Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 7549816, 26 pages
https://doi.org/10.1155/2020/7549816

Research Article

Data Association Methods via Video Signal Processing in Imperfect Tracking Scenarios: A Review and Evaluation

Hui Li,[1] **Yapeng Liu**,[1] **Wenzhong Lin**,[2] **Lingwei Xu**,[1,2] **and Junyin Wang**[1]

[1]*School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China*
[2]*Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350108, China*

Correspondence should be addressed to Lingwei Xu; xulw@qust.edu.cn

Received 24 June 2020; Revised 3 August 2020; Accepted 7 August 2020; Published 31 August 2020

Guest Editor: Junpeng Shi

In 5G scenarios, there are a large number of video signals that need to be processed. Multiobject tracking is one of the main directions in video signal processing. Data association is a very important link in tracking algorithms. Complexity and efficiency of association method have a direct impact on the performance of multiobject tracking. Breakthroughs have been made in data association methods based on deep learning, and the performance has been greatly improved compared with traditional methods. However, there is a lack of overviews about data association methods. Therefore, this article first analyzes characteristics and performance of three traditional data association methods and then focuses on data association methods based on deep learning, which is divided into different deep network structures: SOT methods, end-to-end methods, and Wasserstein metric methods. The performance of each tracking method is compared and analyzed. Finally, it summarizes the current common datasets and evaluation criteria for multiobject tracking and discusses challenges and development trends of data association technology and data association methods which ensure robust and real time need to be continuously improved.

1. Introduction

With the advent of the 5G [1–3], there is a great improvement in transmission speed [4], and a large amount of video signal information in imperfect scenarios has to be processed [5, 6]. The rapid development of neural networks [7] makes the processing of these signals no longer difficult. Multiobject tracking (MOT) has an important position in video signal processing. It has a huge application prospect in automatic driving, intelligent monitoring, radar imaging, and so on [8, 9]. Up to now, most of the MOT methods are based on ideal scenarios, for example, target density is low and occlusion rarely occurs. In practical applications, however, imperfect scenarios may exist, and a model mismatch would appear between the ideal one and the real one. There are still no MOT methods suitable for all scenes, and neural network has low versatility [10], so MOT is still considered to be a very challenging task. MOT can be divided into two directions: MOT based on initial frame and MOT based on detection. However, MOT based on initial frame can only track marked objects in the first frame, and new objects cannot be processed. MOT based on detection is different from that based on initial frame. Firstly, all objects in each frame are detected, and then the object tracking is carried out by using detection results, which is called the data association stage [11]. The latter is the mainstream direction of multiobject. With the application of deep learning in object detection, the detection performance has gradually matured. Therefore, the progress of data association determines the progress of MOT.

Data association is the process of comparing and matching all the objects in the object detection stage, drawing the association track between each frame, and comparing with the association track between real data for learning, so that the network can predict the correct movement track of the infinite approaching object. Simply

speaking, all detected object data are divided into Nm sets, and the probability of the same object in each set is infinitely close to 1 [12]. There are many hypotheses in this process: (1) multiobjective, (2) false detection, (3) missing detection, and (4) ambiguity. If detection results of all the objects in the video are the normal state of real objects, then there is no need for data association. The premise is no missed detection, no false detection, and no observation noise. However, there are always errors in the actual object detection process and lack of prior knowledge of the tracking environment. In addition, the number of objects and the detected data from the false object or the real object cannot be known in advance. Moreover, the relaxation of each hypothesis introduces a layer of uncertainty, and the object of tracking is to find the object closest to the real object among all the uncertainties [13]. It can be seen that data association is a key and difficult point in MOT. The complexity and efficiency of association methods will have a direct impact on the performance of MOT. When the object is a single object, there is no need for data association, but when there are multiple tracking objects and the objects are very similar, the role of data association is particularly important; of course, the process of association will become more complex [14]. Although there are many data association methods [15–20], how to build a more efficient data association method is still one of the research difficulties.

The range of association of MOT data mentioned in the previous paper [21–25] is very wide, and most of them deal with the information collected by sensors, such as different forms of object track information collected by radar, sonar, etc., after pairing and comparing the object track tracked by sensors with the predicted object track, and the correct object track is finally determined [26–29]. The paper focuses on the data association method based on video MOT, which only deals with video information and no other sensors. Of course, the above sensor association method can also be applied to video MOT. The success of the final association depends on the similarity function used to match the object and detect it. Traditionally, the cost function is made by hand based on the representation of color histogram, boundary box position, and linear motion model [30, 31], which cannot be generalized across tasks and used in complex tracking scheme [32]. With the application of deep learning in MOT, data association based on deep learning is a framework that integrates data association methods and deep feature information, so as to get rid of constraints of traditional data association relying on manual production and gradually become the mainstream research direction. This paper expounds the data association method from two parts: traditional data association method and data association method based on deep learning, as shown in Figure 1. Among them, traditional data association methods are described from probability-based data association, feature matching-based data association, transformation-based data association, and hierarchical data association. Data association method based on deep learning is described from data association based on the SOT method, end-to-end data association, and data association based on Wasserstein measurement.

2. Traditional Data Association Methods

In the first section, the range of MOT data mentioned in previous papers is very wide, and most of them deal with sensor information to achieve the purpose of "multiple trajectories determine a trajectory." To put it simply, the true trajectory of the tracked object is judged based on comparison and unity of results of multiple tracking methods. The reason why so many sensors are used for object tracking is that, in MOT applications, such as ballistic missile defense, a small deviation may cause serious consequences. Classic sensor-based data association methods include probability data association filter (PDA), joint probability data association filter (JPDA), and statistical data association filter (the order statistics PDA and OSPDA). Vision-based data association method is a branch of the abovementioned association method and an important direction in computer vision, but there are few review articles in this regard, and this article will make up for this gap. First, this article summarizes from traditional data association methods, and the time axis of the representative methods is shown in Figure 2.

2.1. Probability Class Methods. Joint probabilistic data association filtering (JPDA) method is a very classic method in probabilistic data association. Its purpose is to calculate the association probability between observation data and each object and consider all the effective response waves that may originate from each specific object, but they have different probabilities from different objects. The advantage of the JPDA method is that it does not require any prior information about the object and clutter. It is one of the better methods for tracking multiple objects in a clutter environment. However, as the number of objects and measurements increases, the calculation of JPDA will show a combination of explosions, resulting in complex calculations. Using its advantages and improving its shortcomings, many papers combining the advantages of the JPDA method and methods for improving the JPDA method and other methods of probability class are emerging.

Wang and Nguang [33] proposed a multiobject video tracking method based on improved data association and hybrid Kalman/H∞ filtering. Firstly, they extracted multiple features of the object in the video to form a fusion feature matching matrix. Secondly, they combined improved probabilistic data association and developed a simplified joint probabilistic data association for multiobjective association. Finally, they proposed that Kalman/H∞ filtering of mixed state estimation covariance can achieve robust and fast state estimation of video objects. Results show that, while ensuring the efficiency and accuracy of video object association, the performance in noisy environment is also more reliable and accurate. Improved data association diagram based on SJPDA and IPDA is shown in Figure 3.

Xu et al. [34] describe how to develop a real-time multitarget detection and tracking system for small drones. They selected and implemented the deep learning visual object detection algorithm based on YOLO and the MOT

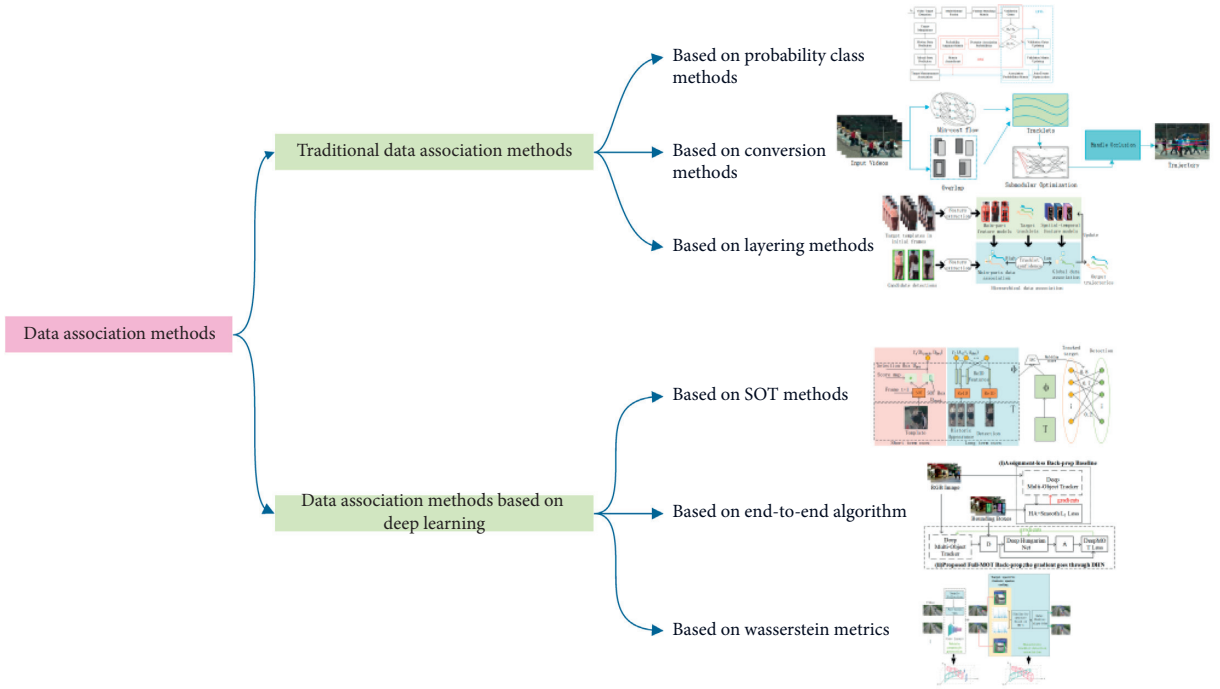

Figure 1: Overall classification chart.

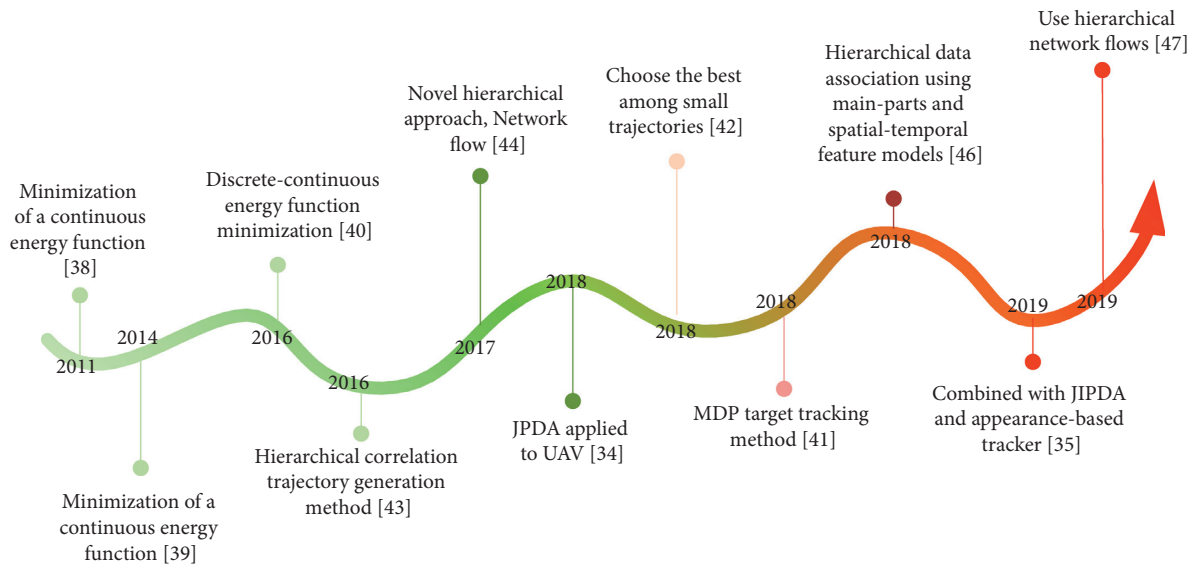

Figure 2: Time axis of traditional data association methods.

algorithm based on JPDA. Results show that JPDA has a good association performance. Bićanić et al. [35] studied the interaction between kinematics and appearance cues in pedestrian MOT and established the consistency of pedestrian MOT. Tracking detection methods based on deep learning detector, appearance-based deep correspondence embedding, and joint integrated probabilistic data association (JIPDA) were studied. Combined with JIPDA and appearance-based tracker, deep correspondence embedding was adopted. A convolutional neural network detector is pretrained on the COCO dataset for accurate pedestrian detection and as an input to JIPDA-based tracking method, where the state consists only of pedestrian motion cues (position and velocity). The proposed pedestrian tracker with motion cues currently ranks first on the 3DMOT2015 online benchmark [36]. Kikuchi [37] proposed a novel object tracking method, which embeds PDAF into the moving field estimation framework to deal with multiframe tracking and physical constraints. Compared with the traditional PDAF, the proposed method can achieve robust estimation using constraints and horizon. The algorithm framework is shown in Figure 4.

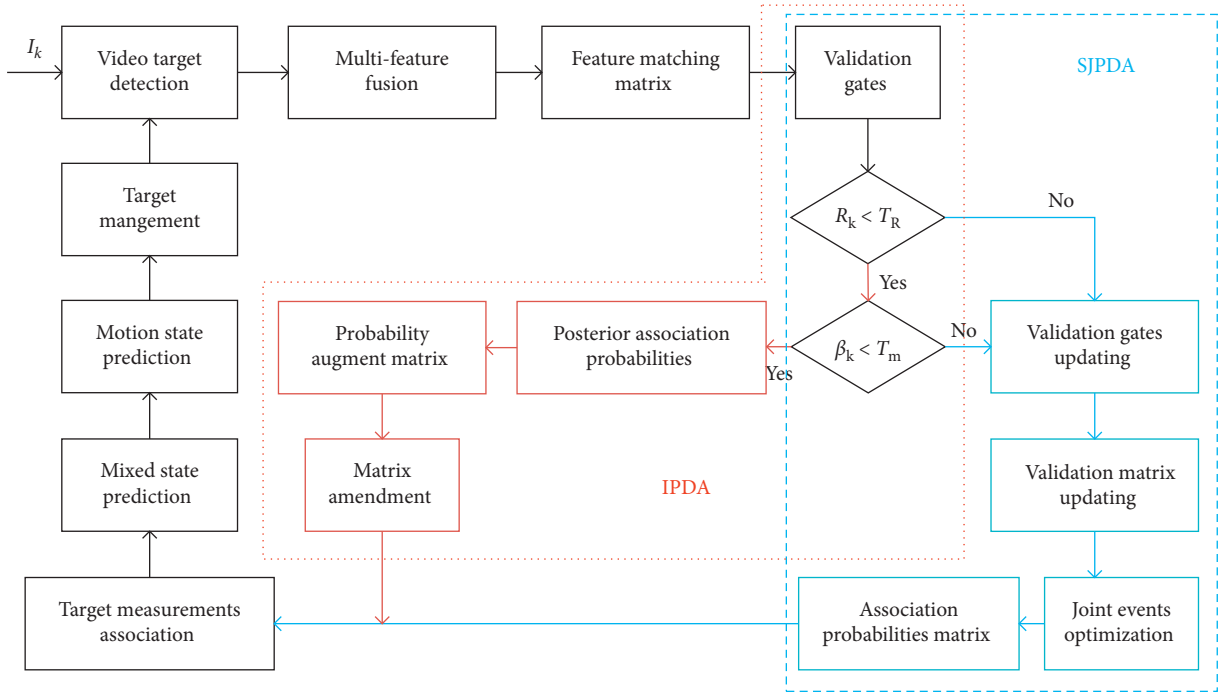

Figure 3: Improved data association diagram.

