitting and false detections in real-time MOT evaluations of public datasets (Picasso and MOT2016). Therefore, it resulted in a high tracking speed (0.01 seconds/frame), a moderate rate of both computational costs and tracking accuracy (57.9%).

Weng and Kitani [22] tried to minimize the computational costs and system complexity for multiple online object tracking. They proposed an approach that combined two filters with CNN to improve data association and object state estimation. The approach incorporated a vast space of the Kalman filter into a full 3D domain to handle 3D location, size, velocity, and object orientation to minimize the computational cost and system complexity [10]. It succeeded with high tracking speed (39.4%) to reduce the computational costs and system complexity but suffered from high false object detections due to the lack of appearance feature extractions in the real-time MOT evaluation of a public dataset (MOT2016). This contributed to a negative overall tracking performance (39.4%). Then, Wang *et al.* [59] combined temporal and appearance features to form a unified framework to reduce the computational cost by grouping tracklets together based on similarities. The approach extended the first architecture of the Siamese network to learn the associating affinities between tracklets. It further combined the appearance model with CCN features and improved the tracking speed (0.5 seconds/frame) and tracking accuracy (56.1%) through clustering and assigning unique individual identities. However, the framework suffered from drifting, object interaction, and occlusions in real-time MOT evaluations of public datasets (MOT2016 and MOT 2017).

To solve computational efficiency, drifting, and occlusions in online multi-object tracking, Chu *et al.* [55] propose an object-specific particle filtering framework for real-time MOT evaluations. The approach tracked each object with

TABLE 3. overview of deep learning methods on modeling target uncertainty in online MOTS.

Algorithm	Training Data Size	Testing Data Size	Year	Technique	Description
Fajardo *et al.* [32]	MOT2015: 4 videos, (1165 Frames, (02:06 minutes))	MOT2015: 3 videos, (868 Frames,(01:13 minutes))	2016	DCNN	Propose deep appearance features method to improve the object data association and affinity in different frames
Wojke *et al.* [62]	MOT2016:7 videos, (5316 Frames,(03:58 minutes)	MOT2016:7 videos, (5919 Frames,(04:13 minutes)	2018	DCNN+ Kalman Filter	Use DCNN and Kalman to find the movement and appearance features of targets.
Gan *et al.* [63]	VOT-2015: 22 videos ,(12568 Frames,(02:58 minutes))	VOT-2015: 12 videos, (6557 Frames,(01:36 minutes))	2018	STAM+CNN	Use STAM and CNN to develop an online MOT method to handle the drift and identity (ID) switch caused by occlusions and integration among targets.
Kampker *et al.* [64]	KITTI:10 videos, (2111 Frames, (01:57 minutes))	KITTI: 6 videos, (1767 Frames,(01:33 minutes)	2018	DCNN	Use DCNN to tackle the high number of object uncertainties in urban areas.
Bewley *et al.* [65]	MOT2016:7 videos, (5316 Frames,(03:58 minutes)	MOT2016:7 videos, (5919 Frames,(04:13 minutes)	2016	Kalman Filter + Hungarian Algorithm	Use Kalman filter features and the Hungarian algorithm to find the association in visual tracks.
Wang *et al.* [66]	MOT2016:7 videos, (5316 Frames,(03:58 minutes)	MOT2016:7 videos, (5919 Frames,(04:13 minutes)	2019	DCNN	Use DCNN to learn target detection and appearance in a shared model.
Zhu *et al.* [41]	MOT2016:7 videos, (5316 Frames,(03:58 minutes)	MOT2016:7 videos, (5919 Frames, (04:13 minutes)	2019	Appearance and Motion Modeling	Combine the Appearance and Motion model+ Template Matching

two constructed CNN-based classifiers. To handle occlusion between objects, it learned spatial attention features based on the visible map using convolution and fully connected layers. Then, the spatial attention map weight features were used to promote the accuracy of the classifier. It reduced time-consuming computation by sharing the CNN feature maps. Then, a single object tracker was incorporated into the spatial-temporal attention mechanism (STAM) procedure and enabled target searching in the next frame. It has enhanced the tracking-speed (0.5 seconds/frame) and handled the interactions very well but could not uniquely differentiate targets that appeared similar in real-time MOT evaluations of public datasets (MOT2015 and MOT 2016). This led to high false alarm and misdetection rates under heavily dense scenes and resulted in unsatisfactory tracking accuracy (46.0%).

C. DEEP LEARNING METHODS ON MODELING TARGET UNCERTAINTY IN AN ONLINE MOT

Online MOT uncertainty is mainly caused by ineffectiveness in associating targets with relevant tracks. This affects the performance of many algorithms in handling object discrimination and direction predetermination processes. In this section, we explain the deep learning methods based on modeling target uncertainty in online MOTs, as shown in Table 3.

Bewley *et al.* [65] used the merits of single object tracking [83] to integrate the Kalman filter features with the Hungarian algorithm to find the association in visual tracks. The approach used CNN-based detection and Faster Region CNN (FR-CNN) in an end-to-end fashion. The FR-CNN shared parameters between two stages to create an efficient framework for detections. However, the approach focused more on efficiency and reliability to handle common frame-to-frame associations than robust detection errors. This led to a failure in handling objects' appearance variations [2], [23], [58] and high IDS (7,318) with a low tracking accuracy (33.4%) under heavy interaction in a real-time MOT evaluation of a public dataset (MOT2016). To increase discrimination, Wojke *et al.* [62] employed the deep feature extraction technique based on a wide residual network (WRN) for the person re-identification process. They normalized the l_2 and 128-dimensional features before the cosine softmax classifier layer [22]. Then, the cosine and emotional Mahalanobis distances are used to fuse dissimilarities. The approach incorporated the Kalman filter to find the movement and appearance features of the target. It further extracted the appearance feature through DCNN and tracked the target individually. Although it has improved tracking performance (61.4%), it struggled to handle target tracking under crowded, distanced views, drifting, and prolonged occlusions. Consequently, this led to high-frequency changes in object IDS rate (12,862) during a tracking process in a real-time MOT evaluation of a public dataset (KITTI).

Inflexible objects have been proven to cause object drifting [29]. Then, Gan *et al.* [63] use the merits of [55] to develop an online MOT approach to handle the drift and identity (ID) switches caused by occlusions and integration among targets. The approach used convolutional layers to extract appearance features [47] and fully connected layers to update a distinguished online target from the background. It further used the interaction of appearance motion with the interaction cues of the target and the online ID assignment scheme based on multi-level features to confirm the trajectory of each target. This technique enhanced the model updates and identity association of the appearance model with STAM and CNN. It also contributed to the approach capabilities of finding appropriate target detections in the previous frame and the effectiveness of linking them to the current frame. However, the similarities, long-term occlusions, and velocity changes [85] on a tracked target mostly led to uncertainties in real-time tracking. Therefore, these factors contributed to the approach's failure to differentiate targets, which led to a high IDS (7,912), fragmentation, and unsatisfactory tracking accuracy (44.0%) on public datasets (VOT and OTB) evaluations. Zhu *et al.* [41] and Liu *et al.* [76] tried to address the problems by combining appearance and motion models. The technique integrated the models onto the Siamese network to learn affinities for tracklets and replace previous features from the IDLA. It further employed the online tracking framework to cascade associations between tracklets and detections in two stages based on target confidence levels (high-to-low). However, it could not track small objects in motion that have similar appearance in real-time MOT evaluations of public datasets (MOT2016 and MOT2017). Therefore, it suffered from bearable IDS (1,871) and challenging tracking accuracy (48.3%). Then, Wang *et al.* [66] tried to learn and track small objects in motion by extending the first architecture of the Siamese network to learn target detection, affinity associations between tracklets, and appearance embedding in a shared model. The approach incorporated the appearance-embedding model into a single-shot detector for simultaneously outputting detections and corresponding embedding. It further used those detections for localization and tracking and then linked tracks onto the appearance model for data associations. It achieved satisfactory tracking accuracy (62.1%) in real-time MOT evaluations of public datasets (MOT2016, MOT2017, and KITTI), but it could not describe the dependencies between tracklets with a similar appearance.

To effectively differentiate objects with a similar appearance Fajardo *et al.* [32] proposed a deep appearance features method to improve the object data association and affinity in different frames that uniquely tracked targets through the CNN framework based on motion and appearance information [80]. The approach struggled to recognize the cropped patches with limited information [86] and hence suffered from false positives and misdetections in real-time MOT evaluations of public datasets (MOT2016 and MOT2017). Therefore, it resulted in moderate frequent object ID changes (4,123) and high tracking accuracy (75.2%). This increased the attention onto tackling the high uncertainty number values in real-time tracking [87]. Kampker *et al.* [64] presented a real-time framework for multi-object detection and maneuver-aware tracking for 3D LIDAR applications to tackle object uncertainty in cluttered urban environments. It combined a sensor occlusion-aware detection method with computationally efficient rule-based filtering and adaptive probabilistic tracking to handle uncertainties arising from the sensing limitation of 3D LIDAR and the complexity of the targets' movement. The technique used algorithm detection as an input 3D point cloud and divided it into non-ground and elevated measurements. This task was accomplished via a slope-based ground removal approach and a subsequent filtering process. It further generated the object hypotheses for the tracking targets in a clustering process. Then, the objects of interest were extracted by means of a subsequent feature-based bounding box fitting and rule-based filtering. However, the technique handled prolonged occlusions and improved the tracking accuracy (86.1%) with a remarkable reduction in IDS (65) in real time MOT evaluations of public datasets (MOT2016 and KITTI).

