A Review of Vision-Based Traffic Semantic Understanding in ITSs

Jing Chen, Qichao Wang, Harry H. Cheng, *Senior Member, IEEE*, Weiming Peng, and Wenqiang Xu

***Abstract*— A semantic understanding of road traffic can help people understand road traffic flow situations and emergencies more accurately and provide a more accurate basis for anomaly detection and traffic prediction. At present, the overview of computer vision in traffic mainly focuses on the static detection of vehicles and pedestrians. There are few in-depth studies on the semantic understanding of road traffic using visual methods. This paper aims to review recent approaches to the semantic understanding of road traffic using vision sensors to bridge this gap. First, this paper classifies all kinds of traffic monitoring analysis methods from the two perspectives of macro traffic flow and micro road behavior. Next, the techniques for each class of methods are reviewed and discussed in detail. Finally, we analyze the existing traffic monitoring challenges and corresponding solutions.**

***Index Terms*— Traffic surveillance analysis, macro-traffic flow, micro-vehicle behaviors, temporal reasoning, computer vision.**

I. Introduction

OVER the last decade, traffic detection has shown impressive results in road operation management. It has been widely used in vehicle behavior understanding [1], [2] and traffic flow analysis and prediction [3], [4]. However, traditional detection algorithms can only acquire information on relatively single traffic tasks because it only uses a single sensor to capture single detection data points. For example, ground loops can only obtain traffic flow volume, and radar can only measure speed. A combination of sensors is often needed to assist in a single traffic detection task [5]–[7]. However, the integration and fusion of different traffic detection data have become a bottleneck due to the features of both multisource and heterogeneous data. Due to the heterogeneity of multimodal data, it is difficult to accurately align them in both the original input space and the feature space. For example, target counting results obtained by ground loops have difficulty corresponding to multiple object detection results of the image. The same problem exists between vision data and point cloud data obtained by LiDAR. Furthermore, due to the multisensor consistency problem, there is no effective method for cross-modal data enhancement. As a result, the visual detection method, which can obtain various traffic parameters and has a wide field of view, has become increasingly popular [8], [9].

At present, most of the research on road surveillance focuses on object detection. In fact, understanding road vehicle behavior can help people more accurately understand road traffic situations and the occurrence of emergencies and provide a more accurate basis for abnormal judgment and flow prediction [10], [11]. In this way, the construction of intelligent transportation systems (ITSs) could also be accelerated [12].

With the rapid development of video-based vehicle behavior analysis, there are still various challenges in the application of ITSs. We summarize current challenges for understanding traffic semantics as follows:

1) Defects such as occlusions and shadows reduce the accuracy of vehicle detection and tracking and thus affect the recognition of vehicle behavior. This situation is particularly evident in peak hours of congestion on city-wide roads. Depth information can be used to solve the problem of partially occluded vehicles to effectively improve the detection accuracy of partially occluded vehicles caused by road congestion and low sampling angles.

2) Traffic flow quantification in a complex road network is critical for analyzing and predicting overall traffic flow. Unfortunately, research on traffic flow quantification has focused primarily on a single junction or section, and few studies have been performed on road networks using visual methods.

3) Limiting the traffic semantics of training samples prevents the current algorithm from thoroughly learning the overall vehicle behavior. Moreover, due to the diversity of traffic scenes and the heterogeneity of traffic agents, existing vehicle behavior recognition algorithms are generally not applicable.

To summarize existing road traffic semantics and vehicle behavior analysis methods, Fig. 1 shows details of the methods used for classifying the semantics of macro traffic flow and micro vehicle behavior considered in this review.

As part of this survey, we review the latest advances in vision-based road traffic semantics. They are divided into macro traffic flow spatiotemporal feature quantification and micro road behavior recognition. There are many reviews

Manuscript received 21 August 2021; revised 21 February 2022 and 10 April 2022; accepted 8 June 2022. Date of publication 17 June 2022; date of current version 7 November 2022. This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY21F020014 and in part by the National Science Foundation of China under Grant 61976074. The Associate Editor for this article was Z. Duric. *(Corresponding author: Jing Chen.)*

Jing Chen, Qichao Wang, and Weiming Peng are with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: cj@hdu.edu.cn; fugowang@hdu.edu.cn; penwm@hdu.edu.cn).

Harry H. Cheng is with the Mechanical and Aerospace Engineering Department, University of California at Davis, Davis, CA 95616 USA (e-mail: hhcheng@ucdavis.edu).

Wenqiang Xu is with the College of Economics and Management, China Jiliang University, Hangzhou 310018, China (e-mail: wenqiang_xu163@163.com).

Digital Object Identifier 10.1109/TITS.2022.3182410

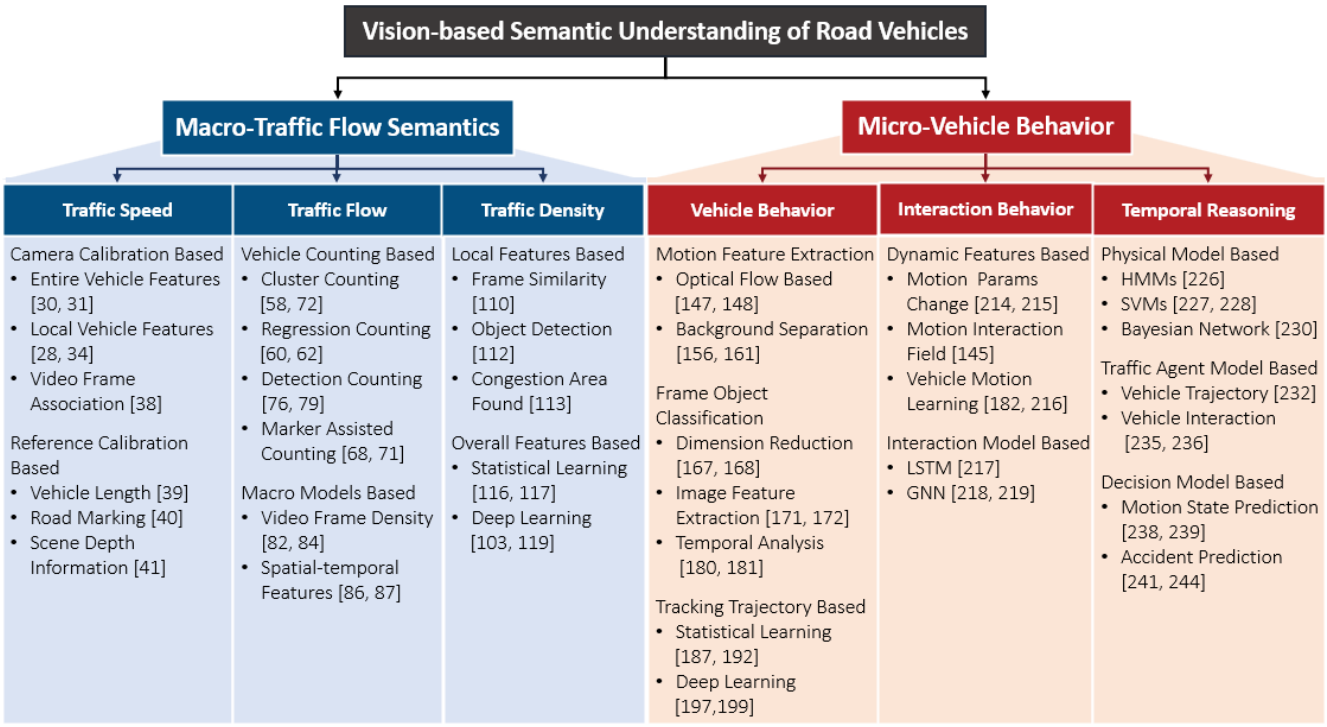

Fig. 1. Detailed classification of macro-traffic flow semantics and micro vehicle behavior.