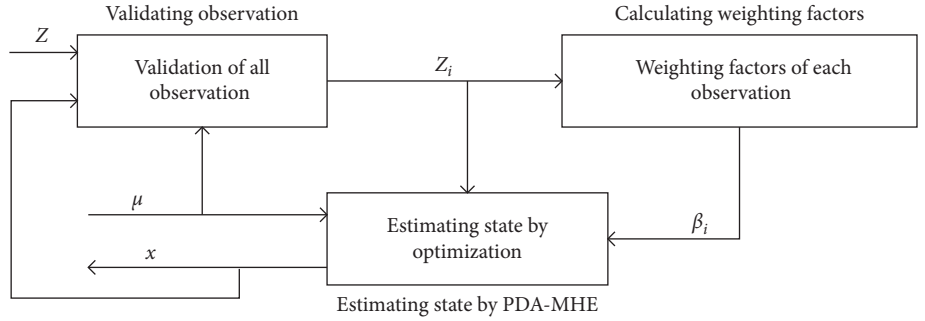

Figure 4: Algorithm framework.

2.2. Conversion Methods. The above two directions of data association methods are directly looking for various methods to solve the data association problem of MOT, but due to the complexity of the scene, it is difficult to achieve object tracking, so the data association problem is converted into other similar problems to solve, and it may produce a good effect.

Andriyenko and Schindler [38] transformed the object tracking problem into a continuous energy function minimization problem. The problem of energy minimization is to find the optimal solution because in the solution space, each solution corresponds to a loss function. The solution process is to express the loss function and find an optimal way to find the optimal solution. Different from other methods, it focuses on finding an energy function that can truly express the problem, rather than a function that facilitates optimization. A suitable optimization framework is constructed to find the "strong" local minimum of the given energy (function). The conjugate gradient method is extended by periodic cross-dimensional jump methods in the framework. These changes allow the search to get rid of "weak" local minima, thereby exploring a larger part of the variable dimension search space, while reducing energy. The influence of different composition energy functions is shown in Figure 5.

In Figure 5, the upper row shows the higher configuration, and the lower row shows the lower configuration. Gray indicates the position of the object with a higher probability.

Milan et al. [39] also convert the MOT problem into a continuous energy minimization problem, which can more fully express the problem rather than global optimization. In addition to image features, the energy function also considers physical constraints such as object dynamics, mutual exclusion, and tracking persistence. In addition, explicit occlusion reasoning is used to process some image features, and the appearance model is used to disambiguate different objects. By alternately between continuous conjugate gradient descent and discrete cross-dimensional jump movement, the strong local minimum of the proposed nonconvex energy is found. This movement can reduce energy, allowing

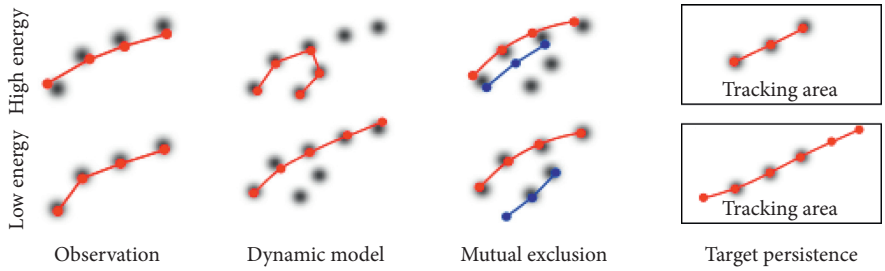

Figure 5: Influence of different composition energy functions.

the search to avoid weak minimums and explore larger parts of the search space with different dimensions. The influence of different components of the energy function is shown in Figure 6.

In Figure 6, the upper row shows a higher configuration, and the lower row shows the lower value for each individual item. Dark, smooth spots indicate the detection location. Different colors of the object location (marked with circles) indicate a distinguishable appearance between objects.

Milan et al. [40] made improvements on the basis of [39] and proposed a unified discrete-continuous energy function minimization. The function includes two tasks: data association and reconstruction of the actual trajectory describing the spatiotemporal motion pattern of each object. Trajectory characteristics are obtained through global label cost, which describes the physical properties of a single track. In addition, to avoid collisions, a pair of label costs is introduced to describe the interaction between objects. By selecting a suitable form for a single energy component, a powerful discrete optimization technique is used to process the data association, and the shape of the single trajectory is updated by continuous energy minimization based on the gradient.

Zuo et al. [41] adopted the Markov decision process (MDP) dynamic object tracking method to turn MOT problem into MDP model strategy problem. An MDP model represents the life cycle of an object. Multiple MDPs represent multiple tracked objects. The life cycle tracking process is shown in Figure 7. In the tracking process, the feature points generated by the traditional TLD method are replaced by the strong angles generated by the Shi-Tomasi angle method, so that there are more stable target feature points in the tracking process. Similar function learning of data association is equivalent to learning of MDP strategy, which uses reinforcement learning method, having dual advantages of online learning and offline learning.

Shen et al. [42] converted the object data association problem into the problem of selecting the optimal small segment from the set of small trajectory segments. The idea is clear and the occlusion problem of the object is effectively solved, but the process is more complicated. The method is divided into two stages. One is to propose a new trajectory selection strategy to solve the occlusion problem; the other is to use the MOT problem as a submodule maximization problem subject to association constraints. Submodule function is selected correctly from the set of candidate trajectories to form object trajectory. The flowchart of the MOT method is shown in Figure 8.

In Figure 8, first, trajectories are generated by overlapping criteria and minimum cost flow, respectively. Then, maximization of the submodule optimization problem is used to solve MOT problem, and the occlusion processing is also embedded in the framework.

2.3. Layering Methods. Layering is the most common idea in software design. The most typical one is the seven-layer model of ISO. Each layer has its own specific functions, and each layer uses some kind of "connection" to ensure that the layers are coordinated and connected with each other, so as to achieve the perfect and orderly operation of the overall function. In the data association of video MOT, the idea of layering is also very popular. However, each layer here is more complementary, each layer can complete the overall task, and the existence of other layers is to make up for defects in a certain layer when completing tasks. For example, when only using the appearance model to associate objects between video frames, the appearance may suddenly change, making the association unsuccessful or wrong. At this time, using the motion model and the appearance model to judge at the same time will get correct association.

Zhu et al. [43] propose a method for generating MOT trajectory using hierarchical association. Firstly, after object detection, the online discriminant analysis apparent model is combined with the AdaBoost algorithm to generate the initial tracking trajectory. Secondly, the discontinuous and segmented trajectories are optimized using the Hungarian algorithm to generate stable and accurate trajectory segments. Finally, an intelligent extrapolation algorithm that minimizes energy is used to achieve a more continuous and smooth trajectory. Framework diagram of hierarchical association MOT method is shown in Figure 9.

Ali et al. [44] proposed a network flow method to connect low-level detection and high-level trajectories. In each step of the hierarchy, a new scoring system, ConfRank, will be used to evaluate the candidate's confidence. The first stage outputs a collection of unconnected reliability detections and short trajectories. For each individual test, determine whether it belongs to existing tracking trajectories or new tracking trajectories. Compared with the latest technology, a good competitive result is obtained by having fewer ID switches on multiple datasets. The algorithm framework is shown in Figure 10.

In Figure 10, on the basis of detections, ConfRank is calculated by the spatiotemporal neighborhood and confidence of

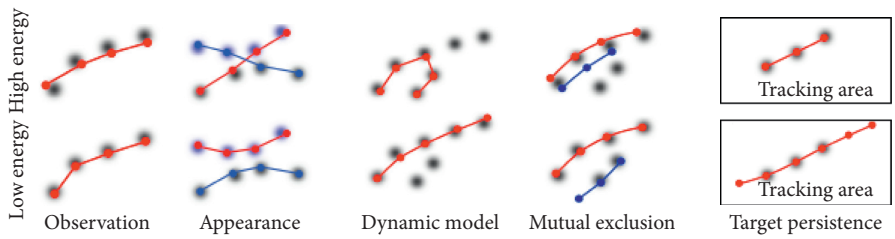

Figure 6: Influence of different components of the energy function.

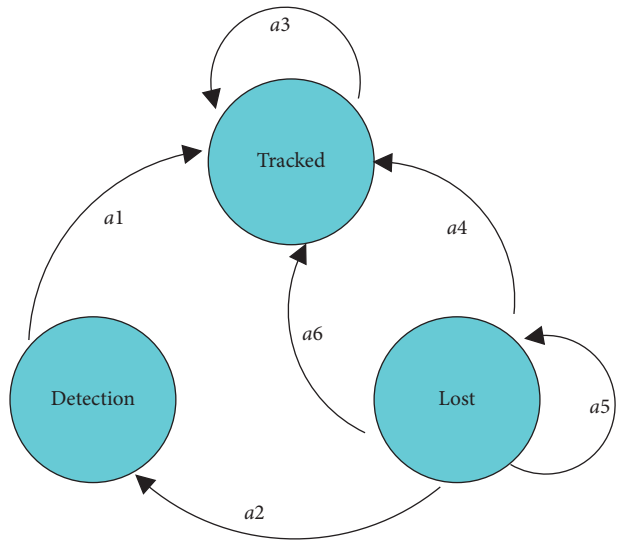

Figure 7: Life cycle tracking process.

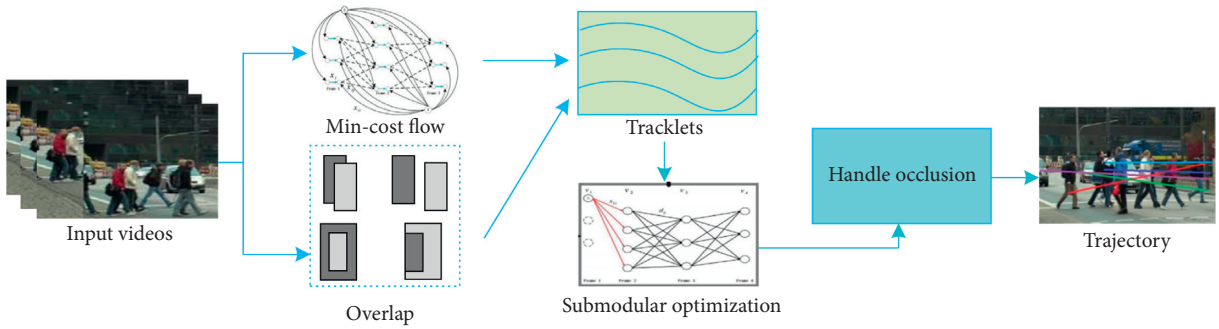

Figure 8: Minimum cost flow method flowchart.

each detection. Overlap and ConfRank score match detections are used to generate initial trajectories (Figure 10(a)). Remaining single detection objects are extended using the single object tracking (SOT) methods (Figure 10(b)). All trajectories are matched to restore the entire trajectory (Figure 10(c)), and its confidence is measured by ConfRank (Figure 10(d)).

Liu et al. [45] proposed hierarchical ideas to improve data association. In order to enable the algorithm to achieve accurate MOT, the proposed method embeds multiple Gaussian uncertainty (MGU) theories into a single motion model for each object and then applies interaction constraints to reassociate small trajectories with lower confidence. This model can not only associate objects more accurately across frames but also dynamically constrain each other globally. The framework of HMOT is shown in Figure 11.

In Figure 11, an online MOT framework under complex conditions is proposed. The framework uses layered ideas to improve data association. MGU is embedded in the motion model of each object, and then interactive constraints are applied to reassociate small trajectories with lower confidence for accurate MOT.

Liu et al. [46] propose a hierarchical data association method using subject and spatiotemporal features. In the tracking process, wavelet confidence is used to achieve hierarchical data association. According to different trajectory confidence values, main-part and spatiotemporal feature

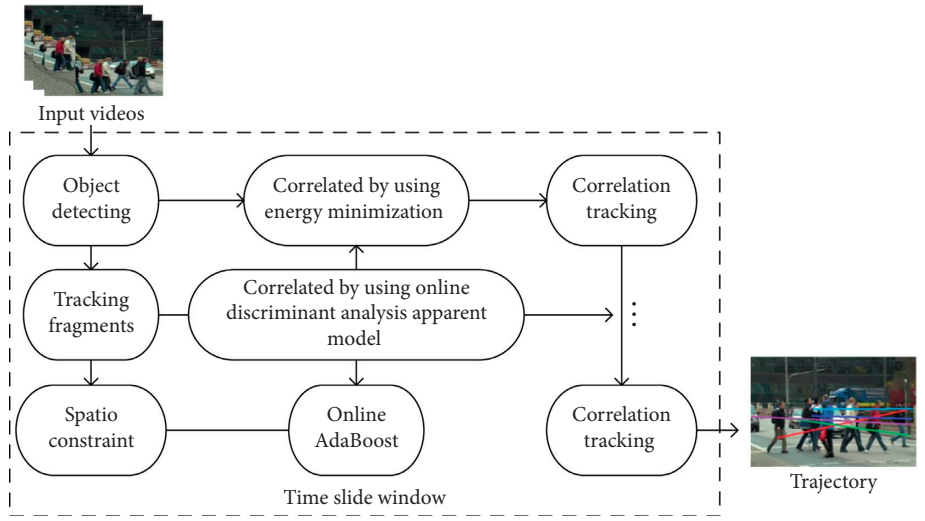

Figure 9: Framework diagram of hierarchical association MOT method.

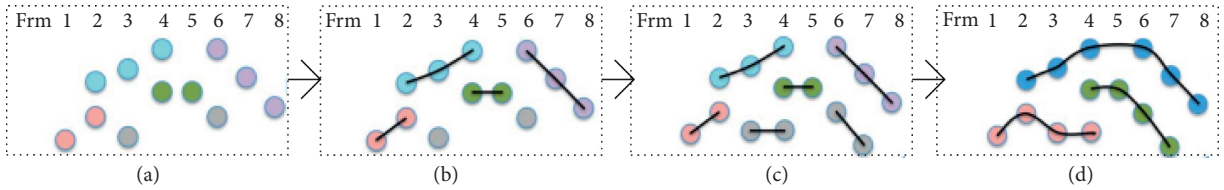

Figure 10: Flowchart of approach.

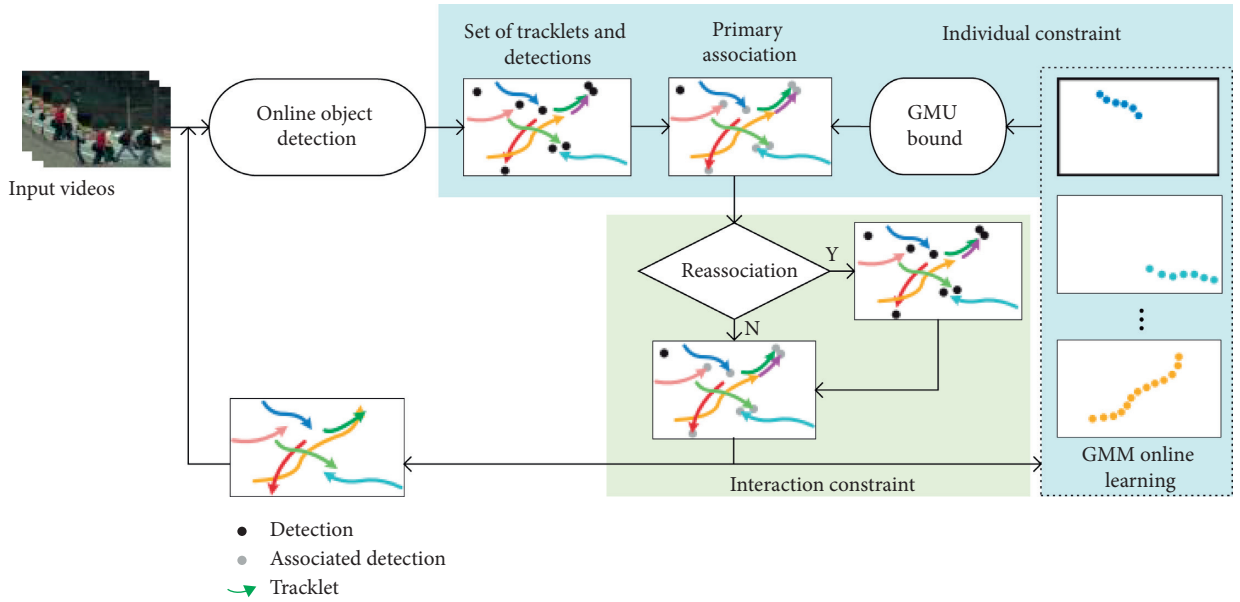

Figure 11: Framework of HMOT.

models are used for subject and global data association, respectively. As a result, the proposed method uses the Hungarian algorithm to obtain the best association pair between short trajectory and detection. The algorithm framework is shown in Figure 12.