D. DEEP LEARNING METHODS WITH CNN, AFFINITY, AND DATA ASSOCIATION IN ONLINE MOTS

The traditional CNN architecture uses the handcrafting of cost functions that hinder the tracking performance in most recent works. It is mostly expanded and integrated with deep learning techniques, as illustrated in Table 4, to handle object affinities and data associations. Hence, in this section, we explore the deep learning methods with CNN, affinity, and data association in an online MOTs.

To enhance target tracklet associations, Schulter *et al.* [17] proposed a formula that enabled the learning of arbitrary parameterized cost functions for all variables with association problems and enhanced the MOT in real-time applications. They constructed an end-to-end deep learning min-cost network flow and defined a loss function of the deep architecture as the weighted I_2 a distance of edge labels. The approach further optimized the algorithm by building network flow with its edges on multilayers to form a deep architecture model. It has been able to track and re-identify the objects under complex scenes in real-time tracking. It further handled the long occlusions and accurately estimated the objects' affinity scores. Therefore, this contributed to a good low IDS rate (65) and high rate achievement in both mostly tracked (58.3%) and tracking accuracy (67.4%) real-time MOT evaluations of public datasets (KITTI, MOT2015 and MOT2016). Kumar *et al.* [67] constructed a complementary graph function to capture the spatial-temporal and appearance information. They further constructed an exclusion graph function to ensure that some detections that occurred simultaneously do not share the same node labels. Then, the appearance information is used to link detections into trajectories. However, this contributed to a high tracking accuracy and a low IDS rate (5) achievement in real-time MOT evaluations of public datasets (APIDIS, PETS-2009 S2/L1, MOT2015 (TUD Stadtmitte, and TUD Crossing)) but struggled to associate objects effectively at crossing scenes.

To solve object association ambiguities in cluttered multi-object scenarios, Scheel *et al.* [71] suggested implementing the Monte Carlo algorithm with a multi-Bernoulli filter to handle the association measurements between objects. They extended the algorithm's object filter to work directly on the raw measurements and process multiple measurements per object. Although the approach achieved high tracking accuracy (74.4%) in a real-time MOT evaluation of a public dataset (KITTI), it failed to calculate the association measurements accurately and resulted in filter divergence. Leal-Taixe *et al.* [68] extended the technique into the Siamese

TABLE 4. Overview of deep learning methods with CNN, affinity, and data association in online MOTS.

Algorithm	Training Data Size	Testing Data Size	Year	Technique	Description
Jiang et al. [8]	PETS09: 1 video (S1.L1),(795 Frames, (01:54 minutes))	PETS09:1 video (S2.L2),(436 Frames, (01:02:minutes))	2018	VSN	Use VSN to analyze objects trajectories across multiple cameras and security analysis of images in various scenarios.
Schulter et al. [17]	MOT2016:7 videos, (5316 Frames,(03:58 minutes)	MOT2016:7 videos, (5919 Frames,(04:13 minutes)	2017	DCNN	Use DCNN to propose a formula to enable the learning of arbitrary parameterized cost functions for all variables.
Wu et al. [27]	DukeMTMC: 4 videos, (1743 Frames, (02:59 minutes))	DukeMTMC 5 videos, (1567 Frames, (35:00 minutes))	2017	SCT+CNN	Use the SCT technique to associate detached detections onto tracks.
Kieritz et al. [36]	MOT2015: 1 video (S1.L1), (795 Frames, (01:54 minutes))	MOT2015:1 video (S2.L2) ,(436 Frames, (01:02:minutes))	2016	Integral Channel Features(ICF)+Appearance Model Learning	Propose an online learning appearance model
Le et al. [37]	PETS09: 1 video (S1.L1), (795 Frames, (01:54 minutes))	PETS09:1 video (S2.L2) ,(436 Frames, (01:02:minutes))	2018	MCT+MDP	Use to collaborate object tracking with a camera network.
Son et al. [40]	MOT2016:7 videos, (5316 Frames, (03:58 minutes)	MOT2016:7 videos, (5919 Frames, (04:13 minutes)	2017	Quadruplet CNN (Q-CNN)	Use Q-CNN to learn and associates detections across video frames using appearance and motion cues.
Kumar et al. [67]	MOT2015: 1 video (TUD Campus), (71 Frames, (00:03 minutes))	MOT2015: 1 video (TUD Crossing), (201 Frames,(00:08 minutes))	2017	Spatial-Temporal and Appearance Information	Capture both the spatial-temporal and appearance information and then used appearance information to link detections into trajectories.
Leal-Taixe et al. [68]	VOT-2015: 22 videos ,(12568 Frames, (02:58 minutes))	VOT-2015: 12 videos, (6557 Frames, (01:36 minutes))	2016	Siamese Network	Use the Siamese network to determine the affinity scores.
Huang and Zhou[69]	VOT-2015: 22 videos ,(12568 Frames, (02:58 minutes))	VOT-2015: 12 videos, (6557 Frames, (01:36 minutes))	2017	R-CNN	Use R-CNN to solve the detached detections.
Yoon et al. [70]	VOT-2015: 6 videos, (3234 Frames, (03:42 minutes))	VOT-2015: 5 videos, (1754 Frames, (02:41 minutes))	2018	DCNN	Use DCNN to solve computational bottleneck and affinity issues.
Scheel et al. [71]	KITTI: 6 videos, (1767 Frames, (01:33 minutes)	KITTI: 15 videos, (664 Frames, (01:00 minutes)	2016	MC Algorithm + MB Filter	Propose to implement the Monte Carlo algorithm and multi-Bernoulli filter to handle the association measurements between objects.
Lee et al. [72]	MOT2016:7 videos, (5316 Frames, (03:58 minutes)	MOT2016:7 videos, (5919 Frames, (04:13 minutes)	2016	CNN+ Lucas-Kande Tracker(LKT)	Use CNN based detector and LKT based motion to compute the likelihood of foreground regions as detections response on different object classes.
Chen et al. [73]	MOT2016: 5 videos, (3966 Frames, (02:58 minutes))	MOT2016: 5 videos, (4269 Frames, (03:10 minutes))	2018	F-CNN+Deep Learning	Propose to handle unreliable detections by collecting candidates from outputs of both detection and tracking processes
Tesfaye et al. [88]	DukeMTMC: 4 videos, (1743 Frames, (02:59 minutes))	DukeMTMC 5 videos, (1567 Frames, (35:00 minutes))	2019	Constrained domain set and three hierarchical layers	Merge the detection boxes into consecutive frames and apply fast constrained domain sets (FCDS) in the layers

network to learn matching features for MOTs and then determined the affinity score. Their approach used three types of Siamese CNN topologies for computational cost, information distribution, and the streaming of the data to form inputs for CNN layers. It compares these three topologies and uses the third architecture to extract the in-depth features. It further used the deep features and motion information with a gradient boosting algorithm to formulate the tracking as a linear programming problem and solved it efficiently. However, it struggled to detect and associate the object tracks under a dense population. This contributed to the unsatisfactory overall performance with tracking accuracy (29.0%) and moderate achievements in both the mostly tracked (48.4%) and IDS rates (639) in a real-time MOT evaluation of a public dataset (MOT2015).

To overcome the detection association problem, Son *et al.* [40] proposed using quadruplet CNN to learn and associate detections across video frames using appearance and motion cues. They extended the Siamese network using quadruplets of image patches as inputs and extracted these patches for three detections from one same object and another different object. The approach further constructed a loss function that temporally learned the smooth appearance embedded with the motion-aware position for metric learning. However, the proposed approach could not associate detections very well under crowded scenes and resulted in high IDS (745). This further resulted in a low mostly tracked rate (14.6%) and tracking accuracy (44.1%) in real-time MOT evaluations of public datasets (2DMOT2015 and MOT2016). Then, Lee *et al.* [72] used a CNN-based detector and Lucas-Kande Tracker (LKT)-based motion to compute the likelihood of foreground regions as the detection response of different object classes. The technique separates the dynamic motion model of a Bayesian filter into entity translations and motion cues. Although this contributed to a better tracking accuracy (62.4) and moderate rate for mostly tracked (31.5%) objects in a real-time MOT evaluation of a public dataset (MOT 2015), it left the proposed approach struggling to associate the tracklets over a long tracking period in the heavy interactive scenes and resulted in high IDS (1,394).