on traffic video analysis. Gao *et al.* [13] investigated convolutional neural network(CNN)-based density map estimation methods for crowd counting. Based on these methods, traffic density estimation can also be achieved. Llorca *et al.* [14] summarized the vehicle speed estimation work based on visual information. In [15], traffic anomaly detection in public places was reviewed. Buch *et al.* [16] introduced automatic number plate recognition (ANPR) methods, vehicle counting, and incident detection in traffic videos. Datondji *et al.* [17] reviewed vehicle detection, tracking, and vehicle trajectory analysis at road intersections. Unlike previous work, our survey is not limited to abnormal detection and individual vehicle behavior analysis (e.g., lane change, make a turn, and stop) but also includes interactions between vehicles and between vehicles and other traffic agents (e.g., pedestrians, cyclists, and motorcyclists). Traffic temporal reasoning based on vehicle behavior analysis has some "predictability" for traffic flow prediction; thus, we also summarize the related work on vehicle behavior prediction. Moreover, we carry out a quantitative analysis of the three main features of macro traffic flow (speed, flow volume, and density).

The contributions of our paper are as follows:

- At present, most reviews of visual methods [18]–[20] in traffic focus on static object detection of road vehicles and pedestrians, with a focus on the position and category information of vehicles and pedestrians in traffic scenes. Other reviews [15], [21]summarized the behavior recognition methods, mainly for understanding driver behavior, autonomous driving vehicles, and other individual behaviors. But they did not understand traffic semantics by visual methods from the perspective of traffic management. Therefore, this paper attempts to categorize research on understanding traffic flow semantics from a traffic flow analysis perspective. The research scope of this review is for large-scale traffic detection scenarios. Sampling data is mainly from roadside sensors rather than cameras installed inside moving vehicles. Therefore, there are few papers on vehicle behavior detection in the field of automatic driving.
- This paper reviews recent advances in computer vision-based road behavior recognition methods and describes their implementation, advantages, and disadvantages. It includes macro spatiotemporal feature quantification and micro road behavior recognition. Unlike previous work, we performed quantitative analysis on the three main characteristics of macro traffic flow (speed, flow volume, and density). In addition, our work is not limited to the anomaly detection and behavior analysis of single-vehicle behavior (such as lane change, turning, etc.) but also includes the interaction between vehicles and between vehicles and other traffic agents. Because temporal reasoning based on road behavior recognition has certain "predictability" for traffic flow prediction, we also summarize the related work of vehicle behavior prediction.
- We summarize the challenges and prospects of understanding road semantics based on visual methods.

The remainder of this survey is organized as follows. Section II analyzes the three features of macro-traffic flow: traffic speed, traffic flow volume, and traffic density. Section III introduces vehicle behavior and the interactions between vehicles and surrounding traffic agents, as well as the complex temporal reasoning of vehicle behavior. Section IV discusses the challenges of vehicle behavior and suggests future directions. Section V summarizes the paper.

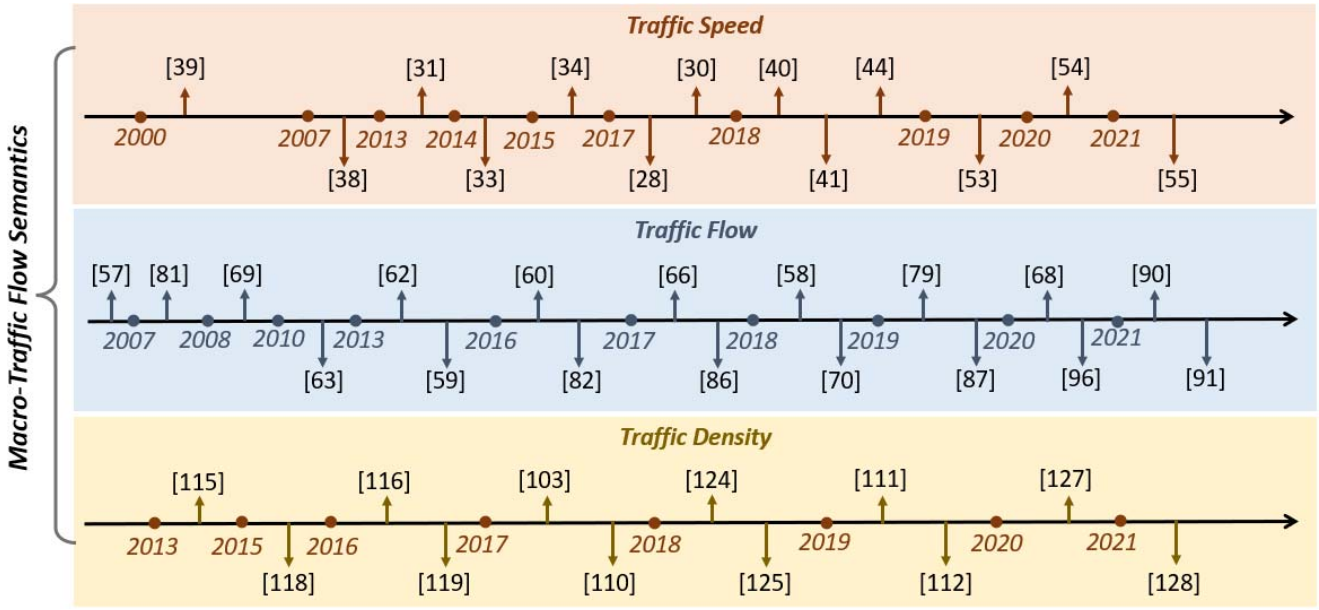

Fig. 2. A brief chronology of macro traffic flow spatiotemporal feature quantification.

II. Quantifying the Spatiotemporal Features of Macro Traffic Flow

Traffic speed, flow volume, and density are three main parameters that describe traffic flow and influence each other. Surveillance video contains rich visual information [22], [23]. The actual status and periodicity of urban traffic can be obtained by quantifying the spatiotemporal features of traffic flow. Surveillance video plays an essential role in traffic flow analysis, prediction, and traffic dispatch management. Fig. 2 is a brief chronology of macro traffic flow spatiotemporal feature quantification. And table VI lists the popular methods for quantifying macro traffic flow.