Lu et al. [47] propose a global optimization method combining nonlinear motion and layered network flow to MOT. Each node in the network represents a trajectory, and each edge represents an affinity score. Nonlinear motion diagram is used to explain the nonlinear motion pattern

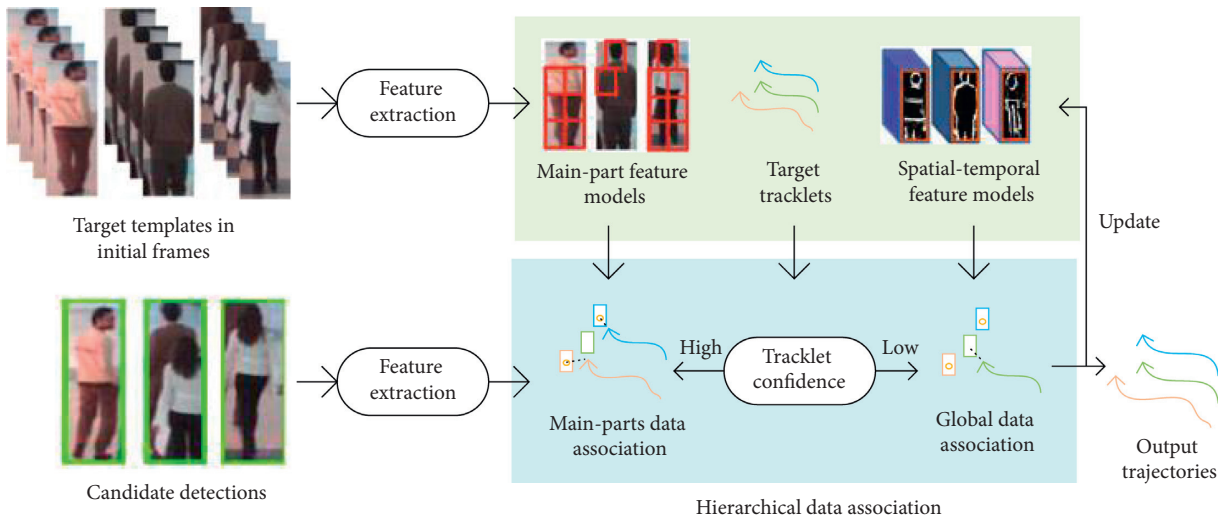

Figure 12: Algorithm framework.

between adjacent small tracks and obtain the motion affinity when there is a time interval. Results show that the method can realize continuous tracking when the direction of motion is changed and occluded completely. The algorithm framework is shown in Figure 13.

Traditional data association method evaluation and representative work are shown in Table 1.

3. Data Association Methods Based on Deep Learning

Deep learning can make computers have human-like intelligence, and their development prospects must be infinite. In recent years, with the application of deep learning in video object tracking, MOT has made great progress. Although its performance is far inferior to that of SOT based on deep learning, its great development prospects in autonomous driving, intelligent monitoring, and other industries as well as the broad applications of deep learning have made research in this area fierce. The participation of a large number of scholars has made relevant achievements endless. Among them, data association is an indispensable link in MOT. Judging from the current development of MOT, data association can be recognized to determine the development process of MOT. Fortunately, driven by current trends, a number of excellent literature works have emerged based on single object tracker data association or object detection based data association.

Compared with traditional data association methods, the performance of data association methods based on deep learning is greatly improved. The reason is that deep learning is a method of computer self-learning the characteristics of objects. The process of feature learning is the process of model construction, which reduces the lack of features caused by manual design. However, the corresponding calculation amount is also greatly increased, and the more complicated the problem is, the higher the calculation power is required. Although the current level of technological development can meet computing requirements, there is still a long way to go to achieve generalization. This is also a direction that deep learning needs to overcome, namely, to achieve high efficiency and low energy consumption.

Through reading a lot of domestic and foreign literature studies, this paper divides data association based on deep learning into three general directions: data association based on SOT method, end-to-end data association, and data association based on Wasserstein metric. The time axis of the representative methods is shown in Figure 14.

3.1. SOT Methods. With the application of deep learning in SOT, its tracking performance has become very high, and the tracking accuracy rate has reached more than 95%. In contrast, there is much room for improvement in multiobject performance. Efficient SOT performance has made many scholars work hard on MOT based on single object tracker, so it has also derived a number of excellent literature works.

Zhu et al. [48] proposed an online MOT method, which merged SOT and data association into unified framework to deal with noise detection and frequent interaction between objects. For data association, dual matching attention network (DMAN) of the spatiotemporal mechanism is proposed. There are three contributions for this method: applying SOT tracker to MOT and introducing a new cost-sensitive tracking loss function based on the latest tracker; a spatial attention network is proposed to deal with the problem of noisy detection and occlusion in MOT; and a temporal attention network is proposed to adaptively assign different levels of attention scores to different observation objects in the trajectory.

In Figure 15, the process is mainly composed of three tasks: detection, SOT, and data association. The state of each object switching between tracking and loss depends on the reliability of tracking. SOT is applied to generate a trajectory of the tracking object, and the data association compares the trajectory with candidate detection in order to allocate the missing object.

Figure 16 consists of a space attention network (SAN) and a time attention network (TAN). Given a candidate

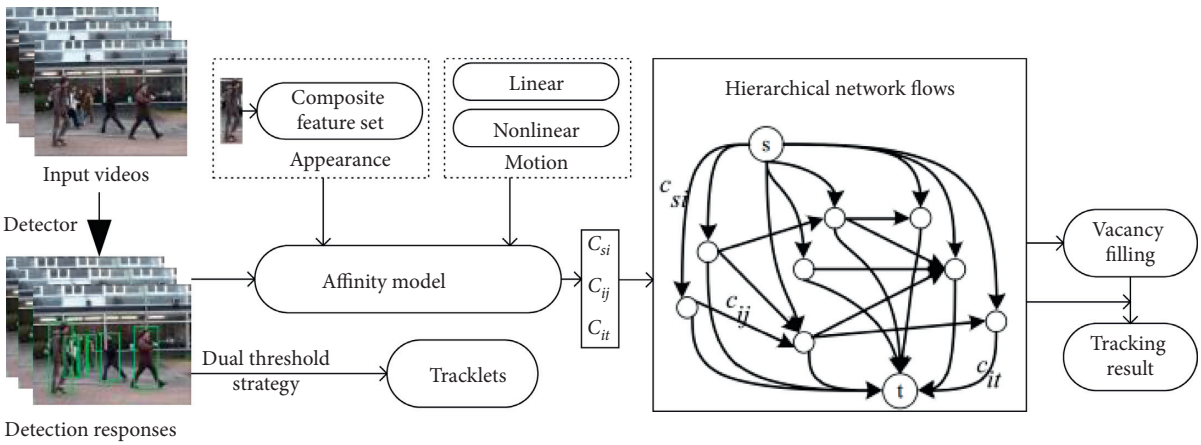

Figure 13: Algorithm framework.

Table 1: Summary and evaluation of previous studies on traditional data association.

Author/organization	Years	Technical characteristics	Dataset	MOTA↑(%)	MOTP↑(%)	IDs↓
Shuoyuan Xu/Cranfield University	2018	JPDA applied to UAV [34]	TUD PETS	56.3 32.7	—	—
Bićanić/University of Zagreb	2019	Combined with JIPDA and appearance-based tracker [35]	3DMOT2015	55.9	64.0	486
Anton Andriyenk/TU Darmstadt	2011	Formulate MOT as minimization of a continuous energy function [38]	terrace2 TUD PETS ped1-c2	88.1 60.5 81.4 48.0	78.1 65.8 76.1 75.5	11 7 15 4
Anton Milan/TU Darmstadt	2014	Formulate MOT as minimization of a continuous energy function [39]	PETS-S2L1 PETS-S2L2 PETS-S2L3	90.6 56.9 45.4	80.2 59.4 64.6	11 73 27
	2016	Discrete-continuous energy function minimization [40]	Datasets[1]	56.3	66.0	42
Guoyu Zuo/Beijing University of Technology	2018	MDP object tracking method [41]	PETS-S2L1	88.1	68.7	5
Jianbing Shen/Beijing Institute of Technology	2018	Choose the best among small trajectories [42]	TUD-Crossing PETS09-S2L2	60.2 21.3	77.2 70.7	32 251
Songhao Zhu/Nanjing University of Posts and Telecommunications	2016	Hierarchical association MOT trajectory generation method [43]	Datasets[2]	—	—	25
Ali Taalimi/University of Tennessee	2017	Novel hierarchical approach, network flow [44]	MOT2015	33.9	71.4	12.1
Junying Liu/Beihang University	2018	Hierarchical MOT [45]	PETS-S2L1 PETS-S2L2 ETH KITTI	91.04 52.83 67.09 72.89	87.85 79.54 81.05 90.61	18 265 85 119
Hongbin Liu/Shandong University	2018	Hierarchical data association using main-part and spatial-temporal feature models [46]	CAVIAR Parking Lot MOT15	88.6 76.2 17.5	89.4 73.9 70.9	3 34 683
Weigang Lu/Jiangnan University	2019	Hierarchical network flows [47]	MOT16	68.2	60.2	953

[1]S2. L1, S2. L2, S2. L3, S2. L1-2, S2. L2-1, Stadtmitte. [2]TUD, PETS2009.

detection and an object tracker sequence as input, SAN extracts combined features by comparing the detection results with each sample in the tracker. TAN infers whether the detected and tracked objects belong to the same object by integrating information from the entire tracker.

Chu and Ling [49] proposed an end-to-end model, named FAMNet. The process is as follows: SOT ⟶ affinity estimation ⟶ data association ⟶ postprocessing. The core lies in completing K-frame joint data association training through local association form.

In Figure 17, subnet in yellow background consists of FAMNet. F_{ik} is a set of features extracted from each frame. See Figure 18 for details.

SOT twin network and the affinity network are merged, SOT backbone network is used as feature extractor, and

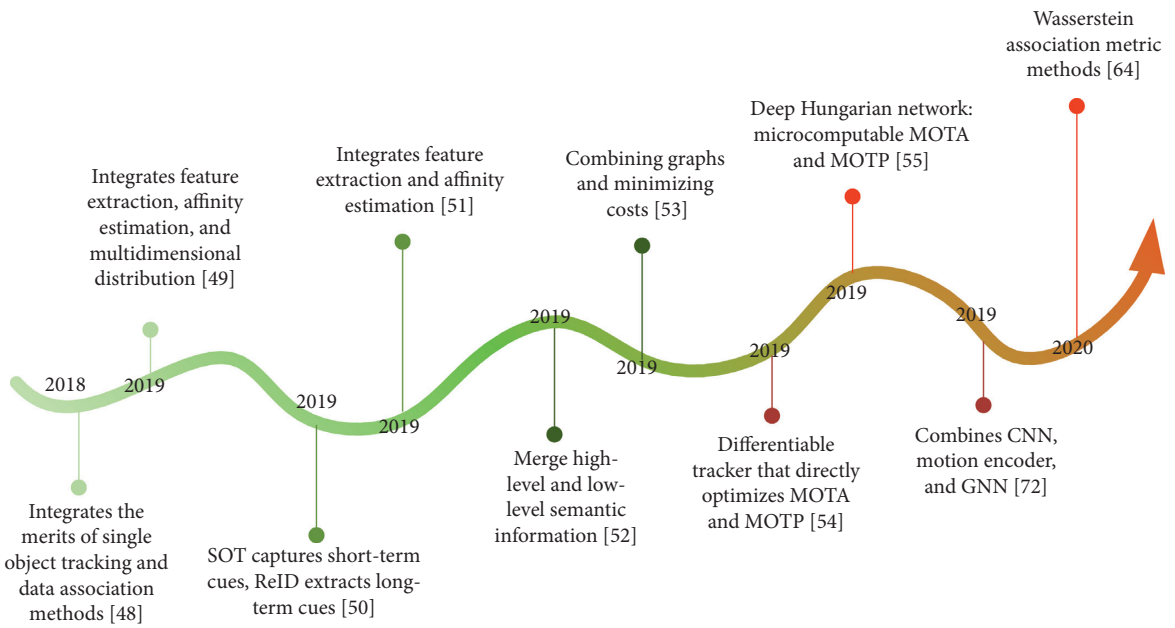

Figure 14: Time axis data association methods based on deep learning.

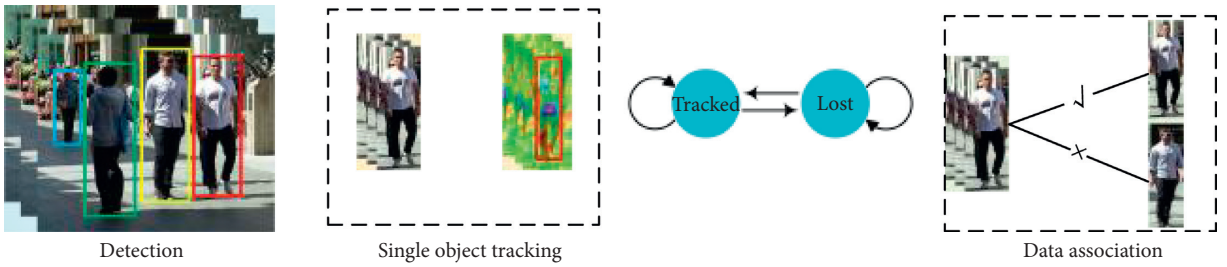

Figure 15: MOT process.

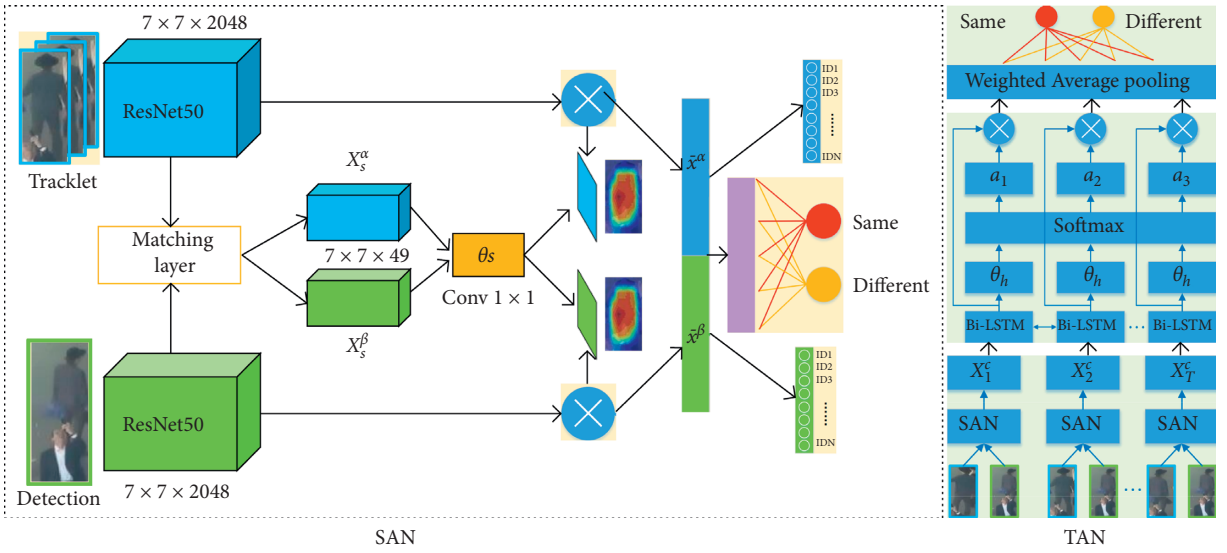

Figure 16: DMAN framework.