Kieritz *et al.* [36] capitalized on the established embedded target appearance process and proposed an online learning appearance model. Their technique combines the appearance model with a simple motion model to estimate the change in position and smooth the trajectory. It used a classifier based on integral channel features to detect persons in each frame. It further used the detector that uses LUV color channels, a histogram of oriented gradients with several bins, and the gradient magnitude to formulate fast detection over every channel. However, the approach experienced deceptive appearances over long track periods and switching between active and inactive states of the trajectories with a low number of associated detections. This hindered an overall performance and resulted in a challenging tracking accuracy (27.1%), a low mostly tracked rate (6.4%), and a moderate IDS rate (1,490) in a real-time MOT evaluation of a public dataset (MOT2015).

Vo *et al.* [86] implemented a multi-sensor generalization labeled multi-Bernoulli (GLMB) filter with two sensors to reduce uncertainty about object existence and state. Zou *et al.* [89] used the established platform to update appearance information based on template matching rather than the learning-based approach. However, these approaches experienced frequent occlusions under heavy interactions and failed to handle appearance variations. This resulted in detached detections in real-time MOT evaluations. Huang and Zhou [69] proposed an online multi-object tracking approach and used a recurrent convolutional neural network (R-CNN) to solve detached detections. To address the data association problem in the paradigm, the technique discarded all the unused data in the video sequence. It reduced the data to a few single measurements per frame and ran the detector. Then, tracklets are associated with each measurement of a corresponding target. This led to considerable target loss due to misdetection and tracks associations in crowded scenes. It also contributed to a high IDS rate and a very low tracking performance in a real-time MOT evaluation of a public dataset (MOT2015).

Wu *et al.* [27] applied a single-camera tracking (SCT) technique to associate detached detections into tracks. They further used the tracks on multi-camera tracking (MCT) to re-identify each track to form trajectories. However, their MCT base technique could not associate the tracks of different cameras and resulted in illumination changes, view angle variation, and object appearance inconsistency. Though this contributed to a satisfactory mostly tracked (51.8%) result, it could not efficiently associate detections and tracks across cameras, hence resulting in poor tracking accuracy (9.65) in a real-time MOT evaluation of a public dataset (MOT2016). To strengthen the target data association across cameras, Le *et al.* [37] proposed the use of Markov decision processing (MDP) to collaborate object tracking with the camera network. The approach extended the MDP to a multiple views framework. Then introduced a novel target association method across cameras. It further collected and associated the tracking outcomes on each camera onto target tracks. This contributed to the effectiveness in handling the appearance similarities [90] under crowded scenes. However, it also led to a low IDS (240) with better overall performance in terms of mostly tracked (62.0%) objects and tracking accuracy (69.8%) in real-time MOT evaluations of public datasets (PETS09-(S1 L1 and S2 L2)).

Houssineau *et al*[9] proposed a new online scheme for evaluating ReID algorithms for object tracking aiming to improve the target ReID process within a camera at different times. The approach considered several issues, such as the open set, dynamic, small gallery set, and multiple camera configurations. However, it could not efficiently capture the scenarios of online tracking for camera networks, open sets, and the dynamic nature of the gallery set due to its limitation on considering only camera scenarios. Then, Tesfaye *et al.* [88] used a constrained domain set and three hierarchical layers to enhance tracking of individual object appearances in each camera. The approach splits the video into small segmentations and generates tracklets. It then merged the detection boxes into consecutive frames [92] and applied fast constrained domain sets (FCDS) in the first layer. In the second, it merged the tracklets into a routine with FCDS in each camera-across. It finally organized all tracks together in the third layer and built a graph of tracklet matching across cameras to verify whether a person appears in one or more cameras across. However, the proposed approach achieved better performance in tracking accuracy (56.6%) and recorded high performance of mostly tracked objects in a

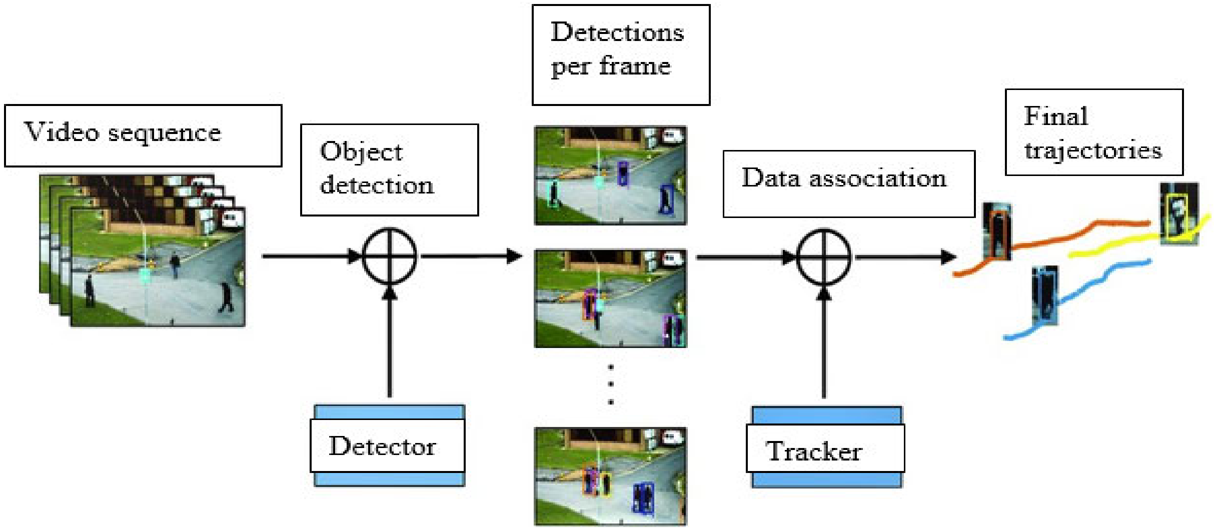


FIGURE 5. Object Tracking based on Quality Detections [76].

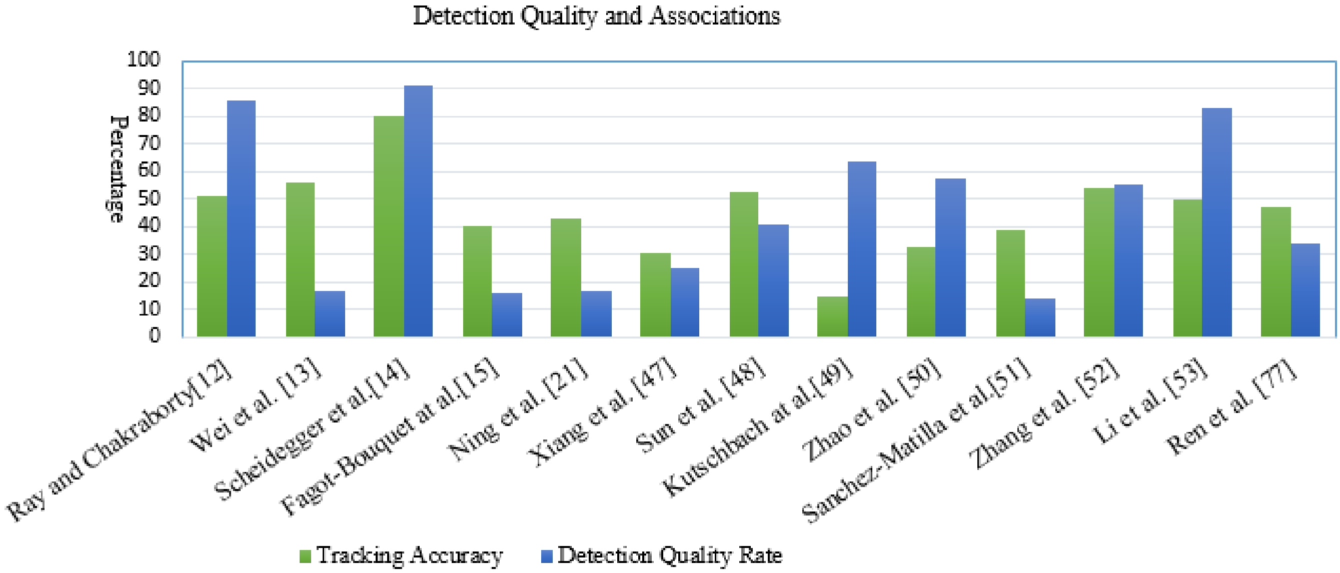


FIGURE 6. Analysis of Deep Learning Algorithms based on Detection Quality and Associations in real-time MOTs.

real-time MOT evaluation of a public dataset (MOT2015), but failed to handle the similarity appearances, could not associate tracklets across cameras, and resulted in most frequent changes in IDS (1,637) under crowded scenes. It also suffered from re-tracking and ReID due to fragmentation.