A. Quantifying Traffic Speed Features

Compared with traditional sensor-based methods such as inductive loops [24]–[26], LiDAR [27]–[29], and radar [30], visual-based traffic speed quantization methods are inexpensive. At the same time, there is much information in RGB videos, which can detect and track vehicles in traffic scenes to obtain speed information [14], [31]–[33].

Specifically, visual methods convert the pixel displacement in an image to the driving distance of the vehicle in the real world. It is divided into two main streams: camera calibration-based and reference calibration-based approaches.

1) Approaches Based on Camera Calibration: By accurately retrieving a camera's intrinsic and extrinsic parameters, one can convert the camera coordinate system to the global coordinate system. Then, actual vehicle displacement can be calculated from the pixel distance. The speed of traffic flow at a given time can be easily calculated. Wicaksono *et al.* [34] estimated the perpendicular view and the distance between a vehicle and camera based on camera calibration. The vehicle displacement during successive frames can be used to realize speed estimation. However, the positions of vehicles in video frames are affected by the scale of the real world. Nurhadiyatna *et al.* [35] projected a real road scene into the pinhole model to estimate the speed to overcome this problem. These works fall into directly detecting and tracking the whole vehicle.

Speed estimation can also be performed by using the local landmark features [36]. For example, Luvizon *et al.* [37] located and tracked license plates to obtain motion vectors. The motion vector is converted into a velocity vector by determining the relationship between the motion of pixels and vehicles in the real world. Similarly, Wu *et al.* [38] took the license plate as the vehicle's feature point for obtaining interframe trajectory displacement to calculate velocity. Sochor *et al.* [32] proposed projecting reference points of the tracked vehicle onto the ground plane while estimating the speed by computing the moving distance. Scene flow can describe 3D motion and enrich semantic understanding of the overall traffic scene [39]. In scene flow, the relative speed of a vehicle can be defined as the vehicle's motion relative to the camera in a certain period. Song *et al.* [40] used intrinsic parameters of the camera and scale information of a vehicle to obtain geometric clues regarding the vehicle. Combined with optical flow estimation approaches [41], the relative velocity of the vehicle can be estimated.

All the above methods directly or indirectly need to track all or parts of a vehicle. Another alternative approach for speed estimation is built on video frame association. After transferring two continuous video frame images from a 2D image coordinate system to a 3D global coordinate system, He *et al.* [42] correlated the same vehicle in two frames to one image to obtain vehicle moving distance and calculate the vehicle speed.

2) Approaches Based on Reference Calibration: Reference-based quantization methods mainly use the geometric length of specific objects in a scene (e.g., vehicle length and road line separation distance) to indirectly obtain the pixel displacement distance of vehicles. The main distinction

TABLE I
VEHICLE SPEED ESTIMATION BASED ON AI CITY CHALLENGE 2018 TASK1

Category	Classification	Method	DR	RMSE
Camera Calibration Based	Entire Vehicle Features	Tang et al.[48]	100%	4.096
	Entire Vehicle Features	Kumar et al.[31]	100%	9.541
	Entire Vehicle Features	Shi et al.[49]	100%	6.667
	Local Vehicle Features	Mao et al.[50]	100%	7.970
Reference Calibration Based	Road Marking	Tran et al.[44]	100%	8.914
	Road Mraking	Feng et al.[51]	100%	6.041
	Road Marking	Sochor et al.[52]	81.48%	6.869
	Road Marking	Huang et al.[53]	81.48%	8.609

between reference-based approaches and camera-calibration-based approaches is that the length of the reference objects can be used directly to obtain the scene scale information. In fact, the quantization of the speed features of the two approaches depends on the displacement of the vehicle pixel.

An efficient method [43] uses the mean vehicle length to obtain the scale information. However, this work assumes that the vehicle is always moving toward or away from the camera, making it difficult to adapt to actual complex traffic scenes. However, it inspired subsequent research that used prior road scene information to calculate vehicle displacement. Tran *et al.* [44] used the length of the white lines and the distance between the white lines on highways to generate equally distributed scan lines perpendicular to the road direction. As a result, the mapping relationship between the pixel distance and the real-world vehicle displacement can be calibrated. Then, the vehicle speed can be estimated from the time interval in which the vehicle passes through two consecutive scanlines.

Unlike methods that choose the actual reference for calibration, Kampelmühler *et al.* [45] took the depth and optical flow information as references. The key idea of this approach is to track the vehicle and extract trajectory, depth, and optical flow features. The depth and motion clues contained in these features are aggregated at the location of the tracked vehicle. The ultimate purpose is to train a fully connected network to estimate vehicle speed accurately.

3) Quantitative and Qualitative Results: Currently, AI CITY CHALLENGE [46] and BrnoCompSpeed [47] are datasets commonly used for vehicle speed estimation. AI CITY CHALLENGE 2018 TASK1 training and test data are collected from highway interchanges and signalized intersections. Every video clip has a frame rate of 30 fps and a resolution of 1980 $\times$ 1080. No additional information (such as camera calibration parameters, camera height above the ground, etc.) is provided except for the videos captured by the camera. Road vehicle detection rate (DR) and root mean square error (RMSE) of the vehicle speed are used for evaluation. The experimental results are shown in Table I.

In the absence of internal and external camera parameters, the real world can be associated with pixel displacement by a perspective change or transformation based on the horizon vanishing point (VP) or lane width reference in the real world. However, this method makes strong assumptions, e.g., the road is straight and flat, the camera has no distortion, and the vehicle's center of mass is fixed. The BrnoCompSpeed dataset includes not only video from traffic surveillance cameras but also metadata such as camera calibration parameters and the length of features in the images. Each video has a resolution of 1920 $\times$ 1080 and a frame rate of 50 fps and is approximately one hour long. The vehicle speed is accurately measured using LiDAR. The results of the speed measurement based on this dataset are shown in Table II.

Kocur *et al.* [58] used 2D bounding boxes and 3D bounding boxes on the BrnoCompSpeed dataset to detect vehicles. According to the results in the table, constructing 3D bounding boxes can improve the speed measurement accuracy. In addition, they also used different geometric VPs to construct a perspective transformation in the traffic scene to obtain the best VP pair. Zhu *et al.* [59] performed 3D vehicle detection with BEV rotated boxes (R-boxes) generated by inverse perspective mapping. Their experiment proved that this method could greatly improve the vehicle location accuracy in the real world.

4) Discussion: The speed estimation methods based on reference calibration exploit the traffic scene information. However, these methods perform poorly in a complex traffic scene with large traffic flow, traffic congestion, and severe vehicle occlusion. The vehicle scale information in these scenes is difficult to perceive and calculate. The methods based on camera calibration use the parametric matrix of the camera to reconstruct the traffic scene, which can prevent the reference calibration methods from failing in a complex scene. However, the video captured by different cameras needs to be recalibrated.

B. Quantifying Traffic Flow Volume Features

Traffic flow volume is the total number of vehicles on a given road in a specific period of time [22], [60]. Flow volume has a spatiotemporal distribution feature, and its value is a random value. From both local and global perspectives, we categorize flow volume quantization methods into vehicle-counting-based and macro-model-based methods.