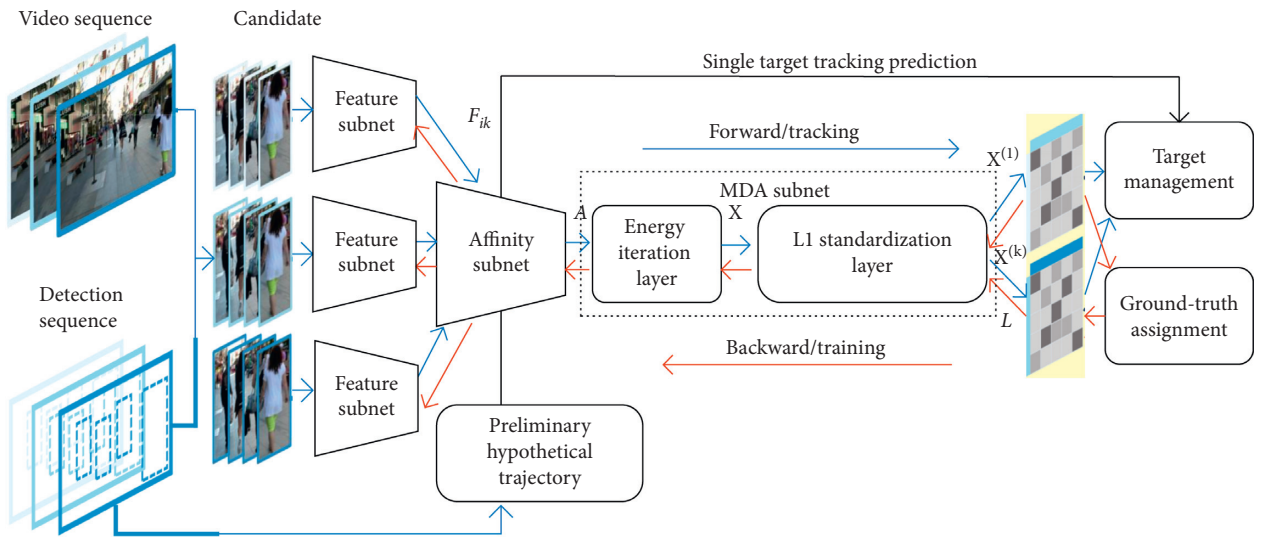

Figure 17: FAMNet-based tracking system.

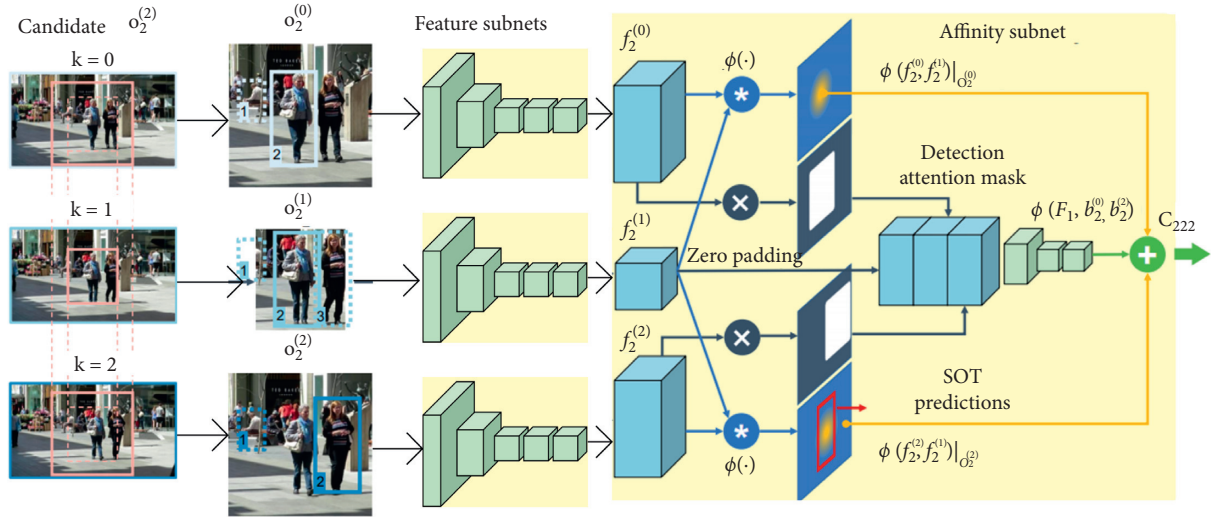

Figure 18: Method framework.

confidence map output by the network is used as the affinity basis. The method framework is shown in Figure 18.

In Figure 18, for the specified object, the current frame ($k = 1$) and the expansion area of the front and back frames are cut, and then the SOT process is performed twice through the twin network to get the upper and lower confidence map output. Among them, there are three detection attention masks. The obtained affinity is the sum of the information obtained by two SOT confidence maps and the affinity obtained by detection attention mask.

Feng et al. [50] proposed a unified MOT framework, including a SOT subnetwork for capturing short-term cues, a ReID subnetwork for extracting long-term cues, and the use of regularization based on short-term and long-term cues. Newton-enhanced decision tree trains a switcher-aware classifier (SAC) for data association. Short-term clues help to detect false negatives (FN), long-term clues avoid serious errors when occlusion occurs, and SAC learns how to combine multiple clues in an effective manner and improve the robustness of data association.

In Figure 19, SOT captures short-term clues and ReID captures long-term clues. SAC predicts their location. Using short-term and long-term cues of the object, detection results, and switchers, detection results match the object.

Short-term clues based on SOT are based on the Siamese-RPN framework, as shown in Figure 20. Based on long-term clues of the pedestrian reidentification ReID network, the main body of the ReID subnet is Inception-v4. ReID features are extracted from the last FC layer before classification.

3.2. End-to-End Methods. Traditional machine learning is often composed of many components, and results are obtained in a pipelined manner. For example, a typical natural language processing (NLP) problem includes multiple independent steps such as word segmentation, part-of-speech

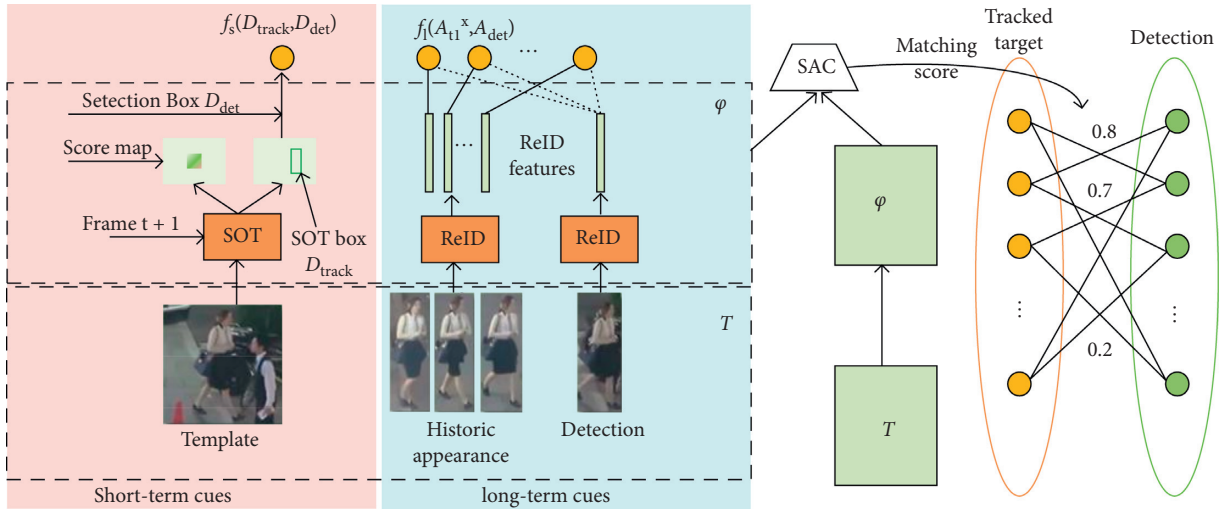

Figure 19: The overall design of the MOT framework.

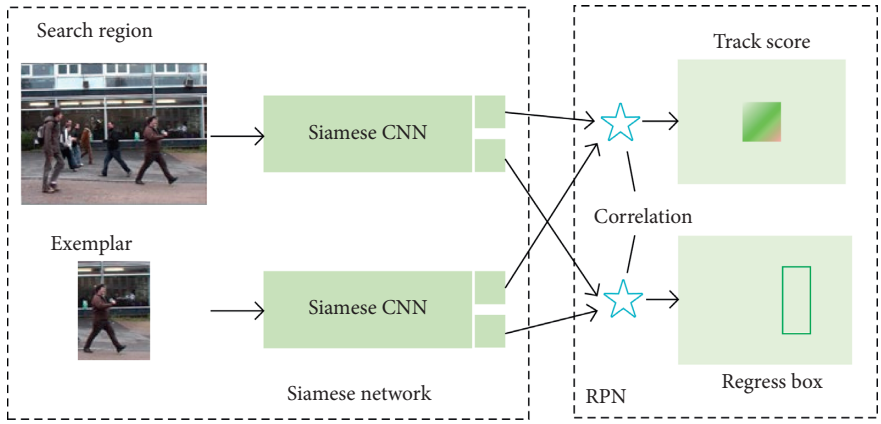

Figure 20: SOT based on the Siamese-RPN framework.

tagging, syntactic analysis, and semantic analysis. Each step is an independent task, and the results will affect the following, thereby affecting the overall training results. The above method is non-end-to-end. In end-to-end training, an error will be generated in the comparison between the prediction result and the real result. And this error will be propagated back in the network to continuously adjust the prediction results so that it keeps approaching the true value. All operations in the middle are included in self-contained neural network and are no longer divided into multiple modules for processing. The above method is end-to-end. End-to-end training network needs large datasets to get a good result. If it is a small dataset, non-end-to-end is better. With the development of technology and the accumulation of time, the scale of the datasets is getting larger and larger, so that the basic conditions for end-to-end deep learning are already available. Recently, there are more and more MOTs based on deep learning, there are features for feature extraction, and some are used to improve the single object tracker, such as Section 3.1. This section introduces the end-to-end data association.

Sun et al. [51] aimed at data association still relying on traditional manual design, such as appearance, motion, spatial proximity, and grouping, to calculate the similarity between different frame objects and proposed the deep association network (DAN) to learn the compact and comprehensive features of predetected objects at multiple levels of abstraction and perform detailed pairing of these features in any two frames to infer the similarity of objects. Aiming at the problem of object occlusion, a twin network with shared weights is used to jointly train a pair of frames that do not need to be continuous, and the features of the object are extracted for object association analysis. In order to track the object in and out of the video frame, an additional column is added to the calculation of the similarity matrix to indicate the in and out of the object. The experimental results on the tracking datasets MOT15, MOT17, and UA-DETRAC show that this method has the best results at that time.

In Figure 21, DNA is divided into two stages: feature extraction stage and affinity estimation. In the feature extraction stage, by inputting a pair of video frames I_t and C_t, I_{t-n} and C_{t-n} that do not need to be continuous with the

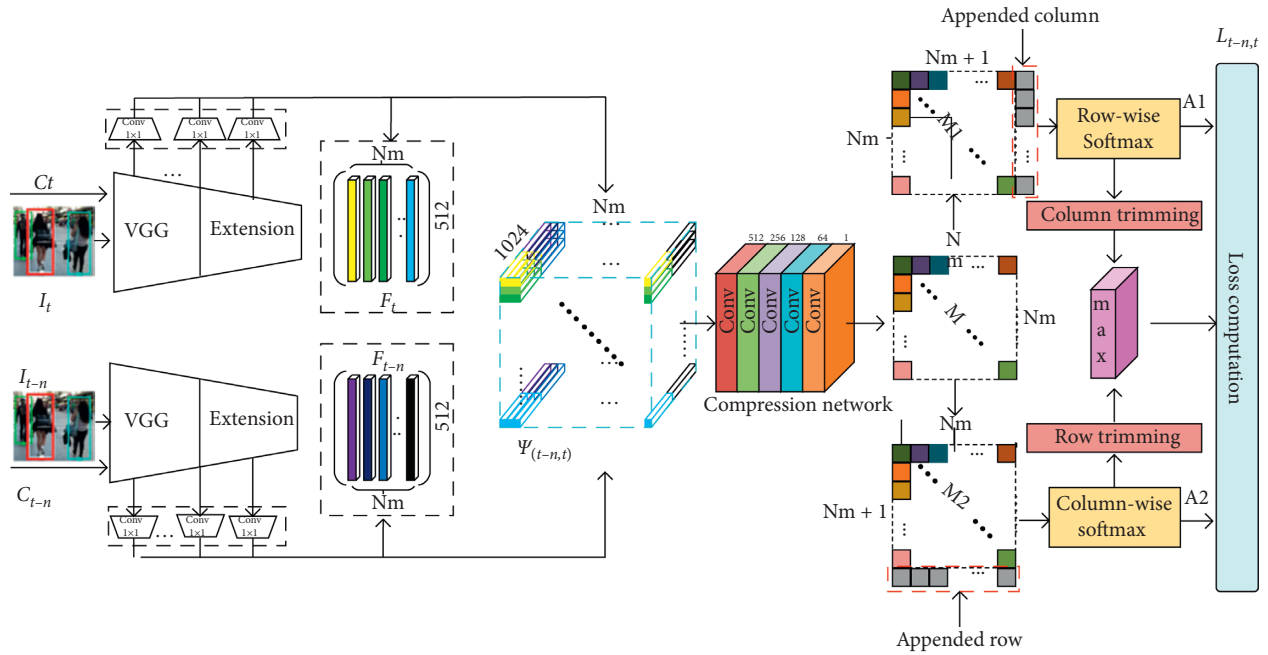

Figure 21: Schematic diagram of DNA.

object center point information, after processing by twin network, features F_t and F_{t-n} of each object in the video frame are obtained n. In the affinity calculation stage, $\psi_{t-n,\,t}$ is generated by connecting and combining F_t and F_{t-n}, and the association matrix is generated after the compressed network is processed, and the corresponding rows and columns are added to represent the in and out video frames of the object, and then the final loss is calculated to obtain final association result.

Han et al. [52] have the same overall framework as in [51], which divides data association into two parts: feature extraction and association evaluation. A framework is proposed to obtain the appearance characteristics of the object in an end-to-end manner, which includes low-level and high-level semantic information. Higher-order feature map is obtained by using the higher-order apparent association between objects in the current frame and the previous frame, and object's higher-order feature is described by a similarity matrix. Use hierarchical data association and Hungarian algorithm to obtain the best matching relationship between objects. It can handle identity exchange due to unreliable detection and local association failures. This is called MOT based on high-order appearance feature fusion (MOT-HAFF). Results show that MOT-HAFF is robust to long-term occlusion tracking.

In Figure 22, MOT-HAFF is divided into two stages: feature extraction stage and affinity estimation. In the feature extraction stage, by inputting a pair of video frames I_t and C_t, I_{t-n} and C_{t-n} that do not need to be continuous with the object center point information, after processing by the twin network, the features F_t and F_{t-n} of each object in the video frame are obtained n, and generate $\Psi_{t,\,t-n}$ by connecting and combining F_t and F_{t-n}. In the affinity estimation stage, $\Psi_{t,\,t-n}$ are input to the OCN module that can process feature maps at multiple spatial frequencies and reduce spatial redundancy. After processing the tensor in OCN, it is input into the GRN module consisting of nine convolutional layers. For trajectory association, n similarity matrices are obtained by calculating appearance feature matrix F_p and F_t storage is used to calculate the future affinity matrix. Trajectory set Γ_t is updated by associating the current frame with n previous frames using the calculated affinity matrix.

Unlike the above literature, Guillem and Leal-Taixé [53] are dedicated to learning better features for MOT for most learning-based work and then use these features in combination with a perfect optimization framework. In this paper, the traditional network flow formula is analyzed, and a completely differentiable framework based on message processing networks (MPNs) is defined. By operating directly within the graph domain, you can globally infer a set of detection results and predict the final solution. Therefore, MOT learning does not need to be limited to feature extraction, but can also be applied to the data association step.

A set of input video frames and detection is shown in Figure 23(a). Figure 23(b) shows graph construction. The nodes in the graph represent detection, and all nodes in different frames are connected by an edge. CNN is used to initialize node embedding in the graph, MLP is used to encode geometric information (not shown in the figure) to initialize edge embedding, and information contained in these embedded items is propagated through the entire graph through neural messaging, and a fixed number of iterations are performed (Figure 23(c)). Once this process is terminated, the embedding generated by neural information transfer is used to divide the edge into active (green) and inactive (red) (Figure 23(d)). During training, the predicted cross-entropy loss w.r.t. ground-truth label is calculated. When inferring, a simple rounding scheme is followed to binarize the classification score and obtain the final trajectory (Figure 23(e)).