In ensuring efficient target tracking and data associations in camera networks, Sharma *et al.* [93] used a camera selection policy to select the candidate camera where the target is likely to appear by fording the ReID queries during the target transition. However, the approach brought affinities and computational complexities. Yoon *et al.* [70] designed an appearance matching network for robust online multiple object tracking to solve the computational bottleneck and affinity issues. The proposed network utilized the structural constraint information to represent the relative information portions and velocity differences between objects and track missed objects under heavy collusion with the aid of the ReID process. Then, Ristani and Tomasi [4] suggested reducing the computational complexity by incorporating standard hierarchical reasoning and sliding temporal techniques onto a tracker. These approaches reduced the IDS rate, but they could track the objects for an extended period. This resulted in poor tracking accuracy in real-time MOT evaluations of public datasets (MOT2015 and MOT2016).

Jiang *et al.* [8] analyzed object trajectories across multiple cameras to allow synthesis data and security analysis of images in various scenarios. Their approach used a multi-camera system without turning parameters from the ground truth and constructed a graph from 2D observations of all camera pairs with no network configuration.

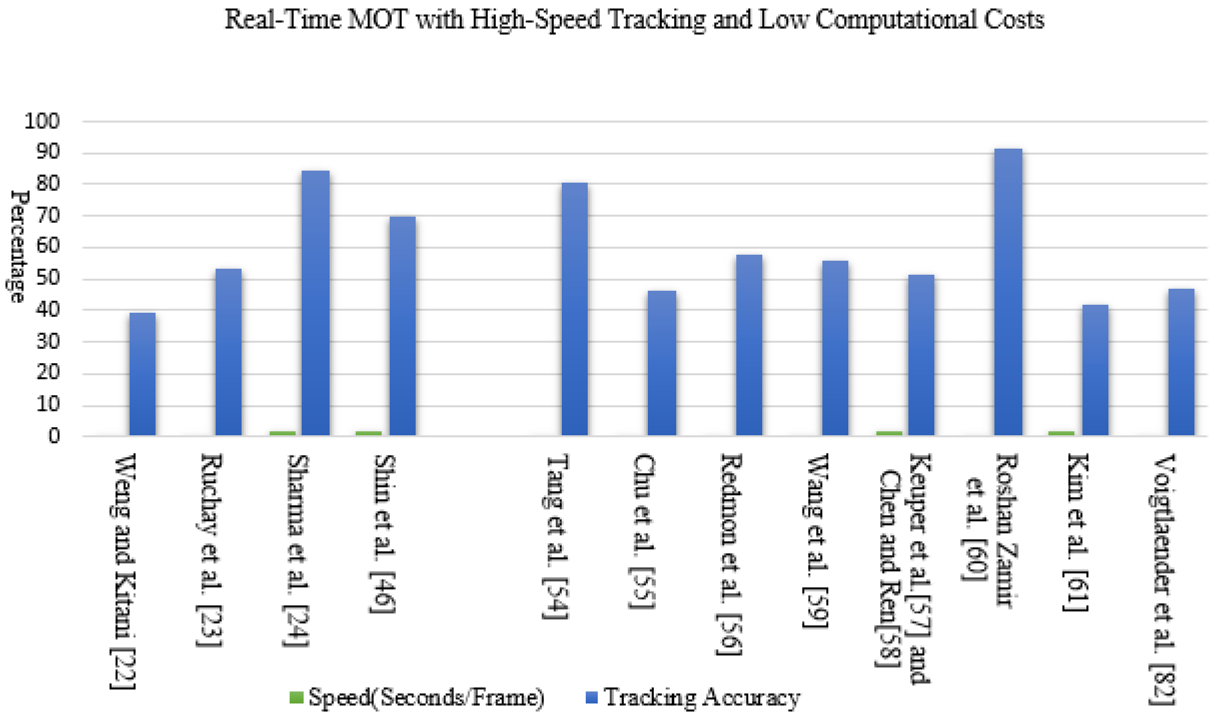


FIGURE 7. **Analysis of Deep Learning Algorithm-based Real-time MOTs with High-Speed Tracking and Low computational Costs.**

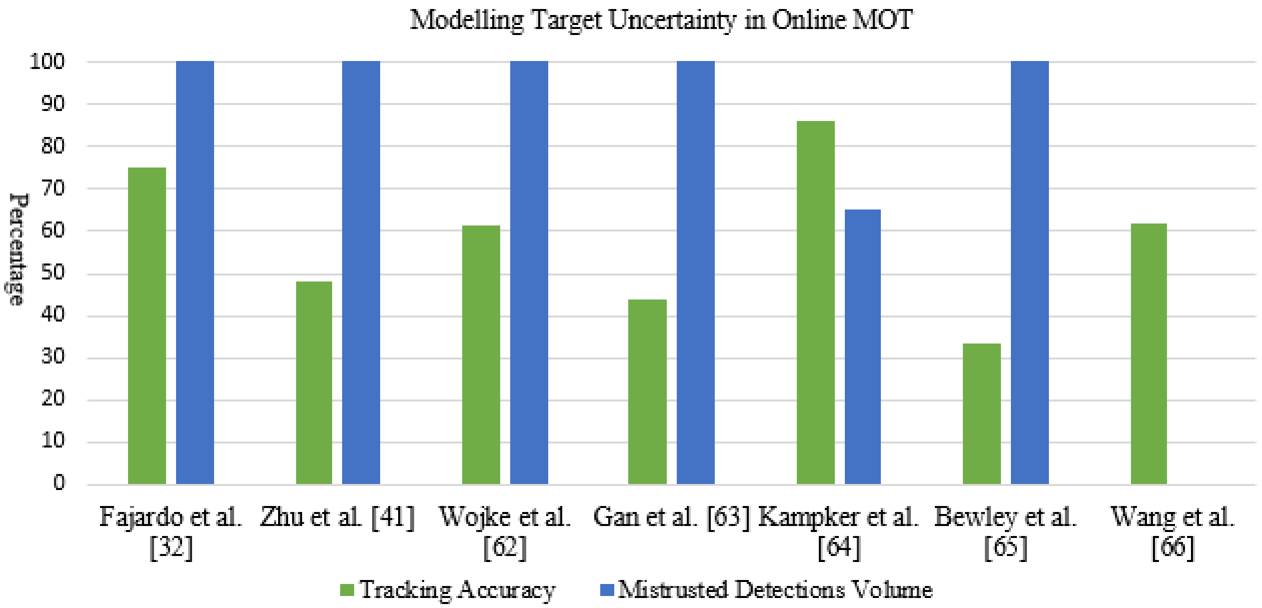


FIGURE 8. **Analysis of Deep Learning Algorithms based on Modeling Target Uncertainty in MOTs.**

The proposed approaches could not efficiently associate the tracklets, especially when they had the same motion and similarities in size and appearance. This resulted in unsatisfactory tracking performance in real-time MOT evaluations of public datasets (PETS09-(S1 L1 and S2 L2)). Then, Chen *et al.* [73] proposed handling unreliable detections by collecting candidates from outputs of both detection and tracking processes. Their approach presented a novel scoring function based on a fully convolutional neural network that shares most computations on the entire image. It further adopted a deeply learned appearance representation to improve the identification ability of a tracker. It also presented a hierarchical data association strategy that utilizes the spatial information and deeply learned person re-identification features to compare tracked objects with their historical features to decide whether the same target was previously identified. However, the

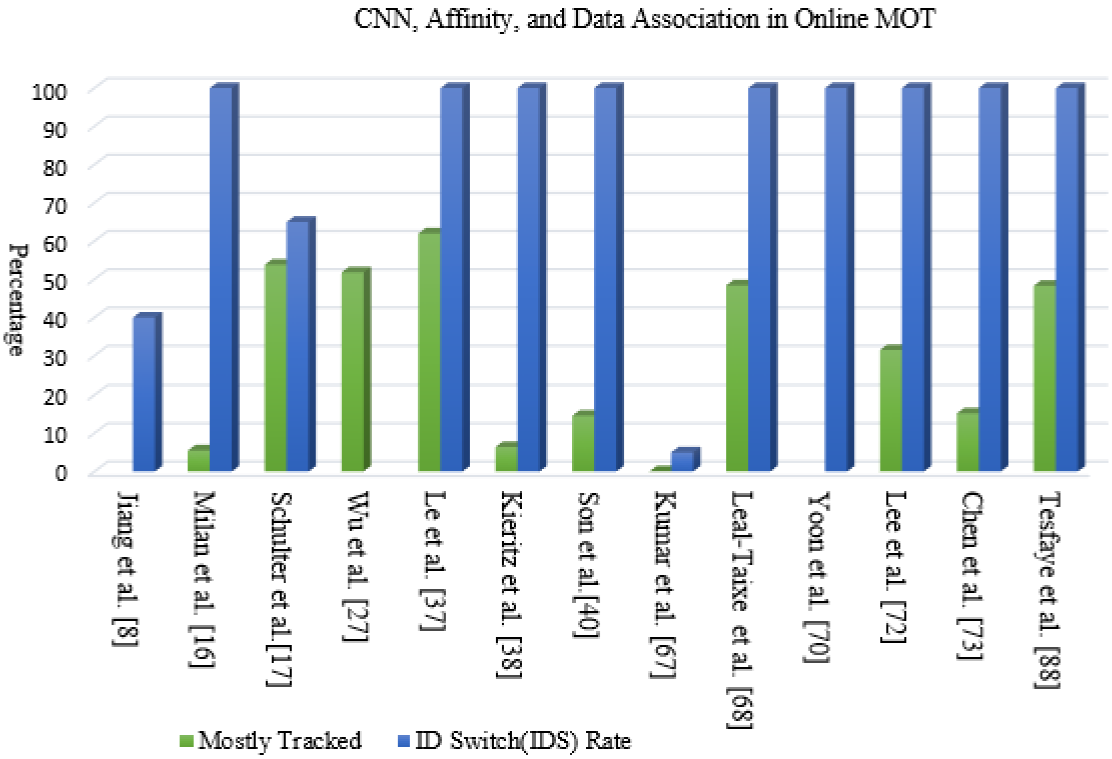


FIGURE 9. Performance on handling the affinity and associations with CNN integrated DLMs.

proposed approach tracker generated more fragmentations when the target suddenly speeds up and affected the target data association. This contributed to a degraded overall performance with vast target losses that led to the unsatisfactory rate for mostly tracked (15.2%) objects and tracking accuracy (47.6%) in a real-time MOT evaluation of a public dataset (MOT2016).