1) Based on Vehicle Counting: The definition of traffic flow volume shows that vehicle counting is the most direct method

TABLE II
SPEED MEASUREMENT BASED ON THE BRNOCOMPSPEED DATASET

Classfication	Methods	VP Pair	Mean Error(km/h)	Median Error(km/h)	Mean Precision(%)	Mean Recall	Year
Entire Vehicle Features	Dubská et al.[54]	-	8.22	7.87	73.48	90.08	2014
	Sochor et al.[32]	-	1.04	0.83	90.72	83.34	2017
	Yu et al.[55]	-	2.77	-	-	-	2018
	Dong et al.[56]	-	2.73	-	-	-	2019
	Transformation2D[57]	VP2-VP3	0.83	0.60	83.53	82.06	2019
	Transformation3D[57]	VP2-VP3	0.86	0.65	87.67	89.32	2019
	ImproveTransformation2D[58]	VP2-VP3	0.92	0.69	84.73	77.58	2020
	ImproveTransformation3D[58]	VP2-VP3	0.79	0.60	87.08	83.32	2020
	Zhu et al.[59]	-	-	-	92.70	97.25	2021

of traffic estimation. We divide the vehicle counting methods into four categories: cluster counting [61]–[63], regression counting [64]–[66], detection counting [67]–[71] and marker-assisted counting [72]–[75]. In the following section, we give a brief introduction to each method.

a) Cluster counting: Cluster counting obtains several typical driving tracks by clustering vehicle tracks on the road and finally counts vehicles on these typical tracks [76]. Specifically, based on the reconstruction of 3D feature point information from coarse to fine clustering, Song *et al.* [62] obtained trajectory-based feature points that can be utilized to count targets. After clustering the trajectory feature points, Rabaud *et al.* [77] then segmented the moving targets to accurately obtain the traffic flow volume count.

b) Regression counting: Regression-based methods learn the mapping relationship between low-level image features and the road target number. For example, Chen *et al.* [66] use a regression function to learn the change in the road target count in the attribute space. Similarly, Liu *et al.* [64] apply a hierarchical classification-based regression (HCR) model to divide traffic scenes into different categories according to the scene density with a two-level classifier.

c) Detection counting: Unlike the clustering-based counting and regression-based counting methods, counting-based detection methods directly detect and track road vehicles. Specifically, the extensive application of CNN object detectors in recent years has made the detection-counting-based method more accurate [78], [79]. In [71], color histograms and geometrical constraints were combined to realize matching between two frames. Then, a bounding box was used to track and count road vehicles. This work can achieve high precision and real-time performance. Abdelwahab *et al.* [80] proposed assigning a serial number to every detected vehicle and tracking and counting vehicles from the time of vehicle detection to the time at which the vehicle disappears from the camera field of view. This strategy can realize the global perception of the vehicle distribution in traffic scenes. In addition, the optical flow can be used to assist in vehicle detection and tracking [81], [82]. In [83], the tracking results based on CNNs and optical flow are combined to detect vehicle motion feature points for tracking and counting.

d) Marker-assisted counting: Methods based on marker-assisted counting use virtual detection lines or virtual regions to count vehicles entering and leaving the field of view. Huynh *et al.* [72] specified a reference line in a video to count the number of vehicle crossings. In [73], a rectangular region was selected as the virtual detection region, and the vehicles that passed through this region were counted. In fact, virtual detection lines and virtual detection regions are appropriate for different flow volume situation. The two methods were analyzed comprehensively in [75]. Virtual detection lines [84] were used for counting when the speed of a vehicle was high. When the traffic is slow due to congestion, it is more likely that the detection line will would count adjacent vehicles as the same vehicle because the vehicles are slow and close to each other. Therefore, the virtual detection region [85] was used for counting. However, due to the limitation of the low-frame-rate camera, the whole vehicle may skip the detection of the virtual line or region, resulting in poor performance.

2) Approaches Based on Macro Models: Instead of counting each vehicle on the road, other works choose to quantify flow volume with the overall traffic scene features. The global representation of a traffic scene is based on a macro view of the traffic flow volume, without relying on counting all vehicles in the video frames. According to whether the rich temporal correlation between adjacent video frames is utilized, the approaches based on the macro model can be divided into density-map-based and spatiotemporal-feature-based approaches. In both methods, density maps play a central role.

a) Density maps: Methods based on density maps output the density map of an entire video frame by using a CNN to extract image features [86]–[89]. Then, the density maps perform flow volume estimation by either the regression or integration method. In [86], a three-branch multicolumn CNN architecture was applied to improve the traffic flow volume estimation performance and capture multiscale information. In contrast to [86], Kang *et al.* [87] showed that an image pyramid combined with an across-scale attention map can also be used to obtain multiscale information. The work in [88] used two fully convolutional neural networks (FCNs) to extract the shallow and deep features of an image for feature fusion, which is suitable for flow volume estimation in a large-scale crowded condition. By using a CNN to generate density maps, Zhao *et al.* [89] divided the traffic estimation problem into two subproblems, i.e., estimating crowd density maps and crowd velocity maps, to learn more features.

TABLE III
TRAFFIC FLOW VOLUME PERCEPTION RESULTS BASED ON THE AI CITY CHALLENGE VEHICLE COUNTING DATASET

Category	Classification	Methods	Score	Year
Vehicle Counting Based	Detection Counting Based	Lu et al.[94]	0.9467	2021
		Ha et al.[95]	0.9459	2021
		Tran et al.[96]	0.9263	2021
		Tran et al.[97]	0.9249	2021
	Marker-Assited Counting	Gloudemans et al.[98]	0.8569	2021
	Detection Counting Based	Kocur et al.[99]	0.8449	2021
	Detection Counting Based	Liu et al.[100]	0.9389	2020
		Ospina et al.[78]	0.9346	2020
	Marker-Assisted Counting	Yu et al.[101]	0.9292	2020
		Folenta et al.[102]	0.8829	2020
		Bui et al.[103]	0.8540	2020

b) Spatiotemporal features: Methods based on spatiotemporal features combine temporal sequence features with density map information to estimate traffic flow volume. To exploit the temporal correlation between density maps of adjacent frames, based on a deep spatiotemporal neural network, Zhang *et al.* [90] connected an FCN and LSTM in a residual learning fashion for flow volume estimation. Fang *et al.* [91] utilized the locality-constrained spatial transformer (LST) module to obtain the mapping relationship of density maps between adjacent frames. In addition, the motion features of road targets can be helpful in enhancing the effect of flow volume estimation. After obtaining the density map, Xiong *et al.* [92] used ConvLSTM [93] to carry out spatiotemporal modeling to integrate motion information in the time dimension.

Methods based on the macro model utilize the whole feature of video frames to estimate traffic flow volume. The inability of the detection and tracking method to accurately count all vehicles when the traffic flow volume is large can be resolved.