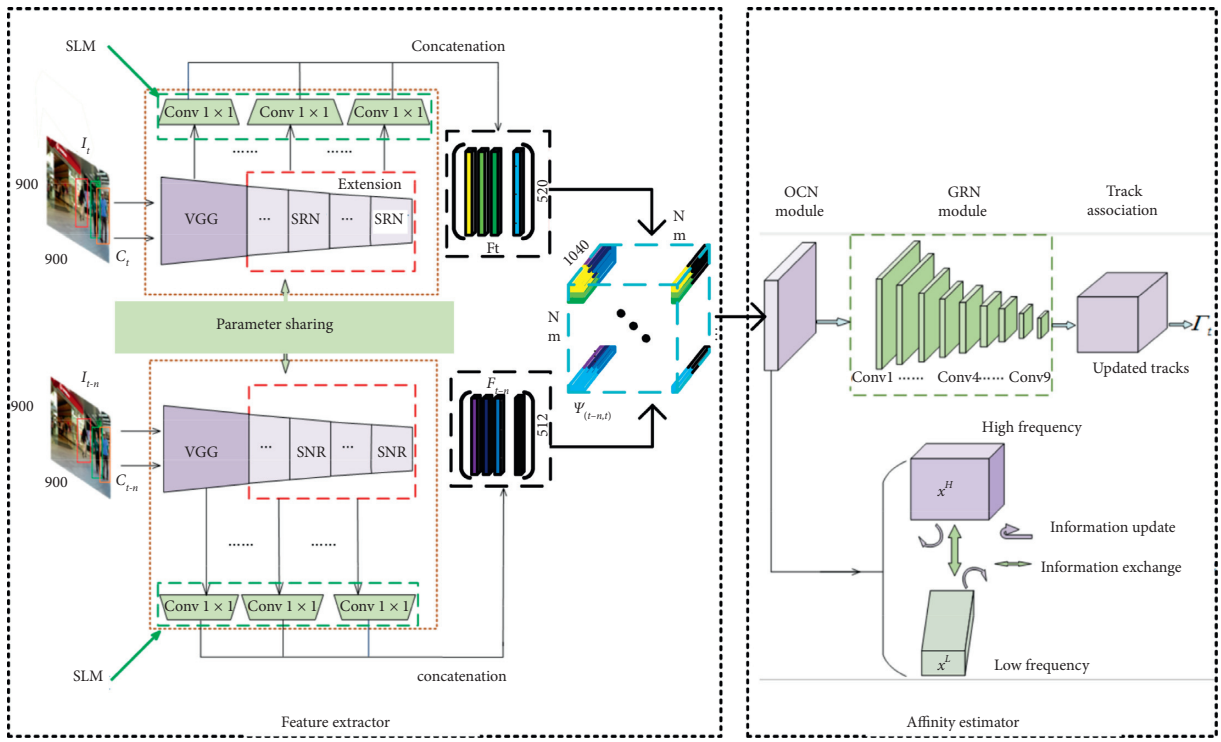

Figure 22: MOT-HAFF framework.

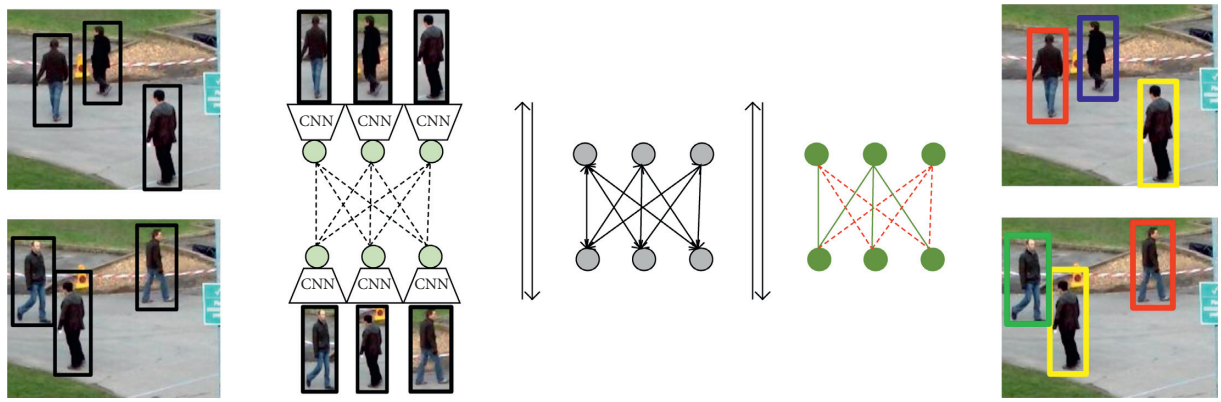

Figure 23: MPN framework. (a) Input. (b) Graph construction + feature encoding. (c) Neural message passing. (d) Edge classification. (e) Output.

For each object of the current frame, the node update part of its message delivery method takes time information into account. That is, for the previous frame and the following frame, the cumulative effect of the edges is calculated separately, and the combination of the links is used instead of the cumulative method.

In Figure 24, the direction of the arrow shows the direction of time at the edge. After performing the edge update, Figure 24(a) shows the starting point, and the intermediate node update embedding is calculated. Figure 24(b) shows the standard node updates in Vanilla MPNs, where the embedding of all neighbors is jointly aggregated. Figure 24(c) shows the proposed update, where the embeddings from past and future frames are aggregated separately and then connected and input into MLP to obtain new node embeddings.

Multiple object tracking accuracy (MOTA) and multiple object tracking precision (MOTP) are two standard and widely used metrics in MOT. It is difficult to directly optimize MOTA and MOTP because these two indicators are nondifferentiable. Xu et al. [54] propose a deep MOT method that can directly optimize MOTA and MOTP. This method simulates the Hungarian algorithm through a bidirectional recursive network. The bidirectional recursive

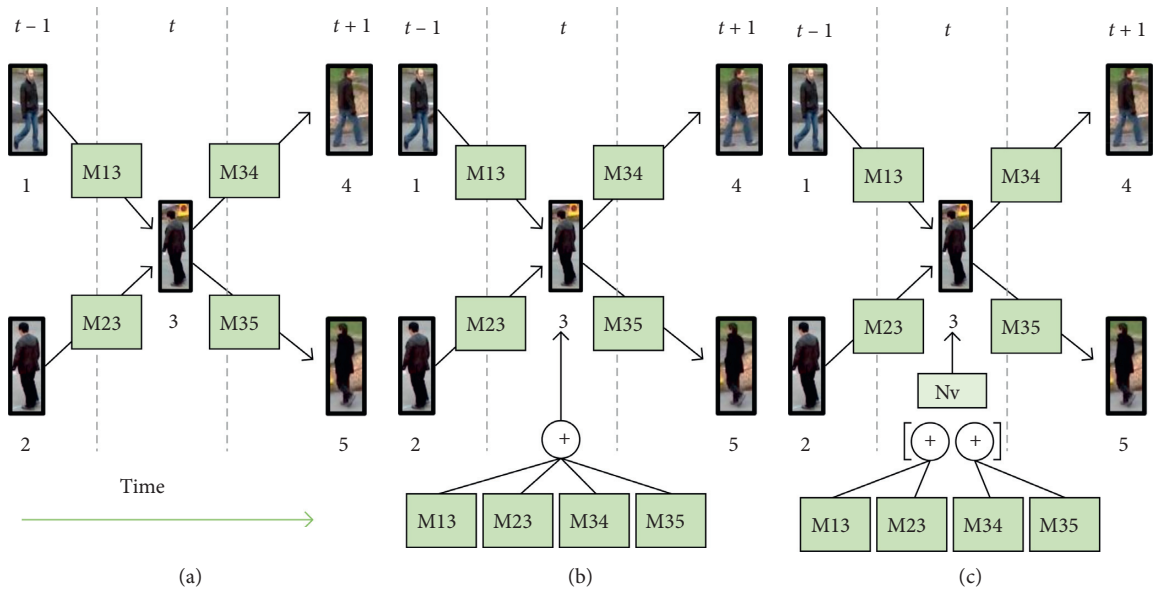

Figure 24: Node updates during messaging. (a) Initial setting. (b) Vanilla node update. (c) Time-aware node update.

network inputs distance matrix from object to hypothesis and outputs the best association from hypothesis to object. The proposed framework is shown in Figure 25.

In Figure 25, taking the example of tracking pedestrians in a video sequence, at time t, several single object trackers are used to track multiple people, providing N_t estimated bounding boxes ($e-\text{bb}$), $X_{t1}, X_{t2}, \ldots, X_{tN}$. These estimates are then compared to the ground-truth bounding boxes ($\text{gt}-\text{bb}$), $O_{t1}, O_{t2}, \ldots, O_{tM}$. The pairwise distance between $\text{e}-\text{bb}$ and $\text{gt}-\text{bb}$ is stored in a distance matrix D_t. The proposed deep Hungarian network (DHN) is then used to estimate the optimal soft-assignment matrix A_t. Finally, expected values of MOTA and MOTP, $\overline{\text{MOTA}}$, and $\overline{\text{MOTP}}$ are computed from D_t and A_t.

The goal of DHN in Figure 25 is to obtain the optimal association matrix A_t from D_t, as shown in Figure 26.

In Figure 26, D_t is a two-dimensional matrix, and two BiRNNs, with different weights and inputs, are sequentially applied in order to receive information from elements in D_t in a row-wise and in a column-wise direction. Firstly, D_t is flattened in the row-wise order and is input into the first BiRNN having hidden units of size h and it outputs a sequence containing $M \times N$ elements. Each element is a vector of size $2 \times h$. Secondly, reshape the output sequence into a tensor, named first-stage hidden representation of size $M \times N$, with a depth of $2 \times h$. After, a column-wise flatten operation is performed before inputting to the second BiRNN, which generates the second-stage hidden representation, size of which is the same as the first-stage representation. At that point, the second-stage hidden representation is flattened and given to three FC layers. They run independently for each of the $M \times N$ vectors of length $2 \times h$, and their size is independent of D_t. It applies the sigmoid function to the output after FC layers. Finally, after reshaping, I_t obtains A_t.

After getting the optimal A_t, it needs to calculate the MOT indicators MOTA and MOTP. Staying in the previous example of tracking pedestrians, if a track is matched to a ground truth (a match means that at time t, a ground-truth identity is assigned to a track. The matching criterion can vary according to applications, commonly, an IoU of two bounding boxes larger than 0.5 is considered as a match), the track is considered as a true positive (TP_t). Otherwise, it is a false positive (FP_t) and a missed ground truth is considered as a false negative (FN_t). For a track marked as TP at time t and at the most recent previous time step, if it is assigned to different ground truth identities, then it counts as identity switch (IDSw_t). The task of deepMOT Loss (DML) in Figure 25 is to calculate the values of FN_t, FP_t, and IDSw_t through A_t and D_t and then obtain MOTA and MOTP. The detailed process is shown in Figure 27.

In Figure 27, by calculating $\overline{\text{FNt}}$, $\overline{\text{FPt}}$, $\overline{\text{TPt}}$, and $\overline{\text{IDSwt}}$, $\overline{\text{MOTAt}}$ and $\overline{\text{MOTPt}}$ are obtained.

The study of Xu et al. [55] is an improved version of [54], and the former has more experiments than the latter. An end-to-end training framework for MOT methods is proposed, which can train the evaluation indicators of MOT. Most of the methods in [55] are derived from [54]. The final experimental results show that after using the deepMOT training framework to train the tracker, the performance of the tracking method can be improved to a certain extent. The method framework is shown in Figure 28.

In Figure 28, MOT training strategy proposed by network structure (DOWN) considers tracking-to-object allocation problem solved by the proposed DHN and approximates the standard MOT loss, rather than the classic training strategy using nondifferentiable HA (UP). The pairwise distance between e-bb and gt-bb is stored in a distance matrix D. The proposed DHN is then used to estimate the optimal soft-assignment matrix $\widetilde{A}$.

Ma et al. [56] proposed deep association network (DAN), learning graph-based training data, which is constructed through the spatiotemporal interaction of objects.

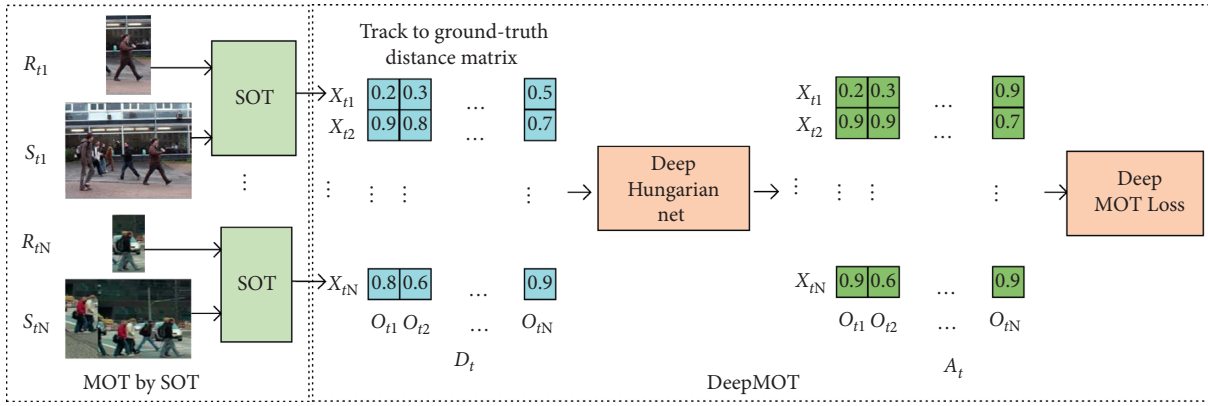

Figure 25: DeepMOT training framework.

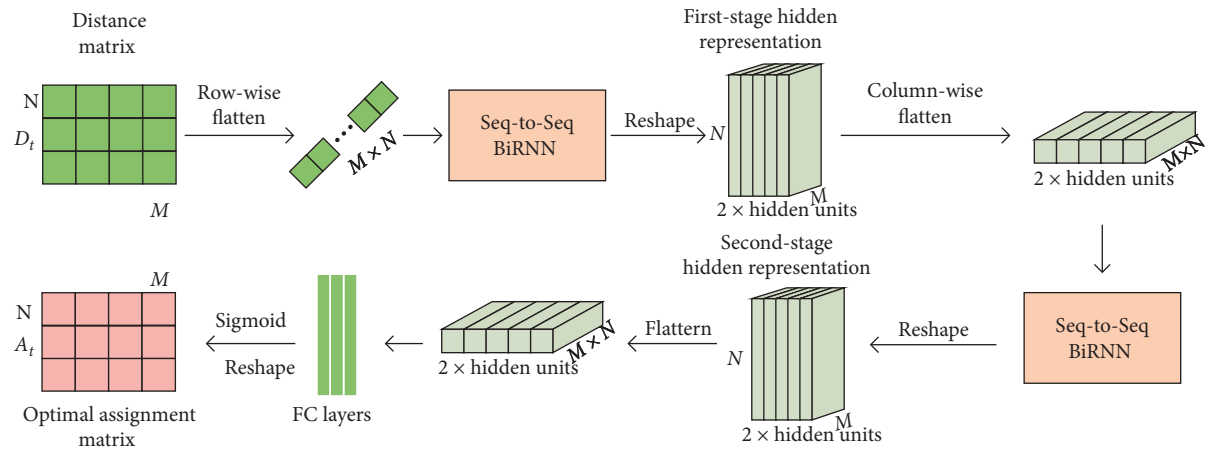

Figure 26: DHN overview.

DAN combines GNN, CNN, and motion encoder (ME). CNNs and MEs extract appearance features and motion features, respectively, and GNN associates the same objects by optimizing the graph structure. The framework of deep association network is shown in Figure 29.

3.3. Wasserstein Metrics Methods. Cosine distance, Euclidean distance, and Mahalanobis distance are all methods to calculate the similarity of two vectors. Cosine distance measures the angle of the space vector. It distinguishes the difference from direction, but is not sensitive to absolute value. Euclidean distance is the most common distance metric. It measures the absolute distance between two points in a multidimensional space. Mahalanobis distance is Euclidean distance that is not affected by dimension, but data premise is as follows: (1) independent distribution of each dimension, and (2) the variance is equal to 1, the mean is 0, and it has been standardized. It can be seen that Mahalanobis distance is also more biased towards the difference in dimensions. Wasserstein distance was first proposed in the last century [57], also known as Earth-Mover (EM) distance, and the expression [58] is as follows:

$$W(P_1, P_2) = \inf_{\gamma \sim \prod(P_1, P_2)} E_{(x,y)\sim y} \|x - y\|. \quad (1)$$

Among them, inf (·) represents the minimum value, Π (*P*1, *P*2) is the set of all possible joint distributions after the combination of *P*1 and *P*2, and *P*1 and *P*2 are the marginal distribution of Π (*P*1, *P*2). (*x*, *y*)~*γ* is sampled from the joint distribution *γ*, and the distance between them $\|x-y\|$ is calculated, so the expected distance value $E_{(x,y)\sim\gamma}[\|x-y\|]$ can also be calculated, and the lower bound $\inf_{\gamma\sim\prod(P_1,P_2)} E_{(x,y)\sim y}$ $\|x - y\|$ of the expected value is found in all possible joint distributions.