IV. DISCUSSION

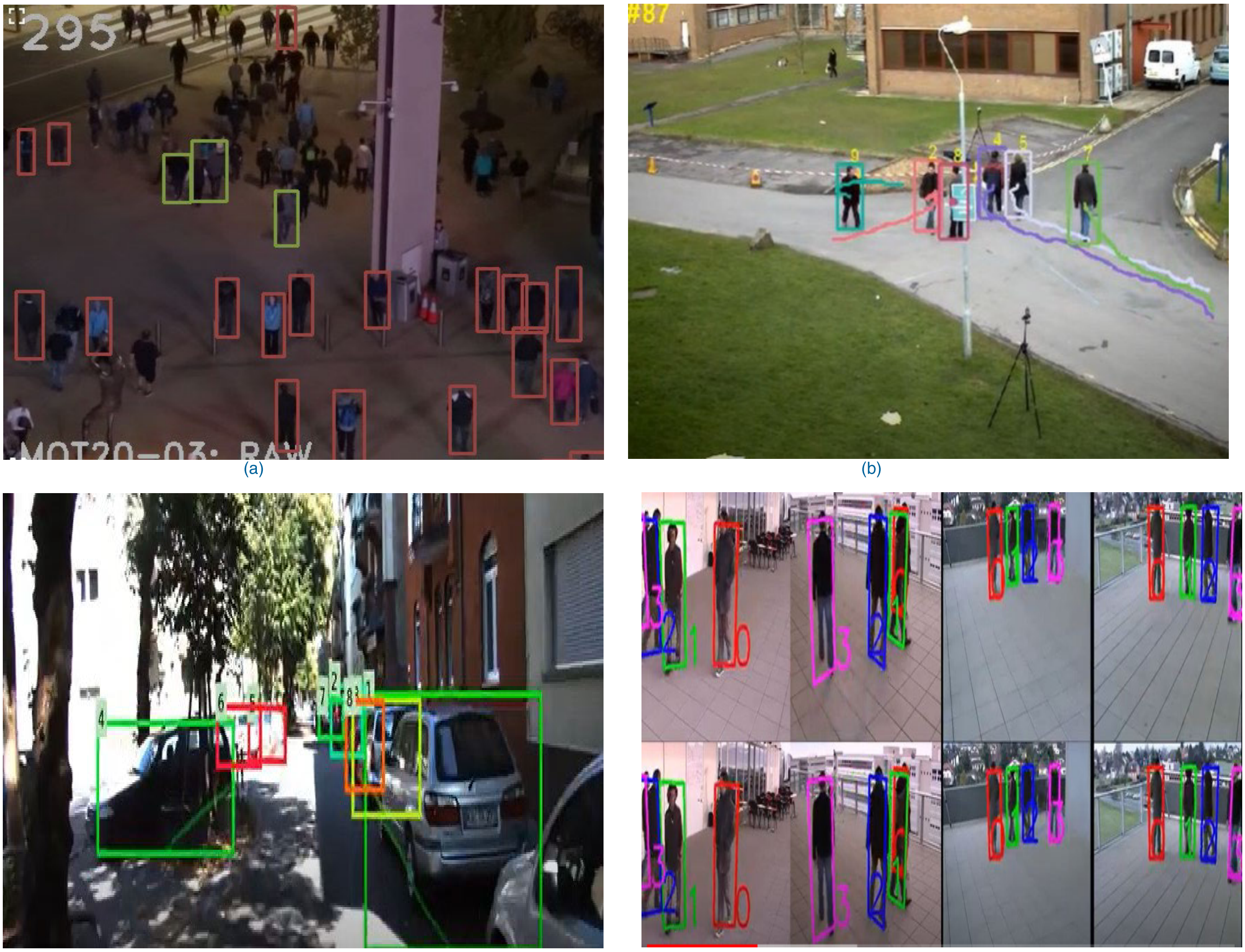
In this systematic review, we provide an overview of the different DLMs for online MOTs in various environments, scenes, and datasets. Based on previous studies, we categorized the proposed approaches into four themes (1. Online MOTs based on detections quality and associations, 2. Real-time MOTs with high-speed tracking and low computational cost, 3. Modeling target uncertainty in online MOTs, and 4. CNN, affinity, and data association in online MOTs).

In the methodology, the real-time MOTs based on deep learning techniques were less frequently accessible. For the performance evaluation, the main challenges were the availability and quality of the evaluation results to include all parameters as per the new standardized MOT evaluation benchmark suggested by [26].

The tracking accuracies of deep learning techniques varied between 14.5% and 86% in video processing under several complex real-world problems [2]. The lowest tracking accuracy of 14.5% was achieved where the distance between the object detected in the first frame and its detection in the next frame increased [49]. This is identified on target detections and tracklets associations in video frames that have complexities in detecting objects with motion variations [37]. It increased the number of problems encountered due to weak data associations and appearance similarities.

DLMs such as [12]–[16], [21], [47]–[53], [77] were combined to construct features and to learn appearance similarities between objects to improve detection quality and associations. In contrast [2], [22]–[24], [46], [54]–[61], [82], classified and assessed the matches between detection and tracklets to quickly track the targets. However, the density of uncertainty in the real-time MOTs persisted. Then [32], [41], [62]–[66], emerged to learn the control of the problem through graphic models and flow optimizations but experienced difficulties in handling the affinities. They used a data-driven mechanism by [8], [17], [27], [36]–[38], [40], [67]–[73], [89], [90] to learn the affinity models for data association and replaced the handcrafted features with a real-time MOT framework, such as Siamese CNN. This contributed to the remarkable progress seen in deep learning techniques based on online MOTs and reduced the numbers of most lost (ML) targets by enforcing a strong foreground and motion differentiation for all moving objects. The performance outcomes of the previous studies have shown that the deep learning approach is the most popular paradigm applied by most researchers in real-time MOTs.

In Table 5, we developed general comparisons and summarized the limitations and strengths of approaches based on each theme. For example, in online MOTs based on detection quality and association, detection and trajectory



(a)

(b)

(c)

(d)

FIGURE 10. Illustration of the detections and feature learning with DLM-based online MOTs. (a). Insufficient detections (poor quality detections) in MOT2020 enter/exit stadium; (b) sufficient detection in PETS09-S211 sequence; (c)Real-time high-speed tracking with a mono-camera in KITTI and (d) multi-person tracking results, color features in deep learning.

construction are based on both current frame information and the previous frame. The proposed deep learning approaches have successfully located detections that correspond to one particular object in different frame sequences. They also proved that low-level detectors are the main factors that thwart the capability of a tracker to estimate the state of the target from ambiguous observations. This is supported by the results of a considerable number of detached detections that could not be associated and hence degraded detections' quality response [2], [22], [23], [54]–[56], [59], [60]. However, background subtraction and feedback detections in each frame could be utilized to improve the tracking performance in deep learning techniques that follow the TBD paradigm. The strategy could also be extended to other DLMs and make them more effective when nearby objects with similar appearances occlude each other in video frames.

Second, real-time MOTs with high-speed and low computational costs and common restrictions, such as making assumptions on surfaces, objects' speed, and directions, led to blurred image capture when objects suddenly sped up. This slowed down the detection rate and led to poor detections that degraded the overall tracking performance. However, the implementation of DCCN, faster R-CCN, filters, and segmentation with DLMs enabled high-speed object tracking with low computational costs, but the methods experienced difficulties in associating the tracks more, especially when the estimated uncertainty was at a low level (lower pixel).

Third, in modeling target uncertainty in online MOTs, the proposed methods had difficulty capturing objects from a mono-camera for online MOTs. This has blurred images and increased uncertainty issues. Then, the independent self-motion that emerged with GLMB reduced the uncertainty about the object's existence state but could not disambiguate between objects and uncertainty due to insufficient data association.