3) Quantitative and Qualitative Results: The AI CITY CHALLENGE vehicle counting dataset can be used to realize traffic flow volume perception by counting the number of different vehicle types (cars, trucks, etc.) turning left, turning right or proceeding straight at complex traffic intersections to design a better signal control strategy and realize scheduling optimization. The evaluation metric of this task is determined by the performance efficiency of the algorithm and the RMSE of the vehicle count result. In addition, the 2021 task requires the algorithm to run in real time. The experimental results are shown in Table III.

Most of these works use the detection-tracking-counting (DTC) framework to obtain the vehicle tracks. For the subsequent counting task, the vehicle trajectory can be matched according to the predefined motion trajectory. Furthermore, the vehicle motion direction can be determined according to the source and sink regions. Compared with the predefined trajectory, this method more easily achieves real-time performance, but its accuracy is reduced.

The UCSD dataset [104] uses cameras to capture pedestrian flow volume on sidewalks rather than specifically counting large crowds. The pedestrian flow volume is variable, and the flow volume changes dynamically from sparse to crowded. The accuracy and applicability of the traffic flow volume estimation algorithm can also be verified by experiments on these datasets. The mean absolute error (MAE) and mean square error (MSE) are used for evaluation. The experimental results are shown in Table IV.

Few works have used this dataset to estimate flow volume due to the low video quality and short video length. As shown in the table, the experimental methods are mostly based on regression region features. In addition, density maps based on CNNs can be extracted for counting, but this method is more suitable for counting in crowded areas.

In addition, cross-camera traffic analysis can be performed based on multitarget multicamera tracking work. In city-scale multicamera vehicle tracking [108]–[110], most of the works perform the following steps: object detection, multitarget single-camera tracking, vehicle ReID, and cross-camera trajectory matching. The vehicle counts and trajectory needed for cross-camera traffic analysis can thus be obtained.

4) Discussion: In traffic flow volume estimation, the performance of methods based on vehicle counting is affected by the poor vehicle detection accuracy due to the use of low-resolution and low-frame-rate videos. Several works have shown that these limitations can be overcome by obtaining the vehicle density map and combining the previous traffic flow motion state.

C. Quantifying Traffic Flow Density

When analyzing the effect of traffic density, road congestion can be predicted in a timely manner [111]–[113]. Traffic density estimation approaches can be divided into local-information-based and holistic-feature-based methods according to whether global spatial information is used.

1) Approaches Based on Local Information: By analyzing the interframe similarity in traffic scenes, the density of the

TABLE IV
PEDESTRIAN FLOW VOLUME EVALUATION RESULTS OBTAINED USING THE UCSD DATASET

Classification	Methods	MAE	MSE	Year
Regression Counting	Kernel Ridge Regression[105]	2.16	7.45	2007
	Gaussian Process Regression[104]	2.24	7.97	2008
	Ridge Regression[106]	2.25	7.82	2012
	Cumulative Attribute Regression[66]	2.07	6.86	2013
Vedio Frame Density	Cross-scene DNN[65]	1.60	3.31	2015
	OPT-RC[107]	2.03	5.97	2017
	FCN-MT[107]	1.67	3.41	2017

vehicle distribution in a local area can be determined. If the two adjacent frames are definitely similar, the traffic density is congested. In [114], the traffic density was estimated by computing the image correlation coefficient of continuous frames. However, this strategy requires images taken from different angles and positions for feature pretraining and has only two classes: NORMAL and CONGESTED. Lam *et al.* [115] used the bounding boxes of vehicles to calculate the intersection over union (IOU) of continuous video frames to obtain traffic density information. In both works, a measure of similarity between continuous frames of a video is used.

In fact, most approaches use a deep learning object detector to detect vehicles in video frames. Qi *et al.* [116] acquired traffic flow features by detecting vehicles at an intersection. Then, they adopted particle swarm optimization (PSO) to optimize the penalty coefficient for traffic flow density estimation. Jiang *et al.* [117] regressed the entire congested area in video frames by quadrilateral object modeling and improved the loss function of YOLOv3 [118]. An advantage of this approach is that congested areas can be identified directly instead of having to detect each vehicle in the traffic scene.

2) Approaches Based on the Overall Features: In scenes with high traffic density, many vehicles are overlapped and occluded, which increases the difficulty of traffic density estimation. Holistic approaches based on the overall features are proposed to solve this problem. These methods attempt to obtain global information (e.g., crowded flows and density maps) from a traffic scene rather than detecting and tracking vehicles. Depending on whether deep learning is used, we categorize these methods into statistical-learning-based and deep-learning-based methods.

a) Approaches based on statistical learning: In most statistical learning methods, traffic flow appearance and motion information are used to classify the traffic scene density. Sobral *et al.* [119] calculated the average crowd density and crowd speed based on vehicle congestion attributes to classify the traffic flow density. This method requires tracking the vehicle with its spatial appearance and dynamical information to obtain the density; thus, the algorithm process is relatively complicated. There is a relatively convenient way to estimate density. Grag *et al.* [120] proposed dividing each lane into blocks and then detecting the occupancy rate of vehicles on these blocks. As a result, the global density information can be obtained. This strategy is straightforward and can be implemented in real time, but its precision is not high. In [121], global crowd features were extracted to construct multidimensional feature vectors. After training based on Gaussian mixture hidden Markov models (GM-HMMs), the maximum likelihood (ML) criterion was applied to traffic flow density classification. These methods all classify the traffic flow density by extracting global features from traffic videos. Limited by the simple implementation algorithm, these methods have poor generalization ability. In particular, they are not suitable for complex traffic scenes (e.g., night and rainy scenes).

b) Approaches based on deep learning: Deep learning methods can extract high-level global image features, giving them a wide range of applicability in understanding traffic semantics under the condition of full and sufficient training samples. In [122], codebook visual descriptors and CNNs were used to extract density features, combined with an SVM, to classify traffic flow density. Luo *et al.* [123] performed semantic segmentation on traffic video with a low frame rate. More specifically, they considered three different CNN models to transform the entire image into feature vectors to jointly estimate the density.

To overcome the challenges brought by the shortcomings of video, e.g., low frame rate, low resolution, high occlusion, and large perspective, Zhang *et al.* [107] obtained the traffic flow density based on rank-constrained regression and FCNs. The traffic flow density was calculated by embedding scene geometry information into the weight matrix in the regression method. Moreover, the density map was obtained based on the end-to-end FCNs. In short, methods based on deep learning object detection and image segmentation can fully understand the overall traffic flow density, and their performance is far better than that of the methods based on statistical learning.

3) Quantitative and Qualitative Results: The TRaffic ANd COngestionS (TRANCOS) dataset [124] is dedicated to detecting traffic congestion. It contains 1,244 images, with a total of 46,796 annotated vehicles, taken by surveillance cameras on the streets of Spain.