Bewley et al. [59] only used the running information of the tracking object in the object association part, which is easy to cause loss of object. Wojke et al. [60] added apparent information of the object on this basis and used cosine distance and Mahalanobis distance to comprehensively calculate the similarity between objects, but there are many false positives for object. Many related papers propose that Wasserstein distance replaces previous Euclidean distance. Mémoli [61] modified the Gromov–Hausdorff distance to produce Gromov–Wasserstein distance, which is more suitable for practical calculations, but retains all the ideal

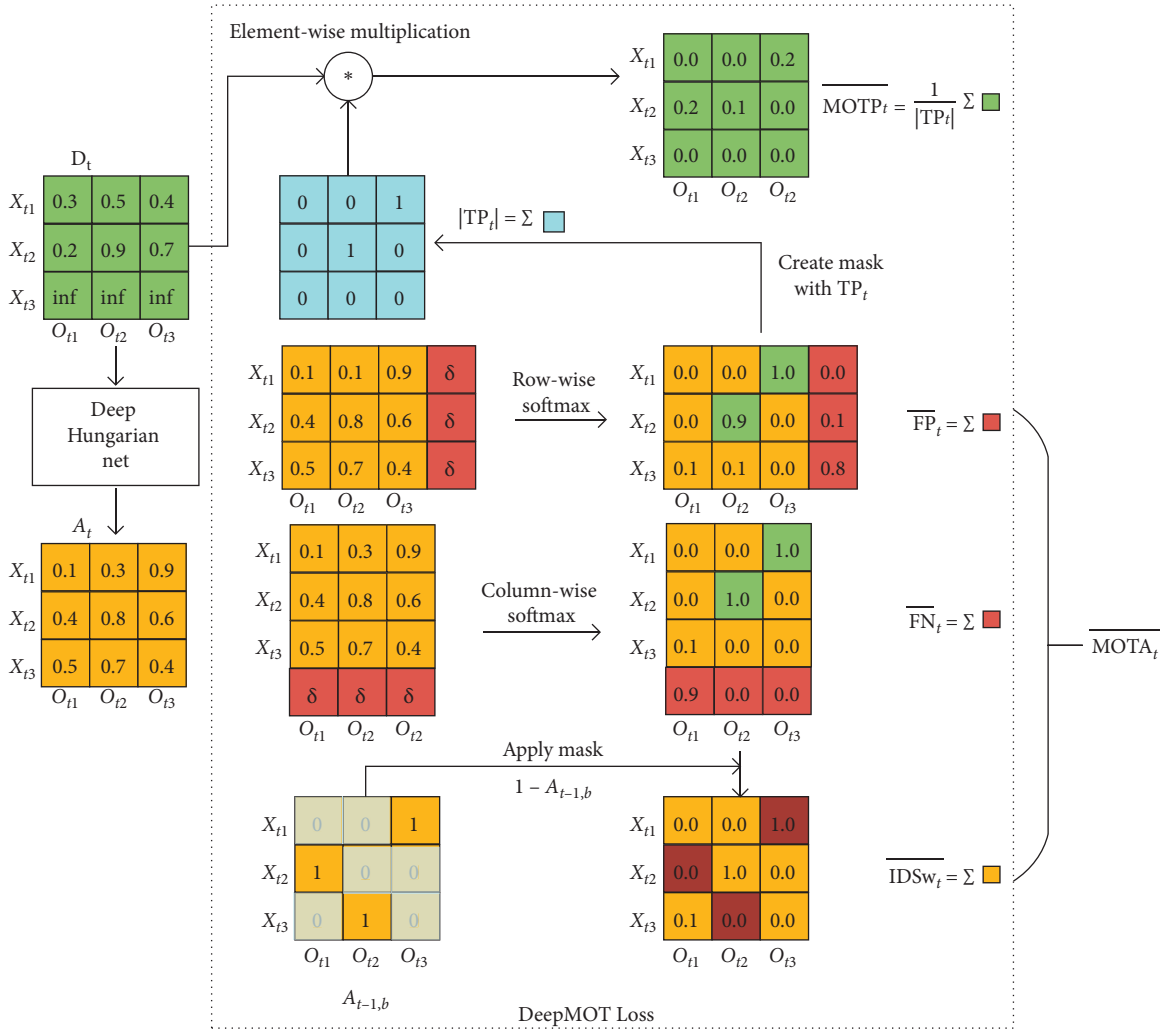

Figure 27: DML example.

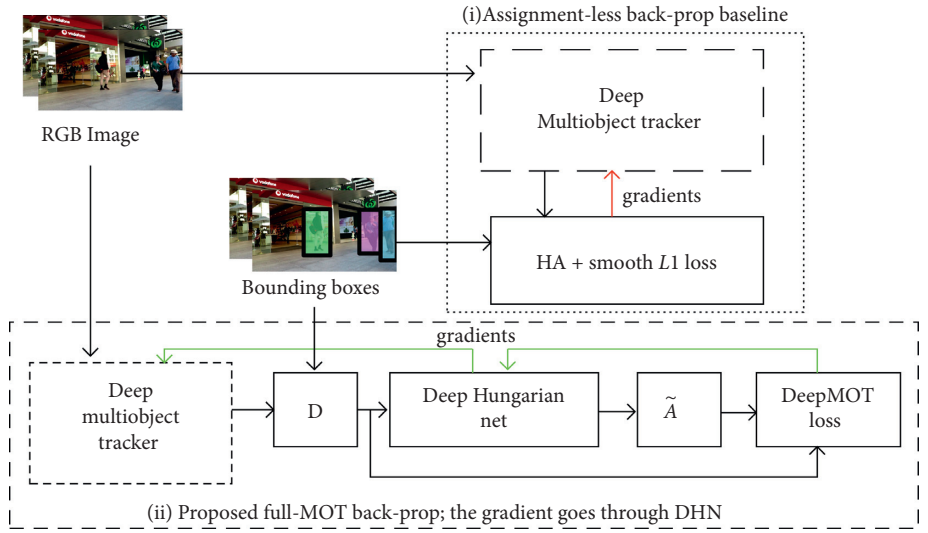

Figure 28: Method of deepMOT framework.

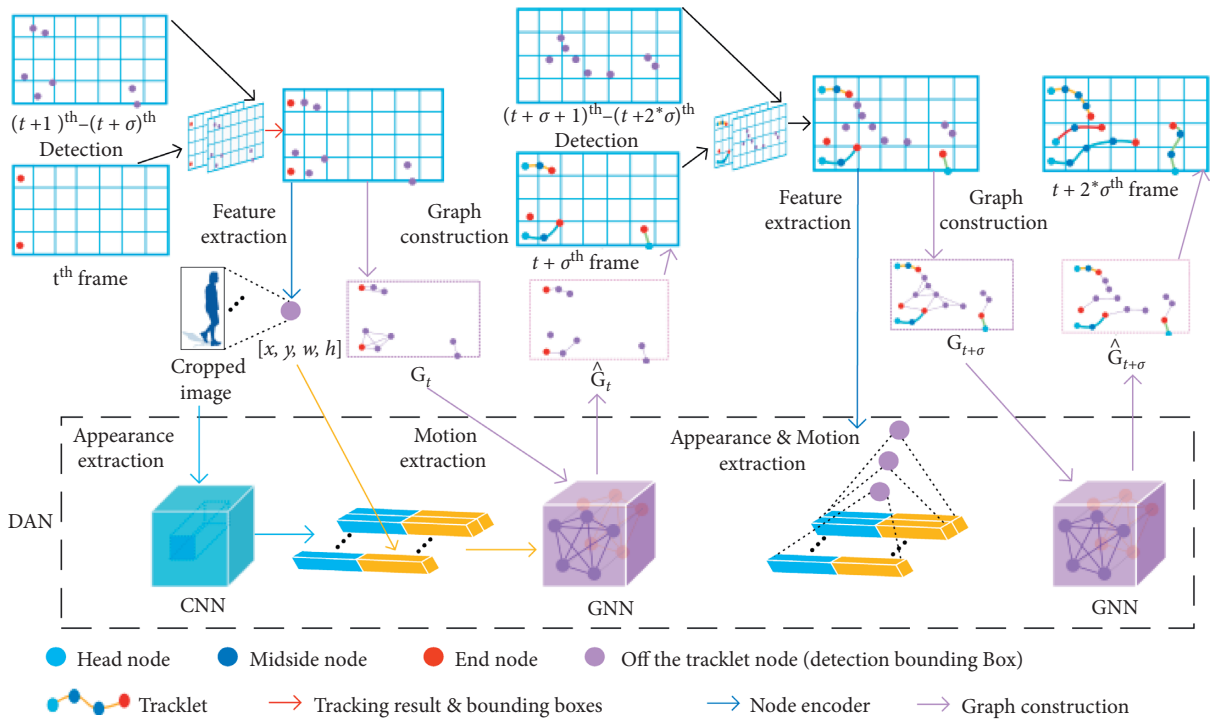

Figure 29: Framework of DAN.

theoretical basis and gives the isomorphic classes in the metric space Strict measure of conjugation. Solomon et al. [62] summarized the previous work and applied Wasserstein distance to many classic graphics problems such as correspondence problems and BRDF. Bonneel et al. [63] combined the color histogram and Wasserstein distance and proposed its application in color space and shape inference. Papadakis and Rabin [64] applied Wasserstein distance to image segmentation problem. In the specific application of vehicle tracking, partial occlusion and similar objects pose a huge challenge. In order to solve the above problem, Zeng et al. [65] proposed a robust MOT method by Wasserstein Association Measurement. The overall network structure is shown in Figure 30. The article combines deep neural network and Wasserstein metric to solve the problem that the traditional method of calculating the Wasserstein metric is too complicated.

In Figure 30, the framework includes two stages: vehicle proposal generation and Wasserstein-based tracklet-detection association. The former uses Faster R-CNN to generate vehicle proposals. The latter expresses each tracklet and detection through TSSC, then calculates the similarity between each tracklet and detection through WD-1, and introduces the Kuhn–Munkres algorithm to optimize the tracklet-detection correlation performance.

Wasserstein metric is currently mostly used in adversarial networks. The better one is Wasserstein GAN network, which can generate high-quality pictures even without using batch standardization. Wasserstein metric is a new direction in video MOT data association, and there is still little relevant literature. However, the accuracy of the final experimental results, Wasserstein metric comprehensively considering characteristics of direction and dimensions, and feasibility of using deep learning to solve Wasserstein distance make data association based on Wasserstein metric have a promising future. And deep learning used in the data association stage of MOT has a great development potential. Therefore, this article describes the preliminary use of deep learning to solve Wasserstein distance as a direction in deep learning-based data association.

Table 2 shows data association methods evaluation and representative work based on deep learning.

4. Benchmark Datasets

Data association is a stage in video MOT. Its fundamental purpose is to improve the performance of object tracking. Therefore, the datasets and evaluation indicators of video MOT are still used. At present, according to the progress of video multiobject development, the datasets can be divided into two stages: classical datasets based on traditional methods and popular datasets based on deep learning. The size of the classic datasets is not large and mainly used to test the performance of the method. Popular datasets are generally larger because deep learning can be said to be a learning process based on datasets. The larger the size of the datasets, the greater its advantages and the higher the learning performance. With the development of multiobject video tracking, the demand for datasets will continue to increase. As a result, domestic laboratories have launched MOT datasets for

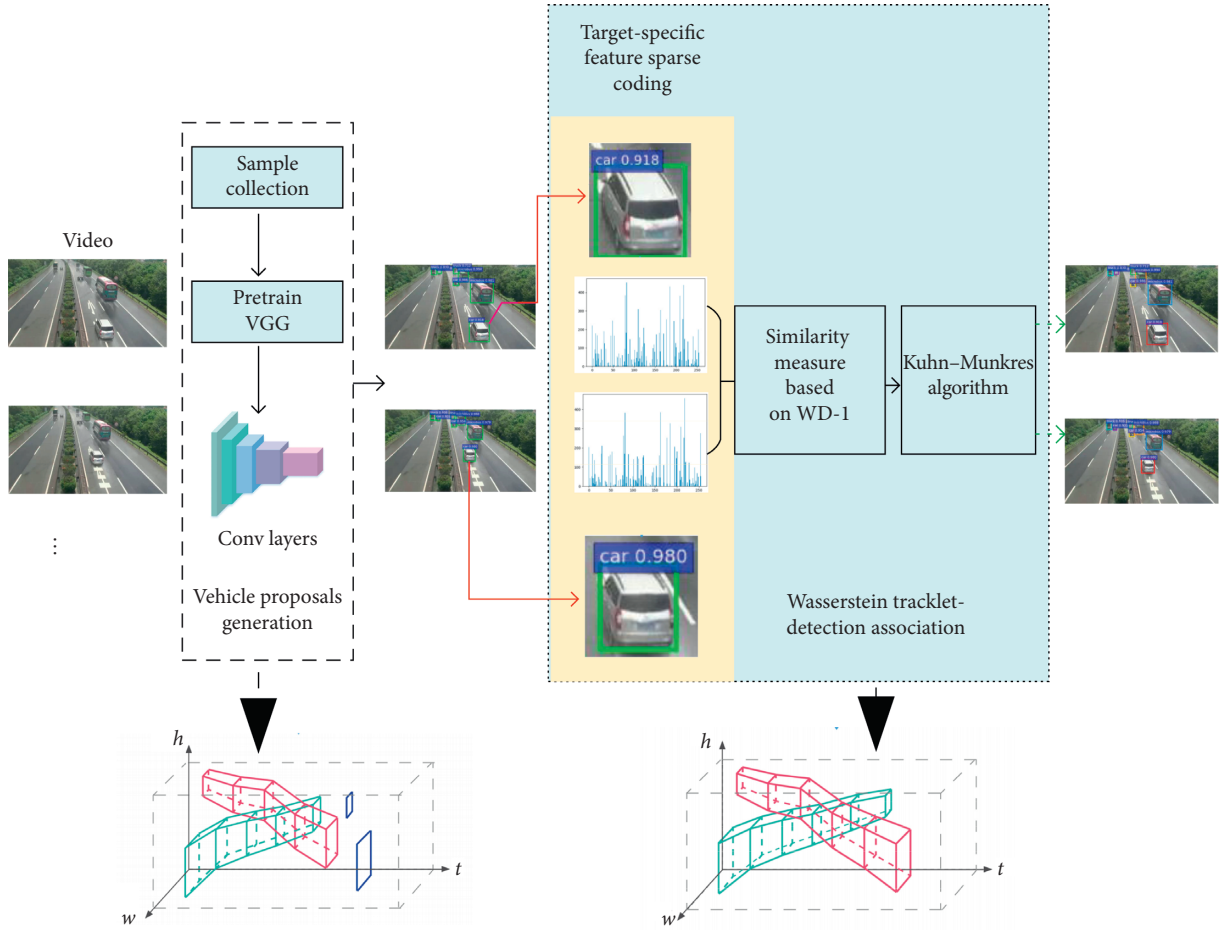

Figure 30: Overall network framework.

various scenes and different tracking objects through video collection and postproduction. At present, large-scale datasets are mainly for pedestrians and vehicles. The main scenes are shopping malls, pedestrian streets, sports, highways, and traffic intersections. Table 3 shows the details. The most commonly used [70, 72, 73] datasets will also be introduced later, and the evaluation indicators for MOT will also be briefly introduced.

4.1. MOT17. MOT17 dataset is contributed by MOT challenge, which is a very influential game in the field of MOT. Since 2005, many algorithms that participate in the competition have been published in the top three computer conferences—CVPR, ICCV, and ECCV. So far, the open datasets are MOT15 to MOT20 series, of which the most commonly used is the MOT17 dataset, which will be introduced in this section.