TABLE 5. Strengths and limitations of various deep learning methods based on online MOTS.

Themes	Advantages	Disadvantages
Online MOTs based on detections quality and associations	- Provide constant output - Able to breakdown and link complex and large data - Capable of associating object detected in a previous frame with its new detections in the current frame and smooth segmentation map	- Complex training - Heavily depends on a template - Foible detections and association for small objects - Lack of object differentiation in crowded scenes - Too slow in handling detections for objects with motion variations
Real-time MOTs with high-speed tracking and low computational cost	- Require few parameters - High detections and tracking accuracy - Highly efficient in providing constant output	- Do not perform very well in highly complex scenes - Ignore the correlation among neighboring pixels - Struggle to handle illumination and motion variations - Require high computing resources
Modeling target uncertainty in online MOTs	- Moderate detections and tracking accuracy - Low adaptability with complex data	- High mistrust and detached detections - Insufficient data association (struggle to handle objects' appearance variations and ReID) - Complex system - Requires a large volume of memory and CPU
CNN, affinity, and data association in online MOTs	- Efficient and robust for achieving consistent output - Able combined modules (appearance and motion) for handling appearance and motion variations - Able to detect and localize the object based on data association history	- Require extensive samples - Require additional time for training and high computing process

Last, in the context of affinity and data association, different techniques are applied to enhance CNN for object association. Their affinity computation in multiple frame sequences could not distinguish objects with similar appearances or pedestrians wearing the same attires. This contributed to the skipping of the detections for small objects distantly captured in the images to be suppressed. It also caused difficulties in target tracking and data association issues across multiple cameras [94], whereby each camera scene needed to be merged in with those of the different cameras on the network.

DLMs had trouble learning the incoming tracks and differentiating various detections, as shown in Figs. 6, 9, and 10(b)-(d), but they signified promising progress towards real-time MOT systems in handling the object observation formulation, affinity, and data association problem. They used the detectors to enable the pass over of the generated detections onto the trackers as the input for data associations. This compensated for the missing detections shown in Fig. 10(a) but struggled with weak detections and a lack of tracklet association for small objects. This caused a high volume of mistrust and detached detections that degraded tracking accuracy, as depicted in Fig. 9. The approaches struggled to handle objects' appearance variations and ReID; they could also not learn the appearance features very well under crowded scenes and motion variations.

However, the proposed approaches could differentiate various detections and effectively learned the incoming tracks, as shown in Figs. 6 and 10(b)-(d). The strategy improved the quality of detection and tracking performance. The approaches with low-quality detections could not maintain the object appearance features and yielded a poor tracking performance. The advocacy is well presented in Fig. 7, where high-speed tracking methods tend to skip the objects traveling slowly and struggle to handle objects' appearance variations and motion variations. The detection accuracy inconsistency caused a failure for the approaches to make decisions on which targets are true incomers or leavers. This affected the overall tracking accuracy rate compared to that with low-speed tracking.

The complexity and uncertainty in these approaches can be seen in Figs. 8 and 10(a), where the multi-camera view calculations of the motion trajectories from entering and exiting the views lie in multifarious aspects. The mistrusted detections crowd the tracking accuracy performance.

Although there is noticeable progress, it is important to note that this review reports only on deep learning techniques based on real-time MOTs, as categorized in Fig. 4.

The DLMs have not been implemented thoroughly to solve real-world problems. Thus, many challenges persist, and more studies need to be conducted to include a broadened scope on vehicles and top-view multiple object tracking using drones

For further research, it would be advisable to compare the traditional CNN with DCNN techniques based on experimental evaluations. It would also be important to include other techniques, such as deep convolutional generative adversarial networks (DCGANs), for a robust algorithm to handle the challenges that have been reported under complex environments.

V. CONCLUSION

This review paper analyzes and summarizes the latest progress and challenges in real-time MOTs. We analyzed several papers on deep learning techniques used in real-time multiple object tracking. We further described and discussed the best results for the four main themes: online MOTs based on detections quality and associations, real-time MOTs with high-speed tracking and low computational cost, modeling target uncertainty in online MOTs, and CNN, affinity, and data association. For each theme, several papers are considered to illustrate the main challenges of the most popular solutions proposed by the authors.

Until now, there has been no review of the various recent DMLs for online MOTs. Deep learning strategies are already widely used in real-time MOTs. Our analysis shows that DLMs improve the handling of multiple object detections and trajectory associations across sequential frames under challenging environments. The results could be used for further improvement of the solutions' efficiency and robustness on surveillance security management systems. They can also be used for further studies in real-time MOT algorithms to promote the sustainable development goal (SDG) 16 by contributing to adequate and timely decision-making by committees and justice institutions that protect and save lives in smart cities.

REFERENCES

[1] S. Moon, J. Lee, D. Nam, H. Kim, and W. Kim, "A comparative study on multi-object tracking methods for sports events," in *Proc. 19th Int. Conf. Adv. Commun. Technol. (ICACT)*, Bongpyeong, South Korea, 2017, pp. 883–885.

[2] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Lecce, Italy, Aug. 2017, pp. 1–6.

[3] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-track: Efficient pose estimation in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 350–359.

[4] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6036–6046.

[5] M. Tiwari and R. Singhai, "A review of detection and tracking of object from image and video sequences," *Int. J. Comput. Intell. Res.*, vol. 13, no. 5, pp. 745–765, Mar. 2017, doi: 10.1109/cis.2009.13.

[6] D. M. Patel, U. K. Jaliya, and H. D. Vasava, "Multiple object detection and tracking: A survey," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 6, no. 2, pp. 809–813, Apr. 2018.

[7] K. Soomro, H. Idrees, and M. Shah, "Predicting the where and what of actors and actions through online action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2648–2657.

[8] X. Jiang, Z. Fang, N. N. Xiong, Y. Gao, B. Huang, J. Zhang, L. Yu, and P. Harrington, "Data fusion-based multi-object tracking for unconstrained visual sensor networks," *IEEE Access*, vol. 6, pp. 13716–13728, Apr. 2018, doi: 10.1109/access.2018.2812794.

[9] R. Martín-Nieto, Á. García-Martín, J. M. Martínez, and J. C. Sanmiguel, "Enhancing multi-camera people detection by online automatic parametrization using detection transfer and self-correlation maximization," *Sensors*, vol. 18, no. 12, p. 4385, Jul. 2018, doi: 10.3390/s18124385.

[10] S. Tian, F. Yuan, and G.-S. Xia, "Multi-object tracking with inter-feedback between detection and tracking," *Neurocomputing*, vol. 171, pp. 768–780, Jan. 2016, doi: 10.1016/j.neucom.2015.07.028.

[11] M. Z. Islam, M. S. Islam, and M. S. Rana, "Problem analysis of multiple object tracking system: A critical review," *IJARCCE*, vol. 4, no. 11, pp. 374–377, Nov. 2015, doi: 10.17148/ijarcce.2015.41183.

[12] K. S. Ray and S. Chakraborty, "An efficient approach for object detection and tracking of objects in a video with variable background," 2017, *arXiv:1706.02672*. [Online]. Available: http://arxiv.org/abs/1706.02672

[13] J. Wei, M. Yang, and F. Liu, "Learning spatio-temporal information for multi-object tracking," *IEEE Access*, vol. 5, pp. 3869–3877, Jan. 2017, doi: 10.1109/access.2017.2686482.

[14] S. Scheidegger, J. Benjaminsson, E. Rosenberg, A. Krishnan, and K. Granstrom, "Mono-camera 3D multi-object tracking using deep learning detections and PMBM filtering," in *Proc. IEEE Intell. Vehicles Symp.*, Changshu, China, Jun. 2018, pp. 433–440.

[15] L. Fagot-Bouquet, R. Audigier, Y. Dhome, and F. Lerasle, *Improving Multi-Frame Data Association With Sparse Representations for Robust Near-Online Multi-Object Tracking* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), B. M. Leibe, J, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 774–790.

[16] A. Milan, L. Leal-Taixe, K. Schindler, and I. Reid, "Joint tracking and segmentation of multiple targets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5397–5406.

[17] S. Schulter, P. Vernaza, W. Choi, and M. Chandraker, "Deep network flow for multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2730–2739.

[18] D. Wang, W. Fang, W. Chen, T. Sun, and T. Chen, "Model update strategies about object tracking: A state of the art review," *Electron.*, vol. 8, no. 11, pp. 1–31, Oct. 2019, doi: 10.3390/electronics8111207.

[19] M. Fiaz, A. Mahmood, and S. Ki Jung, "Tracking noisy targets: A review of recent object tracking approaches," 2018, *arXiv:1802.03098*. [Online]. Available: http://arxiv.org/abs/1802.03098

[20] Z. He, J. Li, D. Liu, H. He, and D. Barber, "Tracking by animation: Unsupervised learning of multi-object attentive trackers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 1318–1327.