Although most works use the MSE or MAE to evaluate experimental results, these evaluation metrics usually result in inaccurate estimations due to vehicle occlusion in an image. Therefore, TRANCOS proposes the grid average mean

TABLE V
TRAFFIC CONGESTION ESTIMATION METHOD RESULTS BASED ON THE TRANCOS DATASET

Category	Classification	Methods	GAME(0)	GAME(1)	GAME(2)	GAME(3)	Year
Overall Features Based	Statistical Learning	Lempitsky et al.[125]	19.76	16.72	20.72	24.36	2010
	Deep Learning	Hydra-3s[126]	10.99	13.75	16.69	13.32	2016
		TraCount[127]	8.12	-	-	-	2016
		FCN-rLSTM[90]	4.38	-	-	-	2017
		Laradji et al.[128]	13.76	16.72	20.72	24.36	2018
		Li et al.[129]	3.56	5.49	8.57	15.04	2018
		PSDDN[130]	4.79	5.43	6.68	8.40	2019
		ADMG[131]	3.79	5.93	8.49	13.51	2020
		KDMG[131]	3.13	4.79	6.20	8.68	2020
		Ciampi et al.[132]	3.30	-	-	-	2021

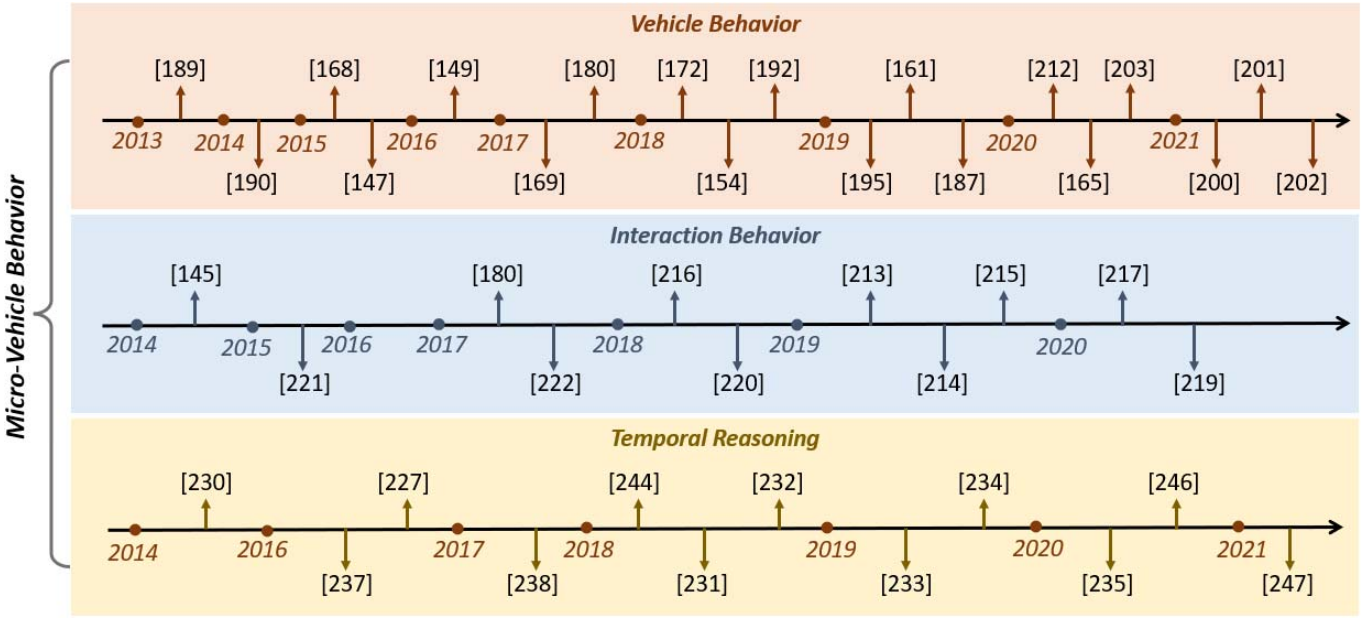

Fig. 3. A brief chronology of micro road behavior recognition.

absolute error (GAME) index, considering the number and locations of all vehicles. The traffic congestion estimation methods based on the TRANCOS dataset are shown in the following table.

In these works, CNNs are used to extract the spatial features of images and generate density maps for counting to estimate the density of traffic congestion. However, there is no temporal information in the TRANCOS dataset. The traffic congestion detection result is shown in Table V.

4) Discussion: In traffic flow density estimation, methods based on local information require high-resolution and high-frame-rate videos. At the same time, the estimation is easily influenced by the occlusion of vehicles and heavy traffic congestion. In contrast, the global spatial information of the image is utilized to solve the influence of vehicle partial occlusion in high-density-based deep learning methods.

D. Conclusions

Works quantifying traffic speed, flow volume, and traffic density have been able to fully understand macro traffic flow. However, constrained by the variety of real traffic scenes (e.g., highways and intersections) and the complexity of weather and lighting conditions, there are still some challenges in practical applications. Quantization of the traffic speed, flow volume, and traffic density plays a central role in macro traffic flow quantification. Furthermore, macro traffic flow analysis, combined with a series of quantitative parameters, such as queue length, time headway, and space headway, will be carried out to comprehensively understand macro traffic flow features.

III. MICRO BEHAVIOR RECOGNITION ON ROADS

In this section, we introduce the approaches for recognizing traffic agent behaviors on roads. We categorize micro behavior understanding into three parts: vehicle behavior understanding, traffic agent interaction behavior recognition, and complex temporal reasoning of vehicle behavior. Details on the categorization of these methods are given below.Fig. 3 is a brief chronology of micro road behavior recognition.

TABLE VI
POPULAR METHODS FOR QUANTIFYING MACRO-TRAFFIC FLOW

Traffic Flow Features	Classifiers	Specific Method	Descriptions
Traffic Speed	Camera Calibration Based	Entire Vehicle Features[34, 35]	Detect and track the displacement of a car within a specified time
		Local Vehicle Features[32, 37, 38]	Vehicle tracking is realized through local features such as license plate and headlights
		Video Frame Association[42]	Observe the displacement of a vehicle in a continuous video frame
	Reference Calibration Based	Vehicle Length[43]	The scene scale information is inferred to realize displacement calculation by the vehicle length
		Road Marking[44]	The length and width of the road marker can also be used to calculate the displacement of a vehicle in a given time
		Scene Depth Information[45]	Distance information is measured by depth estimation
Traffic Flow Volume	Vehicle Counting Based	Cluster Counting[61–63]	Cluster trajectory feature points are used to count the number of vehicles.
		Regression Counting[64–66]	Learn the mapping relationship between the underlying image features and the number of objects
		Detection Counting[67–71, 80, 83]	An object detector detects road vehicles to realize counting
		Marker-Assisted counting[72–75, 84, 85]	Count the number of vehicles crossing a specified area or marked line of a road
	Macro Models Based	Video Frame Density[86–89]	The density map of the congested road area is extracted by CNNs to realize counting
		Spatiotemporal Features[90–93]	Flow volume is estimated by combining traffic temporal information and a density map
Traffic Density	Local Features Based	Frame Similarity[114, 115]	Compare the similarity of adjacent video frames to determine the congestion degree of a traffic scene
		Object Detection[116]	Density estimation is achieved by vehicle object detection per frame
		Congestion Area Found[117]	Detection of the congestion areas in a scene rather than the detection of each vehicle
	Overall Features Based	Statistical Learning[119–121]	Estimate scene density with the overall traffic flow appearance and motion information
		Deep Learning[107, 122, 123]	The global features of video frames are extracted based on CNNs