MOT17 has a total of 42 video sequences, 21 of which are used for training and 21 for testing; most of the video resolution is 1920×1084 pixels, within one minute, and there are more pedestrians and pedestrian occlusions; each video sequence is captured by stationary or moving cameras in a constrained environment. It uses all the video sequences of MOT16, but has a more accurate ground truth, and each sequence has 3 sets of detection: DPM, Faster R-CNN, and SDP. Part of the training set based on SDP detection is shown in Figure 31.

4.2. KITTI_Tracking. The KITTI dataset was cofounded by Karlsruhe Institute of Technology in Germany and Toyota's American Institute of Technology in 2011. It is currently the world's largest evaluation dataset for computer vision methods in autonomous driving scenarios. The original data have a total of 180 GB of data, including real image data collected from urban, rural, and highway scenes. Each image can have up to 15 vehicles and 30 pedestrians, as well as various degrees of occlusion and truncation. The entire dataset is sampled at a frequency of 10 Hz and consists of 389 pairs of stereo images and optical flow maps, a 39.2 km visual ranging sequence, and images of over 200 k 3D labeled objects. The KITTI data acquisition platform includes 2 grayscale cameras, 2 color cameras, a 3D lidar, 4 optical lenses, and 1 GPS navigation system. The collection platform is shown in Figure 32.

Table 2: Summary and evaluation of previous studies on deep learning.

Author/organization	Years	Technical characteristics	Dataset	MOTA↑(%)	MOTP↑(%)	IDs↓
Ji Zhu/Shanghai Jiao Tong University	2018	Unified framework, including SOT and data association [48]	MOT16	46.1	73.8	532
			MOT17	48.2	75.9	2194
Peng Chu/Temple University	2019	Feature extraction, affinity estimation, and multidimensional distribution are integrated in a single network [49]	MOT2015	40.6	71.1	778
			MOT2017	52.0	76.5	3072
			KITTI-Car	77.1	77.8	123
			UA-DETRAC	19.8	36.7	617
Weitao Feng/SenseTime Group Limited	2019	SOT captures short-term cues, ReID extracts long-term cues [50]	MOT2016	49.2	74.0	606
			MOT2017	52.7	76.2	2167
			MOT2016p	69.6	78.5	768
ShiJie Sun/University of Western Australia	2019	End-to-end data association is divided into two stages: feature extraction and affinity estimation [51]	MOT2017	52.4	52.4	8431
			MOT2015	38.3	38.3	1648.08
			UA-DETRAC	20.0	20.0	518.2
Guang Han/Nanjing University of Posts and Telecommunications	2019	End-to-end data association, extract features that merge high-level and low-level semantic information [52]	MOT2015	38.41	72.03	1547
			MOT2017	48.8	76.9	5601
Brasó/Technical University of Munich	2019	End-to-end data association, combining graphs and minimizing costs [53]	2DMOT2015	48.3	—	504
			MOT2016	55.9	—	431
			MOT2017	55.7	—	1433
Yihong Xu/Univ. Grenoble Alpes	2019	Propose a differentiable tracker that directly optimizes MOTA and MOTP [54]	MOT2017	44.3	76.0	4861
	2019	End-to-end data association, a deep Hungarian network is proposed for microcomputable MOTA and MOTP [55]	MOT2017	53.7	77.2	1947
			MOT2016	54.8	77.5	645
Cong Ma/Peking University	2019	End-to-end; combines CNN, ME, and GNN [56]	DukeMTMC	86.7	—	928
			MOT16	48.6	—	594
Yanjie Zeng/South China University of Technology	2020	Multivehicle tracking with Wasserstein association metric method [65]	VECHSV	72.4	85.3	670
			DETRAC	68.5	86.5	151

Table 3: Classification of video MOT datasets.

Name	Number of goals	Number of boxes	Camera mode	Scale	Tracking category	Features	Scenes
TownCenter [66]	16	—	Stable	4500	Pedestrian	Simple; annotation is complete; clear; blocked frequently	1
PETS09-S2L1 [67]	8	—	Stable	795	Pedestrian	Sparse crowd; high-speed nonlinear mode; blocked frequently	1
TUD-Stadtmitte [68]	—	—	Stable	179	Pedestrian	Low angle of view; severe mutual occlusion; complete occlusion	1
Parking Lot [69]	14	—	Stable	1000	Pedestrian	Parking lot; mutual blocking is more serious than TUD	2
PETS09-S2L2 [39]	—	—	Stable	168	Pedestrian	Medium-density crowds; high speed; blocked severely	1
MOT16 [56]	517	110407	Mobile/stable	5316	Pedestrian	Many scenes; comprehensive data; large amount of data	7
MOT17 [70]	1638	564228	Mobile/stable	15948	Pedestrian	Many scenes; more comprehensive than MOT16, magnitude of the data is larger	7
MOT20 [70]	2332	1336920	Stable	8931	Pedestrian	Wide scene at night; high crowd density	3
UA-DETRAC [71]	8250	1210000	Stable	140000	Vehicle	Canon EOS 550D camera records at 25 fps; rich scenes; large data volume	24
KITTI [72]	—	330000	Mobile	180 GB	Pedestrian/vehicle	Vehicle-mounted camera; up to 15 vehicles and 30 pedestrians in each image; various degrees of occlusion and truncation	>3
MOTs [73]	977	65213	Mobile	10870	Pedestrian/vehicle	Pixel-level relabeling on the KITTI_Tracking and MOTS challenge	>10
DukeMTMC [56]	1404	65213	Stable	36411	Pedestrian	Large HD video dataset, typical MOT scenario	8

Training set

Sample	Name	FPS	Resolution	Length	Tracks	Boxes	Density	Description
	MOT17-13-SDP	25	1920 × 1080	750 (00:30)	110	11642	15.5	Filmed from a bus on a busy intersection
	MOT17-11-SDP	30	1920 × 1080	900 (00:30)	75	9436	10.5	Forward moving camera in a busy shopping mall
	MOT17-10-SDP	30	1920 × 1080	654 (00:22)	57	12839	19.6	A pedestrian scene filmed at night by a moving camera
	MOT17-09-SDP	30	1920 × 1080	525 (00:18)	26	5325	10.1	A pedestrian street scene filmed from a low angle
	MOT17-05-SDP	14	640 × 480	827 (01:00)	133	6917	8.3	Street scene from a moving platform
	MOT17-04-SDP	30	1920 × 1080	1050	83	47557	45.3	Pedestrian street at night, elevated viewpoint
	MOT17-02-SDP	30	1920 × 1080	600 (00:20)	62	18581	31.0	People walking around a large square

Figure 31: Partial training set based on SDP detection.

Grayscale cameras (a) Color cameras (b) 3D lidar (c) Optical lenses (d)

Figure 32: Data acquisition platform.

Sample of KITTI dataset is shown in Figure 33, showing the diversity of the dataset. Branch graph of KITTI data is shown in Figure 34.

4.3. MOTs. In 2019, computer vision laboratory of Aachen University of Technology in Germany proposed MOT and segmentation network Track R-CNN in the paper and published MOTS dataset with only two types of objects, car and pedestrian. This dataset relabels KITTI_Tracking and MOT_Challenge datasets. KITTI_Tracking dataset is annotated as a 2D box. In order to meet MOTS task, KITTI_MOTS dataset adds an instance segmentation Mask label and divides 20 sequences in the original KITTI dataset into a training set and a test set. MOTS_Challenge dataset uses the same method, but there are only four sequences, which are densely populated and heavily occluded, and belong to the challenge-level task data in MOT.

The visualization of KITTI_Tracking dataset and MOTS_Challenge dataset is shown in Figure 35 and Figure 36, respectively.

4.4. Evaluation Criteria for MOT. In the past, there was no universal evaluation standard for MOT. Since 2008, Keni and Rainer [74] introduced the calculation process of two comprehensive indicators MOTA and MOTP. These two indicators have advantages and disadvantages, but as a comprehensive indicator, it is still widely used in academia so far.

(1) MOTA: accuracy of MOT is reflected in the determination of the number of objects and the accuracy of the relevant attributes of the object. It is used to count the accumulation of errors in tracking, including FP, FN, and IDSw. In general, the larger the value, the better the tracking effect:

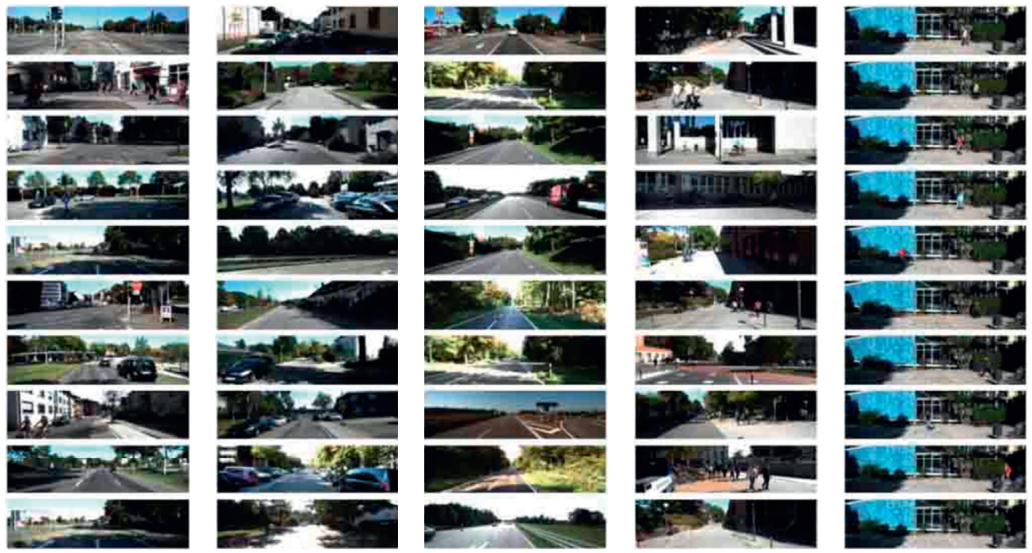

Figure 33: Sample of KITTI dataset.

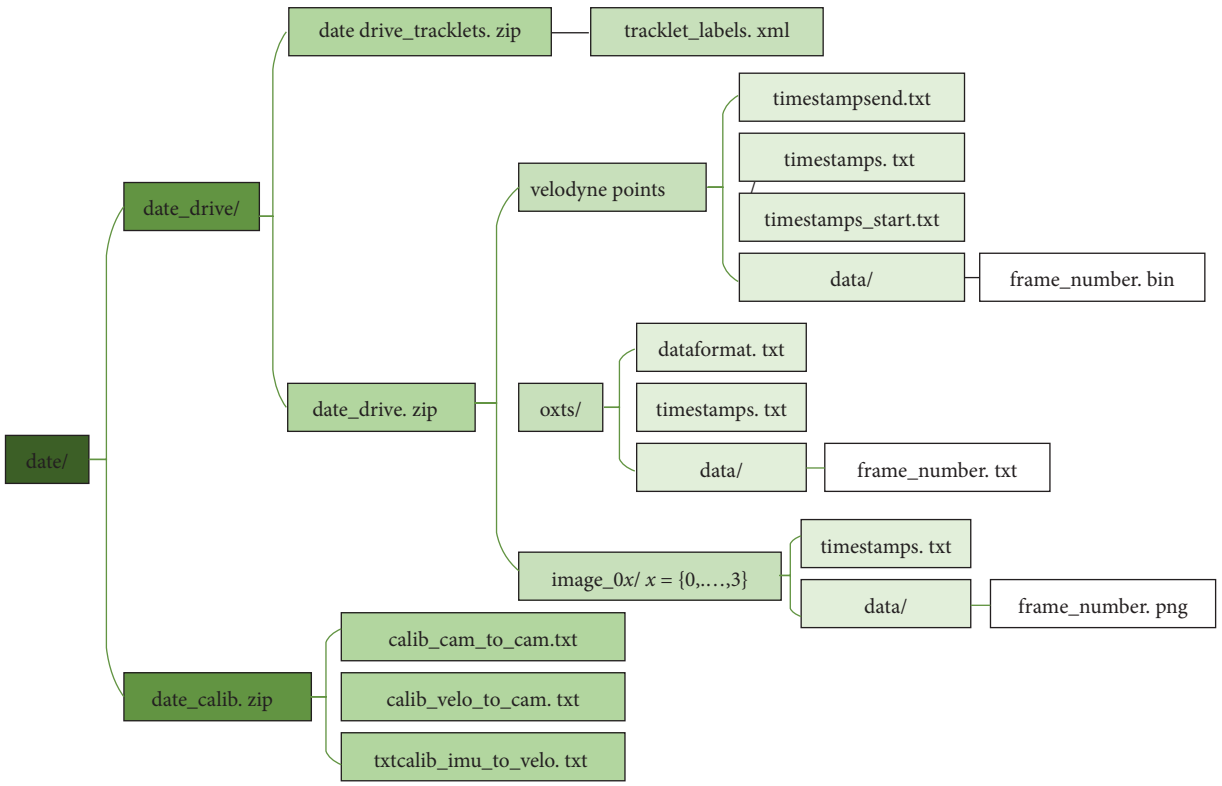

Figure 34: Branch graph of KITTI data.

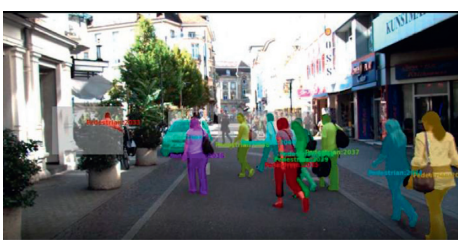

Figure 35: Example of KITTI_Tracking dataset visualization.

Figure 36: Example of MOTS_Challenge dataset visualization.

$$\mathrm{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}. \quad (2)$$

m_t: FP, the missing number, that is the object does not match its assumed position in the t^{th} frame.

f_{pt}: FN, the number of misjudgments, that is the hypothetical position given in the t^{th} frame does not match the tracking object.

mme_t: IDSw, the number of mismatches, that is the number of times ID switching occurs in the tracking object in the t^{th} frame, which mostly occurs in the case of occlusion.

g_t: GT, the number of ground-truth objects, and the total number of GTs in the entire video.

(2) MOTP: accuracy of MOT is used to measure whether the object position is accurate; generally speaking, the larger the value, the better the tracking effect:

$$\mathrm{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}. \quad (3)$$

where c_t represents the number of matches between the t^{th} frame object and the hypothesis and d_t^i represents the distance between the t^{th} frame object and its paired hypothesis location, that is, the matching error.

In addition to the above commonly used main evaluation standards, there are the following standards:

(1) MT (Mostly Tracked): a track can be considered as MT if it is tracked to more than 80%. No matter how the ID changes on this trajectory (for example, it changes during prediction), as long as this trajectory accounts for more than 80% of the actual trajectory, it can be considered as MT. The larger the value, the better the tracking effect.

(2) ML (Mostly Lost Tracking): if a track is tracked below 20%, it is considered ML; the smaller the value, the better the tracking performance.

(3) PT (Partially Tracked): in addition to MT and ML, others are considered to be PT; the smaller the value, the better the tracking performance.

(4) FM (Fragmentation): the number of times the real trajectory is interrupted; the smaller the value, the better the tracking performance.

5. Summary and Outlook

5.1. Challenges in Data Association Research. Video MOT has a wide range of application scenarios in the fields of driverless, intelligent monitoring, and so on. The importance of data association in MOT is self-evident, which plays a decisive role in MOT development. In recent years, with the application of deep learning in data association, performance and robustness of object association have been greatly improved, but there is still a lot of room for improvement from real application in real scenes. At present, there are still some challenges in data association:

(1) Occlusion problem: when the objects are associated with each other, or between objects or trajectories, especially in the case of crowding, occlusion by objects themselves or other objects will lead to wrong or even failed association, and object ID may be exchanged or even the tracked object may be lost.

(2) Similar interference in class: before calculating the similarity degree of objects between frames, there needs to extract features of the object first. Then, it may happen that the extracted object feature information is not rich enough, which makes it impossible to distinguish similar objects in the class, and finally leads to the association between wrong objects.

(3) Object missing, not aligned: when twin network is used to calculate the similarity between two frames of objects, because of the absence and misalignment of objects, it will lead to object association failure even if the similarity of the same object cannot meet threshold requirements through simple feature superposition and fusion.

The above is just the main problem of data association. There are still many problems to be solved in detail, including excessive parameters and calculation. Real time, accuracy, and generality will always be the direction of data association. In addition, how to use the existing technology to make real life more intelligent, convenient, and low-cost, high-performance operation requirements is also an urgent problem to be solved.