[21] G. Ning, Z. Zhang, C. Huang, X. Ren, H. Wang, C. Cai, and Z. He, "Spatially supervised recurrent convolutional neural networks for visual object tracking," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Baltimore, MD, USA, May 2017, pp. 1–4.

[22] X. Weng and K. Kitani, "A baseline for 3D multi-object tracking," 2019, *arXiv:1907.03961*. [Online]. Available: http://arxiv.org/abs/1907.03961

[23] A. N. Ruchay, V. I. Kober, and I. E. Chernoskulov, "Real-time tracking of multiple objects with locally adaptive correlation filters," in *Proc. Image Process., Geoinformation Technol. Inf. Secur.*, Samara Oblast, Russia, 2017, pp. 214–218.

[24] S. Sharma, J. A. Ansari, J. K. Murthy, and K. M. Krishna, "Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Brisbane, QLD, Australia, May 2018, pp. 3508–3515.

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

[26] L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu, "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *Comput. Vis. Image Understand.*, vol. 193, Apr. 2020, Art. no. 102907, doi: 10.1016/j.cviu.2020.102907.

[27] C.-W. Wu, M.-T. Zhong, Y. Tsao, S.-W. Yang, Y.-K. Chen, and S.-Y. Chien, "Track-clustering error evaluation for track-based multi-camera tracking system employing human re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1416–1424.

[28] V. P. Bhuvana, M. Schranz, C. S. Regazzoni, B. Rinner, A. M. Tonello, and M. Huemer, "Multi-camera object tracking using surprisal observations in visual sensor networks," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 50, Apr. 2016, doi: 10.1186/s13634-016-0347-x.

[29] S. Bei, Z. Zhen, L. Wusheng, D. Liebo, and L. Qin, "Visual object tracking challenges revisited: VOT vs. OTB," *PLoS ONE*, vol. 13, no. 9, Sep. 2018, Art. no. e0203188, doi: 10.1371/journal.pone.0203188.

[30] S. Lee and H. Hong, "Use of gradient-based shadow detection for estimating environmental illumination distribution," *Appl. Sci.*, vol. 8, no. 11, pp. 1–13, Nov. 2018, doi: 10.3390/app8112255.

[31] A. K. M. Azad and M. Misbahuddin, "Web-based object tracking using collaborated camera network," *Adv. Internet Things*, vol. 8, no. 2, pp. 13–25, 2018, doi: 10.4236/ait.2018.82002.

[32] S. Fajardo, F. R. García-Galvan, V. Barranco, J. C. Galvan, and S. F. Batlle, "Multi-person tracking based on faster R-CNN and deep appearance features," *Vis. Object Tracking Deep Neural Netw.*, vol. 1, p. 13, Dec. 2016, doi: 10.5772/intechopen.85215.

[33] J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon, "Bayesian multi-object tracking using motion context from multiple objects," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 33–40, doi: 10.1109/WACV.2015.12.

[34] S. Li, G. Battistelli, L. Chisci, W. Yi, B. Wang, and L. Kong, "Computationally efficient multi-agent multi-object tracking with labeled random finite sets," *IEEE Trans. Signal Process.*, vol. 67, no. 1, pp. 260–275, Jan. 2019, doi: 10.1109/tsp.2018.2880704.

[35] I. A. Iswanto and B. Li, "Visual object tracking based on mean-shift and particle-Kalman filter," *Procedia Comput. Sci.*, vol. 116, pp. 587–595, 2017, doi: 10.1016/j.procs.2017.10.010.

[36] H. Kieritz, S. Becker, W. Hubner, and M. Arens, "Online multi-person tracking using integral channel features," in *Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Colorado Springs, CO, USA, Aug. 2016, pp. 122–130.

[37] Q. C. Le, D. Conte, and M. Hidane, "Online multiple view tracking: Targets association across cameras," in *Proc. 6th Workshop Activity Monitor. Multiple Distrib. Sens. (AMMDS)*, Newcastle, U.K., Jul. 2018, pp. 1–12, Paper ffhal-01880374f.

[38] J. Ju, D. Kim, B. Ku, D. K. Han, and H. Ko, "Online multi-object tracking with efficient track drift and fragmentation handling," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 34, no. 2, p. 280, Jan. 2017, doi: 10.1364/josaa.34.000280.

[39] J. Wang, X. Zeng, W. Luo, and W. An, "The application of neural network in multiple object tracking," in *Proc. Int. Conf. Comput. Sci. Softw. Eng. (CSSE)*, Jul. 2018, pp. 258–264, doi: 10.12783/dtcse/csse2018/24504.

[40] J. Son, M. Baek, M. Cho, and B. Han, "Multi-object tracking with quadruplet convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3786–3795.

[41] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M. H. Yang, *Online Multi-Object Tracking With Dual Matching Attention Networks* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 379–396.

[42] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Brit. Med. J.*, vol. 339, no. 7716, pp. 332–336, Nov. 2009, doi: 10.1136/bmj.b2535.

[43] A. Osep, A. Hermans, F. Engelmann, D. Klostermann, M. Mathias, and B. Leibe, "Multi-scale object candidates for generic object tracking in street scenes," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Stockholm, Sweden, May 2016, pp. 3180–3187.

[44] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic, "On pairwise costs for network flow multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 5537–5545.

[45] A. Bai and R. Simmons, "Multi-object tracking and identification via particle filtering over sets," in *Proc. 20th Int. Conf. Inf. Fusion (FUSION)*, Xia'n, China, Mar. 2017, pp. 10–13.

[46] J. Shin, H. Kim, D. Kim, and J. Paik, "Fast and robust object tracking using tracking failure detection in kernelized correlation filter," *Appl. Sci.*, vol. 10, no. 2, p. 713, Jan. 2020, doi: 10.3390/app10020713.

[47] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4705–4713.

[48] S. Sun, N. Akhtar, H. Song, A. S. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 9, p. 1, May 2019, doi: 10.1109/tpami.2019.2929520.

[49] T. Kutschbach, E. Bochinski, V. Eiselein, and T. Sikora, "Sequential sensor fusion combining probability hypothesis density and kernelized correlation filters for multi-object tracking in video data," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Lecce, Italy, Aug. 2017, pp. 1–5.

[50] D. Zhao, H. Fu, L. Xiao, T. Wu, and B. Dai, "Multi-object tracking with correlation filter for autonomous vehicle," *Sensors*, vol. 18, no. 7, pp. 1–17, Mar. 2018, doi: 10.3390/s18072004.

[51] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, "Online multi-target tracking with strong and weak detections," in *Proc. Comput. Vis. ECCV Workshops*, in Lecture Notes in Computer Science, G. Hua and H. Jégou, Eds. Cham, Switzerland: Springer, 2016, pp. 84–99.

[52] Z. Zhang, J. Wu, X. Zhang, and C. Zhang, "Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on DukeMTMC project," 2017, *arXiv:1712.09531*. [Online]. Available: http://arxiv.org/abs/1712.09531

[53] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," 2018, *arXiv:1812.11703*. [Online]. Available: http://arxiv.org/abs/1812.11703

[54] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Subgraph decomposition for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 5033–5041.

[55] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4846–4855.

[56] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[57] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele, "Motion segmentation & multiple object tracking by correlation co-clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 1–13, Oct. 2018, doi: 10.1109/TPAMI.2018.2876253.

[58] L. Chen and M. Ren, "Multi-appearance segmentation and extended 0-1 programming for dense small object tracking," *PLoS ONE*, vol. 13, no. 10, pp. 1–14, Aug. 2018, doi: 10.1371/journal.pone.0206168.

[59] G. Wang, Y. Wang, H. Zhang, R. Gu, and J. N. Hwang, "Exploit the connectivity: Multi-object tracking with TrackletNet," in *Proc. 27th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, Oct. 2019, pp. 482–490.

[60] A. R. Zamir, A. Dehghan, and M. Shah, *GMCP-Tracker: Global Multi-Object Tracking Using Generalized Minimum Clique Graphs* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), A. Fitzgibbon, S. Lazebnik, P. S. Perona, and Y. C. Schmid, Eds. Berlin, Germany: Springer, 2012, pp. 343–356.

[61] C. Kim, F. Li, and J. M. Rehg, *Multi-Object Tracking With Neural Gating Using Bilinear LSTM* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 208–224.

[62] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 3645–3649.

[63] W. Gan, S. Wang, X. Lei, M.-S. Lee, and C.-C.-J. Kuo, "Online CNN-based multiple object tracking with enhanced model updates and identity association," *Signal Process., Image Commun.*, vol. 66, pp. 95–102, Aug. 2018, doi: 10.1016/j.image.2018.05.008.

[64] A. Kampker, M. Sefati, A. S. A. Rachman, K. Kreisköther, and P. Campoy, "Towards multi-object detection and tracking in urban scenario under uncertainties," in *Proc. 4th Int. Conf. Vehicle Technol. Intell. Transp. Syst.*, Setúbal, Portugal, 2018, pp. 156–167.