A. Vehicle Behavior Recognition

For vehicle behavior recognition, several approaches have been developed, from the early manual extraction of video spatiotemporal features [133], [134] to deep-learning-based methods [135]–[137]. With the different features in behavior recognition as input, we classify these methods into motion-feature-extraction-based, video-frame-object-classification-based, and tracking-trajectory-analysis-based methods. Motion features contain rich vehicle dynamic behavior information, which is extracted by optical flow and background modeling. In the methods based on video frame object classification, vehicle behavior features (e.g., vehicle position and motion pattern) are used to discover vehicle behavior changes and abnormal behavior. The methods based on tracking trajectory analysis identify vehicle behavior by tracking vehicle trajectories. In Table IX, we summarize the representative approaches of vehicle behavior recognition.

1) Motion Feature Extraction: Methods based on motion feature extraction are widely used in traffic surveillance (e.g., vehicle tracking [138]–[140], motion estimation [141], and anomaly detection [142], [143]). In traffic scenes, the most obvious property is motion. The common methods of extracting vehicle motion features include optical flow and background separation.

a) Approaches based on optical flow: Optical flow is one of the most commonly extracted low-level features. Optical flow [144]–[146] and histograms of oriented gradient (HOG) [147], [148] can capture motion and appearance changes, effectively representing the overall motion features of traffic scenes. These methods are used mainly in cases of abnormal driving, traffic accidents, etc.

i) Sparse optical flow: Sparse optical flow can be used to extract motion information according to the motion amplitude variation of local pixels, which usually requires a set of pixels (e.g., Harris corner points) for tracking variation. In [149], with the help of both local and global optical flow methods [150], vehicle motion information (e.g., speed and direction) was extracted, with which the motion interaction field (MIF) of traffic scenes was generated to detect and localize traffic accidents. Ahmadi *et al.* [151] clustered optical flow features to learn motion patterns in the traffic phase for detecting abnormal vehicle behavior, e.g., abnormal driving and not following traffic laws. In [152], the global amplitudes of optical flow in each frame were utilized to obtain the optical flow descriptors of traffic scenes. The descriptors and Fisher vector representing the spatiotemporal visual volumes were employed to detect traffic violations. In addition, the histogram of oriented optical flow directions (HOF) can be obtained by weighted statistics and coding the optical flow directions. Hasan *et al.* [153] presented an approach that uses the HOG and HOF as spatiotemporal appearance descriptors to encode appearance and motion information, respectively. Then, the regularized vehicle behavior is learned based on a feedforward autoencoder. These works all use sparse optical flow to extract local motion information. The advantage of this kind of method lies in the calculation speed, but the speed advantage is offset by the loss of some optical flow features, which will affect the accuracy of vehicle behavior recognition.

ii) Dense optical flow: The dense optical flow can be used to calculate the offset of pixels based on the movement of all the pixels in two frames. Ullah *et al.* [154] extracted dense optical flow through the Farneback optical flow algorithm [155] to obtain the motion flow field. Traffic accidents are detected based on the change in momentum in the motion flow field combined with the smoothed particle hydrodynamics (SPH) method. The Farneback optical flow algorithm can also be used to preprocess the dense optical flow feature set. The next step is to send it into the 3D convolutional neural network (3DCNN) as a temporal feature for traffic accident detection [156], [157]. Compared with sparse optical flow, dense optical flow can be used to extract the motion features and achieve a better effect more fully.

b) Approaches based on background separation: Based on background modeling, the background is separated from the video frame and compared with the current video frame to analyze the vehicle behavior. In traffic scenes, abnormal vehicles, e.g., broken down vehicles and those involved in an accident, become part of the background as stationary objects. Therefore, the background separation method can be performed to remove moving vehicles and analyze the abnormal vehicles in the background.

The Gaussian mixture model can typically remove moving vehicles and achieve background object extraction. To detect stationary vehicles in the background, the Gaussian mixture model (GMM) method is used to extract the background and pass it into the anomaly detection module [158], [159]. Wei *et al.* [160] performed MOG2 [161] to remove moving vehicles in the foreground, retaining only stopped vehicles for detection. The purpose of this strategy is to determine the positions of abnormal vehicles on the road and the start time of abnormal events. Based on [160], a road mask area is generated to remove the influence of moving vehicles and other areas (e.g., parking lots). Then, the Faster-RCNN [162] and multiple object tracking (MOT) algorithms are utilized to obtain vehicle trajectories for anomalous vehicle detection [163], [164].

In contrast, Bai et al [165] did not simply preserve stationary vehicles but overlaid each frame through background and spatial modeling to obtain the vehicle position distribution. Finally, the perspective detection module and the spatiotemporal matrix discrimination module were used to detect traffic accidents. Unlike any of the listed methods, Biradar *et al.* [166] proposed a CNN module for the deep background estimation model. Thus, the stationary vehicles in the background can be retained and localized. Based on the background separation methods, the areas unrelated to vehicle motion feature extraction can be effectively removed. Without these areas, the motion features related to vehicle behaviors can be conveniently extracted from the traffic scene for subsequent processing.

2) Video Frame Object Classification: Video frame object classification methods obtain the feature representation by extracting traffic video frame features [167]–[169]. In this way, the rich semantic information in each frame can be fully used to understand the vehicle behavior. Commonly used video frame object classification methods include data dimensional reduction, video frame image feature extraction, and sequential frame temporal feature extraction.

a) Video data dimensional reduction: Data dimensional reduction methods are used to analyze vehicle behavior by extracting the principal components of the feature data. In [170], after extracting lane feature vectors using the HOG, principal component analysis (PCA) was proposed to detect lane change behavior. Compared with the linear dimensional reduction method of PCA, isometric mapping (Isomap) was adopted based on the manifold learning method in [171]. Isomap was performed to recover low-dimensional manifold data from high-dimensional video frame data to detect vehicle lane change behavior.

In addition, the low-rank matrix can be used to represent the degree of redundancy in image information. According to the average velocity amplitude of image blocks, Xia *et al.* [172] calculated the average motion vector amplitude (AMVA) of nonoverlapping blocks to extract the motion matrix for detecting vehicle accidents. In [173], perceptual video summarization technology was proposed to select keyframes after analyzing the original video structure and the content spatiotemporal redundancy. By taking appearance as a perceptual feature, the strategy transforms the vehicle behavior analysis problem into an optimization problem. In conclusion, the dimensional reduction methods can retain the original energy as much as possible while compressing the feature dimension. Thus, the efficient analysis of vehicle behavior in a traffic scene can be realized.

b) Video frame image feature extraction: There are abundant traffic semantics in video frames; thus, most of the works use the video frame feature extraction method

to understand vehicle behavior. The sparse autoencoder can extract high-level features representing vehicle behavior. Wang *et al.* [174] extracted features by the stacked sparse autoencoder (SSAE) model and then used a softmax classifier to detect vehicle lane departure. The dynamic topic model can be applied directly to analyze activity in videos without the need to classify vehicle behaviors. Isupova *et al.* [175] proposed extracting structured visual features based on the Markov clustering topic model (MCTM) to describe vehicle activities and behaviors. However, most of these methods are suitable only for simple vehicle behavior analysis in a single traffic scene.