5.2. Outlook. This paper introduces traditional data association and data association based on deep learning, in which data association based on deep learning is the focus. Traditional data association calculation is small and can be run on CPU, but the performance is not good, and most of them need human intervention, which is time-consuming and labor-consuming. Data association based on deep learning has better performance, and it seldom needs human intervention. The network will complete most of the work

independently, but it needs a large amount of data and a large amount of calculation, and the training process needs to use GPU. With the upgrading of computer hardware, accumulation of related datasets, and improvement of methods, data association based on deep learning ushers in greater prospects for development, and video MOT will also make new breakthroughs.

(1) Creating a good dataset and evaluation criteria can help improve association performance. Although there are already related datasets at this stage, there is still no better choice for a general video tracking datasets of real scene. In addition, the problem of datasets annotation is still a time-consuming and labor-consuming project. Although there are some semiautomatic or automatic annotation tools, the effect is not very good.

(2) When dataset conditions are satisfied, a good method determines the performance of data association. Combining the advantages of traditional data association and deep learning is a better development direction. At present, although there are related research directions, the way of combination is still superficial and the performance is not high. Therefore, this is also one of the future research directions.

(3) Data association can be divided into two stages: feature extraction and affinity estimation. With the development of deep learning, the performance of the two stages has been greatly improved, but it is easy to make mistakes when encountering occlusion, missing, misalignment, and so on. For the former stage, with deepening of the network, how to retain more useful details and build a feature model based on 3D level and feature information including the object multiattitude is a problem worthy of deep thought. For the latter stage, there is still a long way to go for a more perfect combination of spatiotemporal attention mechanism, motion model, and appearance model.

(4) In object association, embedding SOT module to deal with occlusion, missing, and misaligned features is a promising direction. SOT module can decompose complex problems, and single object has been proved to be effective in dealing with occlusion problems.

(5) With hardware upgrading, computing speed will be improved, which speeds up the new method to show its effect, support more complex methods, shorten achievement transformation time, and promote the rapid improvement of data association performance, that is, significant improvement of MOT performance.

In the future, with the development of deep learning and improvement of the quality and quantity of datasets, data association will have a huge improvement and then shine a dazzling light in driverless and intelligent monitoring, making the whole society more intelligent.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (no. 61702295), Shandong Province Natural Science Foundation (Grant no. ZR2017BF023), Shandong Province Postdoctoral Innovation Project (Grant no. 201703032), and the Opening Foundation of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University (Grant no. MJUKF-IPIC202009).

References

[1] X. Wang, L. Wan, M. Huang, C. Shen, Z. Han, and T. Zhu, "Low-complexity channel estimation for circular and non-circular signals in virtual mimo vehicle communication systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 3916–3928, 2020.

[2] F. Wen and J. Shi, "Fast direction finding for bistatic emvs-mimo radar without pairing," *Signal Processing*, vol. 173, Article ID 107512, 2020.

[3] H. Wang, L. W. Xu, Z. Q. Yan, and T. Aaron Gulliver, "Low complexity mimo-fbmc sparse channel parameter estimation for industrial big data communications," *IEEE Transactions on Industrial Informatics*, 2020.

[4] L. Xu, J. Wang, H. Wang et al., "BP neural network-based abep performance prediction for mobile internet of things communication systems," *Neural Computing and Applications*, pp. 1–17, 2019.

[5] T. Liu, F. Wen, L. Zhang, and K. Wang, "Off-grid doa estimation for colocated mimo radar via reduced-complexity sparse bayesian learning," *IEEE Access*, vol. 7, pp. 99907–99916, 2019.

[6] W. Zheng, X. Zhang, Y. Wang, M. Zhou, and Q. Wu, "Extended coprime array configuration generating large-scale antenna co-array in massive mimo system," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7841–7853, 2019.

[7] L. Xu, T. Quan, J. Wang et al., "GR and bp neural network-based performance prediction of dual-antenna mobile communication networks," *Computer Networks*, vol. 172, no. 5, pp. 1–10, 2020.

[8] X. Zhang, L. Xu, L. Xu, and D. Xu, "Direction of departure (dod) and direction of arrival (doa) estimation in mimo radar with reduced-dimension music," *IEEE Communications Letters*, vol. 14, no. 12, pp. 1161–1163, 2010.

[9] F. Wen, Z. Zhang, K. Wang, G. Sheng, and G. Zhang, "Angle estimation and mutual coupling self-calibration for ULA-based bistatic MIMO radar," *Signal Processing*, vol. 144, pp. 61–67, 2018.

[10] L. Xu, X. Yu, H. Wang et al., "Physical layer security performance of mobile vehicular networks," *Mobile Networks and Applications*, vol. 25, no. 2, pp. 643–649, 2020.

[11] R. A. Singer and J. J. Stein, "An optimal tracking filter for processing sensor data of imprecisely determined origin in surveillance systems," in *Proceedings of the IEEE Conference on Decision and Control*, pp. 171–175, Miami Beach, FL, USA, December 1971.

[12] S. S. Blackman, "Multiple hypothesis tracking for multiple target tracking," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 1, pp. 5–18, 2004.

[13] G. Zhai, H. Meng, Z. Zhong et al., "A multiple hypothesis tracking method for extended target tracking," in *Proceedings of the International Conference on Electrical and Control Engineering*, pp. 109–112, Wuhan, China, June 2010.

[14] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Systems*, vol. 29, no. 6, pp. 82–100, 2009.

[15] L. Svensson, D. Svensson, M. Guerriero, and P. Willett, "Set JPDA filter for multitarget tracking," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4677–4691, 2011.

[16] R. Streit, "JPDA intensity filter for tracking multiple extended objects in clutter," in *Proceedings of the 19th International Conference on Information Fusion*, pp. 1477–1484, Heidelberg, Germany, July 2016.

[17] A. Çakıroğlu, "Tracking variable number of targets with joint probabilistic data association filter," in *Proceedings of the 24th Signal Processing and Communication Application Conference*, pp. 2017–2020, Ankara, Trukey, 2016.

[18] Y. Zhu, J. Wang, and S. Liang, "Efficient joint probabilistic data association filter based on kullback-leibler divergence for multi-target tracking," *IET Radar, Sonar & Navigation*, vol. 11, no. 10, pp. 1540–1548, 2017.

[19] Y. Xie, H. W. Kim, H. J. Kim et al., "Reduction of computational load for implementing iJIPDA filter," in *Proceedings of the International Conference on Information Fusion*, pp. 1–6, Xi'an, China, July 2017.

[20] S. He, H. Shin, and A. Tsourdos, "Joint probabilistic data association filter with unknown detection probability and clutter rate," in *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 559–564, Daegu, South Korea, November 2017.

[21] Y. Xia, K. Granstrcom, L. Svensson et al., "Performance evaluation of multi-bernoulli conjugate priors for multi-target filtering," in *Proceedings of the 20th International Conference on Information Fusion*, pp. 1–8, Xi'an, China, July 2017.

[22] R. Streit, "Analytic combinatorics and labeling in high level fusion and multihypothesis tracking," in *Proceedings of the 21st International Conference on Information Fusion*, pp. 1–5, Cambridge, UK, July 2018.

[23] F. Meyer and M. Z. Win, "Data association for tracking extended targets," in *Proceedings of the IEEE Military Communications Conference*, pp. 337–342, Norfolk, VA, USA, November 2019.

[24] Y. Zhu, J. Wang, S. Liang, and J. Wang, "Covariance control joint integrated probabilistic data association filter for multi-target tracking," *IET Radar, Sonar & Navigation*, vol. 13, no. 4, pp. 584–592, 2019.

[25] S. Krishnaswamy and M. Kumar, "A window-based tensor decomposition approach to data-association for multitarget tracking," in *Proceedings of the American Control Conference*, pp. 5148–5153, Philadelphia, PA, USA, July 2019.

[26] T. Gallion, C. McKinney, and S. Chakravarty, "Pedestrian tracker utilizing sensor fusion and a gaussian-mixture probability hypothesis density filter," in *Proceedings of the SoutheastCon*, pp. 1–6, Huntsville, AL, USA, April 2019.

[27] S. Liang, Y. Zhu, H. Li et al., "Nearest-neighbour joint probabilistic data association filter based on random finite set," in *Proceedings of the International Conference on Control, Automation and Information Sciences*, pp. 1–6, Chengdu, China, October 2019.

[28] S. A. Memon, M. Kim, M. Shin, J. Daudpoto, D. M. Pathan, and H. Son, "Extended smoothing joint data association for multi-target tracking in cluttered environments," *IET Radar, Sonar & Navigation*, vol. 14, no. 4, pp. 564–571, 2020.

[29] S. He, H.-S. Shin, and A. Tsourdos, "Distributed joint probabilistic data association filter with hybrid fusion strategy," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 1, pp. 286–300, 2020.

[30] C. Kim, F. Li, A. Ciptadi et al., "Multiple hypothesis tracking revisited," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4696–4704, Santiago, Chile, December 2015.

[31] S.-I. Oh and H.-B. Kang, "Multiple objects fusion tracker using a matching network for adaptively represented instance pairs," *Sensors*, vol. 17, no. 4, p. 883, 2017.

[32] E. Baser, V. Balasubramanian, P. Bhattacharyya et al., "Fantrack: 3D multi-object tracking with feature association network," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 1426–1433, Paris, France, June 2019.

[33] H. Wang and S. K. Nguang, "Multi-target video tracking based on improved data association and mixed kalman," *IEEE Sensors Journal*, vol. 16, no. 21, pp. 7693–7704, 2016.

[34] S. Xu, A. Savvaris, S. He et al., "Real-time implementation of YOLO + JPDA for small scale UAV multiple object tracking," in *Proceedings of the International Conference on Unmanned Aircraft Systems*, pp. 1336–1341, Dallas, TX, USA, June 2018.

[35] B. Bićanić, M. Oršić, I. Marković et al., "Pedestrian tracking by probabilistic data association and correspondence embeddings," in *Proceedings of the 22th International Conference on Information Fusion*, pp. 1–6, Cambridge, UK, 2019.

[36] L. Leal-Taixé, A. Milan, I. Reid et al., "MOT challenge 2015: towards a benchmark for multi-target tracking," 2015.

[37] T. Kikuchi, "Visual object tracking by moving horizon estimation with probabilistic data association," in *Proceedings of the IEEE/SICE International Symposium on System Integration*, pp. 115–120, Honolulu, HI, USA, January 2020.

[38] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1265–1272, Colorado Springs, CO, USA, June 2011.

[39] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 58–72, 2014.

[40] A. Milan, K. Schindler, and S. Roth, "Multi-target tracking by discrete-continuous energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2054–2068, 2016.

[41] G. Zuo, T. Du, and L. Ma, "Dynamic target tracking based on corner enhancement with Markov decision process," *The Journal of Engineering*, vol. 2018, no. 16, pp. 1617–1622, 2018.

[42] J. Shen, Z. Liang, J. Liu et al., "Multiobject tracking by submodular optimization," *IEEE Transactions on Cybernetics*, vol. 49, no. 6, pp. 1990–2001, 2019.

[43] S. Zhu, C. Sun, and Z. Shi, "Multi-target tracking via hierarchical association learning," *Neurocomputing*, vol. 208, no. 1, pp. 365–372, 2018.

[44] T. Ali, L. Liu, H. Qi et al., "Addressing ambiguity in multi-target tracking by hierarchical strategy," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 3635–3639, Beijing, China, September 2017.

[45] J. Liu, X. Cao, Y. Li et al., "Online multi-object tracking using hierarchical constraints for complex scenarios," *IEEE*

Transactions on Intelligent Transportation Systems, vol. 19, no. 1, pp. 151–161, 2018.

[46] H. Liu, F. Chang, and C. Liu, "Multi-target tracking with hierarchical data association using main-parts and spatial-temporal feature models," *Multimedia Tools and Applications*, vol. 78, no. 20, pp. 1–21, 2019.

[47] W. Lu, Z. Zhou, L. Zhang et al., "Multi-target tracking by non-linear motion patterns based on hierarchical network flows," *Multimedia Systems*, vol. 25, no. 4, pp. 383–394, 2019.

[48] J. Zhu, H. Yang, N. Liu et al., "Online multi-object tracking with dual matching attention networks," in *Computer Vision—ECCV 2018*, pp. 366–382, Springer, Berlin, Germany, 2018.

[49] P. Chu and H. Ling, "Famnet: joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6172–6181, Seoul, South Korea, 2019.

[50] W. Feng, Z. Hu, W. Wu et al., "Multi-object tracking with multiple cues and switcher-aware classification," 2019, https://arxiv.org/abs/1901.06129.

[51] S.J. Sun, N. Akhtar, H.S. Song et al., "Deep Affinity network for multiple object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[52] G. Han, Y. Gao, and N. Sun, "Multi-target tracking based on high-order appearance feature fusion," *IEEE Access*, vol. 7, pp. 173393–173406, 2019.

[53] B. Guillem and L. Leal-Taixé, "Learning a neural solver for multiple object tracking," 2019, https://arxiv.org/abs/1912.07515.

[54] Y. Xu, Y. Ban, X. Alameda-Pineda et al., *Deepmot: a Differentiable Framework for Training Multiple Object Trackers*, DeepAI, London, UK, 2019.

[55] Y. Xu, Y. Ban, X. Alameda-Pineda et al., "How to train your deep multi-object tracker," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[56] C. Ma, L. Yuan, Y. Fan et al., "Deep association: end-to-end graph-based learning for multiple object tracking with conv-graph neural network," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 253–261, Ottawa, Canada, June 2019.

[57] I. Olkin and F. Pukelsheim, "The distance between two random vectors with given dispersion matrices," *Linear Algebra and Its Applications*, vol. 48, pp. 257–263, 1982.

[58] M. Arjovsky, S. Chintala, L. Bottou et al., "Wasserstein GAN," 2017, https://arxiv.org/abs/1701.07875.

[59] L. Bewley, "Simple online and real-time tracking," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 3464–3468, Phoenix, AZ, USA, September 2016.

[60] N. Wojke, A. Bewley, and D. Paulus, "Simple online and real-time tracking with a deep association," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 3645–3649, Beijing, China, September 2017.

[61] M. Mémoli, "Gromov-wasserstein distances and the metric approach to object matching," *Foundations of Computational Mathematics*, vol. 11, no. 4, pp. 417–487, 2011.

[62] J. Solomon, F. De Goes, P. Gabriel et al., "Convolutional wasserstein distances: efficient optimal transportation on geometric domains," *ACM Transactions on Graphics*, vol. 34, no. 4, p. 66, 2015.

[63] N. Bonneel, G. Peyre, and M. Cuturi, "Wasserstein barycentric coordinates: histogram regression using optimal transport," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1–10, 2016.

[64] N. Papadakis and J. Rabin, "Convex histogram-based joint image segmentation with regularized optimal transport cost," *Journal of Mathematical Imaging and Vision*, vol. 59, no. 2, pp. 161–186, 2017.

[65] Y. Zeng, X. Fu, and L. Gao, "Robust multivehicle tracking with wasserstein association metric in surveillance videos," *IEEE Access*, vol. 8, pp. 47863–47876, 2020.

[66] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3457–3464, Providence, RI, USA, June 2011.

[67] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1918–1925, Providence, RI, USA, June 2012.

[68] B. Yang and R. Nevatia, "An online learned CRF model for multi-target tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2034–2041, Providence, RI, USA, June 2012.

[69] A. R. Zamir, A. Dehghan, and M. Shah, "GMCP-tracker: global multiobject tracking using generalized minimum cliquegraphs," *Computer Vision—ECCV 2012*, Springer, Berlin, Germany, pp. 343–356, 2012.

[70] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "A simple baseline for multi-object tracking," 2020, https://arxiv.org/abs/2004.01888.

[71] S. Lyu, M.-C. Chang, D. Du et al., "UA-DETRAC 2018: report of AVSS2018 & IWT4S challenge on advanced traffic monitoring," in *Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 1–6, Auckland, New Zealand, November 2018.

[72] X. Qi, Z. Liu, Q. Chen et al., "3D motion decomposition for RGBD future dynamic scene synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7665–7674, Long Beach, CA, USA, June 2019.

[73] P. Voigtlaender, "Mots: multi-object tracking and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7934–7943, Long Beach, CA, USA, June 2019.

[74] B. Keni and S. Rainer, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, Article ID 246309, 2008.