[65] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 3464–3468.

[66] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. Comput. Vis. ECCV*, in Lecture Notes in Computer Science, A. B. Vedaldi, H, T. Brox, J. M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 107–122.

[67] A. Kumar K. C, L. Jacques, and C. De Vleeschouwer, "Discriminative and efficient label propagation on complementary graphs for multi-object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 61–74, Jan. 2017, doi: 10.1109/TPAMI.2016.2533391.

[68] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, "Learning by tracking: Siamese CNN for robust target association," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Las Vegas, NV, USA, Jun. 2016, pp. 418–425.

[69] J. Huang and W. Zhou, "Online multi-target tracking using recurrent neural networks," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, San Francisco, CA, USA, Jul. 2019, pp. 4225–4232.

[70] Y.-C. Yoon, Y.-M. Song, K. Yoon, and M. Jeon, "Online multi-object tracking using selective deep appearance matching," in *Proc. IEEE Int. Conf. Consum. Electron. Asia (ICCE-Asia)*, Jeju, South Korea, Jun. 2018, pp. 206–212.

[71] A. Scheel, C. Knill, S. Reuter, and K. Dietmayer, ''Multi-sensor multi-object tracking of vehicles using high-resolution radars,'' in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Gothenburg, Sweden, Jun. 2016, pp. 558–565.

[72] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee, *Multi-Class Multi-Object Tracking Using Changing Point Detection* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee, Eds. Amsterdam, The Netherlands: Springer, 2016, pp. 68–83.

[73] L. Chen, H. Ai, Z. Zhuang, and C. Shang, ''Real-time multiple people tracking with deeply learned candidate selection and person re-identification,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Diego, CA, USA, Jul. 2018, pp. 1–6.

[74] L. Hou, W. Wan, J.-N. Hwang, R. Muhammad, M. Yang, and K. Han, ''Human tracking over camera networks: A review,'' *EURASIP J. Adv. Signal Process.*, vol. 2017, no. 1, p. 43, Jun. 2017, doi: 10.1186/s13634-017-0482-z.

[75] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[76] P. Liu, X. Li, H. Liu, and Z. Fu, ''Online learned siamese network with auto-encoding constraints for robust multi-object tracking,'' *Electronics*, vol. 8, no. 6, p. 595, May 2019, doi: 10.3390/electronics8060595.

[77] L. Ren, J. Lu, Z. Wang, Q. Tian, and J. Zhou, *Collaborative Deep Reinforcement Learning for Multi-Object Tracking* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 605–621.

[78] A. Osep, W. Mehner, P. Voigtlaender, and B. Leibe, ''Track, then decide: Category-agnostic vision-based multi-object tracking,'' in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Brisbane, QLD, Australia, May 2018, pp. 3494–3501.

[79] L. Xiong, X. Zhang, J. Liao, and G. Yang, ''Multi-object tracking based on HOG template matching and non-maximum convergence algorithm,'' *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 10, no. 1, pp. 233–242, Jan. 2017, doi: 10.14257/ijsip.2017.10.1.23.

[80] V. Carletti, A. Greco, A. Saggese, and M. Vento, ''Multi-object tracking by flying cameras based on a forward-backward interaction,'' *IEEE Access*, vol. 6, pp. 43905–43919, May 2018, doi: 10.1109/access.2018.2864672.

[81] N. M. Al-Shakarji, G. Seetharaman, F. Bunyak, and K. Palaniappan, ''Robust multi-object tracking with semantic color correlation,'' in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Lecce, Italy, Aug. 2017, pp. 1–7.

[82] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, ''MOTS: Multi-object tracking and segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 7934–7943.

[83] C. Liu, R. Yao, S. H. Rezatofighi, I. Reid, and Q. Shi, ''Multi-object model-free tracking with joint appearance and motion inference,'' in *Proc. Int. Conf. Digit. Image Comput. Techn. Appl. (DICTA)*, Sydney, NSW, Australia, Nov. 2017, pp. 1–8.

[84] J. H. Yoon, C.-R. Lee, M.-H. Yang, and K.-J. Yoon, ''Structural constraint data association for online multi-object tracking,'' *Int. J. Comput. Vis.*, vol. 127, no. 1, pp. 1–21, Apr. 2018, doi: 10.1007/s11263-018-1087-1.

[85] Y.-C. Yoon, A. Boragule, Y.-M. Song, K. Yoon, and M. Jeon, ''Online multi-object tracking with historical appearance matching and scene adaptive detection filtering,'' in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Auckland, New Zealand, Nov. 2018, pp. 1–6.

[86] B.-N. Vo, B.-T. Vo, and M. Beard, ''Multi-sensor multi-object tracking with the generalized labeled multi-Bernoulli filter,'' *IEEE Trans. Signal Process.*, vol. 67, no. 23, pp. 5952–5967, Dec. 2019, doi: 10.1109/TSP.2019.2946023.

[87] L. Wen, D. Du, S. Li, X. Bian, and S. Lyu, ''Learning non-uniform hypergraph for multi-object tracking,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 8981–8988, doi: 10.1609/aaai.v33i01.33018981.

[88] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah, ''Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets,'' *Int. J. Comput. Vis.*, vol. 127, no. 9, pp. 1303–1320, May 2019, doi: 10.1007/s11263-019-01180-6.

[89] Y. Zou, W. Zhang, W. Weng, and Z. Meng, ''Multi-vehicle tracking via real-time detection probes and a Markov decision process policy,'' *Sensors*, vol. 19, no. 6, p. 1309, Mar. 2019, doi: 10.3390/s19061309.

[90] K. A. Shiva Kumar, K. R. Ramakrishnan, and G. N. Rathna, ''Inter-camera person tracking in non-overlapping networks,'' in *Proc. 11th Int. Conf. Distrib. Smart Cameras*, Sep. 2017, pp. 55–62, doi: 10.1145/3131885.3131912.

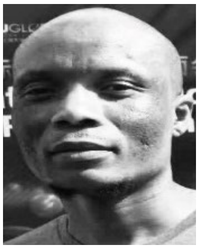
[91] J. Houssineau, D. E. Clark, S. Ivekovic, C. S. Lee, and J. Franco, ''A unified approach for multi-object triangulation, tracking and camera calibration,'' *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2934–2948, Jun. 2016, doi: 10.1109/TSP.2016.2523454.

[92] A. Bathija, ''Visual object detection and tracking using YOLO and SORT,'' *Int. J. Eng. Res. Technol.*, vol. 8, no. 11, pp. 705–708, Mar. 2019.

[93] A. Sharma, S. Anand, and S. K. Kaul, ''Reinforcement learning-based querying in camera networks for efficient target tracking,'' in *Proc. Int. Conf. Automated Planning Scheduling*, vol. 29, no. 1, pp. 555–563, Nov. 2020.

[94] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, ''Virtual worlds as proxy for multi-object tracking analysis,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 4340–4349.

[95] N. T. L. Anh, F. M. Khan, F. Negin, and F. Bremond, ''Multi-object tracking using multi-channel part appearance representation,'' in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Lecce, Italy, Aug. 2017, pp. 1–6.



LESOLE KALAKE received the B.S. degree in computer science and statistics and the B.Sc. degree (Hons.) in applied population science from the University of KwaZulu-Natal, South Africa, in 2004 and 2015, respectively, and the M.Sc. degree in information systems from the Kobe Institute of Technology, Japan, in 2017. He is currently pursuing the Ph.D. degree with the School of Information Engineering, Shanghai University. His research interests include machine learning and video/image processing.

WANGGEN WAN (Senior Member, IEEE) was born in Nanchang, Jiangxi, China, in 1961. He received the M.S. and Ph.D. degrees in electronic and information engineering from Xidian University, Xi'an, China, in 1988 and 1992, respectively. He was a Visiting Scholar with the Department of Computer Science, Minsk Radio Engineering Institute, formerly USSR, from 1991 to 1992, and a Postdoctoral Research Fellow with Xian Jiaotong University, from 1993 to 1995. He became a Visiting Professor with the Hong Kong University of Science and Technology, and Hong Kong Polytechnic University, from 1998 to 2004. He joined Shanghai University as a Full Professor, in June 2004. He is currently a full-time Professor and a Deputy Dean with the School of Communication and Information Engineering, Shanghai University. He has published one book, over 150 articles, and ten patents. His current research interests include multimedia signal processing, data mining, embedded systems, and system-on-chip design in a multimedia systems, digital audio/video processing, computer architecture, embedded systems, and system-on-chip design. He is an IET Fellow and an ACM Professional Member.

LI HOU received the B.S. degree in communication engineering and the M.S. degree in power electronics from the Liaoning University of Technology, in 2003 and 2006, respectively, and the Ph.D. degree in communication and information systems from Shanghai University, in 2017. In 2006, she joined the School of Information Engineering, Huangshan University, where she has been an Associate Professor. Her current research interests include machine learning, video/image processing, and big data mining.