In recent years, CNN-based object detectors have made great progress. Vehicle behavior analysis based on CNNs has become the mainstream method of video frame object classification. In [176], object detectors such as Fasters-RCNN [162] and the SSD [177] were used to detect traffic violations, e.g., running a red light and making a wrong turn. Wei *et al.* [178] detected the lane change behavior of vehicles with ResNet [179] in real time. In [180], a depth-enhanced feature pyramid network was designed to identify traffic accidents in video frames. In [159], Faster-RCNN with InceptionV2 [181] and ResNet101 [179] was used to identify abnormal vehicles that were stationary for a long time in the video background.

By image context feature aggregation, the performance of behavior recognition can be improved. Leng *et al.* [182] proposed a context-aware Faster-RCNN method that takes advantage of interior and contextual features. By relying on learning the context information around obstacles, the recognition results of small and occlusion obstacles can be improved. Unlike most works that use a single-object detector, Nguyen *et al.* [164] applied multiple adaptive detectors (RetinaNet [183] used in dark scenes; Faster-RCNN used in bright scenes) to detect stationary vehicles and locate abnormal events.

In addition to vehicle detector methods, semantic segmentation methods can be adopted to extract traffic image features. In [184], three structured networks (namely, STRIPE-NET, CONTEXT-NET, and REFINER-NET) were used in a scene segmentation task to detect obstacles. This network architecture can integrate multiple spatial scale features, exploit road scene information, and preserve the details needed for small-scale obstacle detection.

c) Sequential frame temporal feature extraction: In recent works, recurrent neural networks (RNNs) have played an important role in video frame temporal feature extraction. In [185], RNNs based on a temporal attention mechanism were designed to detect vehicle lane changes. Specifically, the model was trained by using an attention layer over multiple LSTM cells.

Moreover, CNN-LSTM networks can extract features from deep CNNs and input them into temporal networks to process the spatiotemporal features. Xu *et al.* [186] classified multimodal vehicle behaviors by learning a vehicle's previous motion state based on an end-to-end FCN-LSTM network. Vehicle behaviors contain discrete actions, e.g., move straight, stop, turn left, and turn right, and continuous actions, e.g., lane following. Some works have obtained temporal features by inputting the segmentation mask of traffic videos into a temporal neural network. For example, Yurtsever *et al.* [187] sent video frames marked with a vehicle semantic segmentation mask into a CNN-LSTM network. The spatiotemporal features were extracted for vehicle lane change behavior risk analysis. Unlike most works, Xu *et al.* [188] designed a temporal recurrent network (TRN) using a spatiotemporal decoder to predict future vehicle behavior. Combining past and future vehicle behavior, the current vehicle behavior (e.g., left turn, right turn, and travel straight) can be jointly identified.

All the above methods integrate combined spatiotemporal features; we can also analyze the spatial and temporal features independently. In [189], a two-stream dynamic-attention recurrent convolutional network architecture was proposed to provide the risk level classification for each traffic video frame. The two-stream method consists of a spatial stream and a temporal stream. The spatial stream analyzes individual video frames to calculate appearance features. The temporal stream addresses the optical flow information between consecutive frames to obtain the advanced motion features. In other words, RNNs can realize the long-term memory of the corresponding vehicle state information. Thus, by extracting the traffic video temporal features, analysis of the temporal vehicle behaviors can be realized.

3) Tracking Trajectory Analysis: A vehicle's trajectory is one of the most important features in vehicle behavior recognition [190]–[192]. The position of a vehicle in a traffic video frame is determined by detecting and tracking the vehicle, and then, its trajectory can be obtained by concatenating its positions. We categorize methods based on tracking trajectory analysis into those based on statistical learning and those based on deep learning.

a) Approaches based on the statistical learning method: Using the statistical learning method, vehicle trajectories can be extracted and modeled to represent vehicle behaviors. The Dirichlet process mixture model (DPMM) is a nonparametric Bayesian model. It can infer vehicle behavior from surveillance videos without the need to specify parameters in advance. Santhosh *et al.* [193] proposed clustering vehicle trajectories to detect anomalous behaviors based on the DPMM method. Kumaran *et al.* [194] used a color gradient to represent the vehicle trajectory extracted from a video. To learn the vehicle motion pattern, the modified Dirichlet mixture model (mDPMM) was also applied to cluster the vehicle trajectories. At the same time, the t-distributed stochastic neighbor embedding method was used to separate abnormal trajectories.

The reconstruction-based method learns the vehicle motion pattern from the normal vehicle trajectories. Li *et al.* [195] performed sparse reconstruction on a vehicle trajectory and calculated the reconstruction coefficients and residuals to distinguish between normal and abnormal vehicle behaviors. In addition, the hidden Markov model (HMM) can learn the dynamic features of vehicle behavior. Aköz *et al.* [196] proposed clustering vehicle trajectories to learn normal vehicle behavior based on a continuous HMM. If the vehicle position and speed deviate greatly from those of a normal vehicle, the

vehicle is considered abnormal. All of these methods learn motion patterns from vehicle trajectories to identify vehicle behavior.

Without needing to obtain motion patterns through trajectory clustering, Wang *et al.* [197] proposed extracting motion information by the short local trajectory method based on the sparse topic model. Combined with the Fisher kernel method, motion information was quantified into visual words to represent the trajectory. Therefore, semantic information of traffic scenes can be extracted for behavior recognition. In addition, the autoencoder can conduct representation learning on traffic videos. Dinesh *et al.* [198] employed the intersection points of different vehicle trajectories, combined with the feature representation learned from the deep stacked autoencoder, to detect vehicle accidents. However, the performance of this strategy is not satisfactory because of the complex traffic scenes and the traffic pattern changes at the intersections. In other words, these works adopt the strategy of clustering or reconstructing vehicle trajectories so that they can learn the motion patterns of trajectories to understand vehicle behaviors.

b) Approaches based on the deep learning method: Vehicle tracking methods based on deep learning can easily obtain the vehicle trajectory and then analyze vehicle behavior through these trajectories [199], [200]. Wang *et al.* [201] employed the TrackletNet tracker (TNT) [202] to extract the trajectory of anomalous vehicles and estimate the start time of abnormal behavior, e.g., vehicle accidents and breakdown stopping. In [163], an MOT algorithm combined with a single-object tracker (SOT) module was designed to extract anomalous vehicle trajectories. In [136], after simple online real-time tracking (SORT) was used to obtain vehicle trajectories, the vehicle trajectory classification method of contrastive likelihood estimation (CPLE) based on ML estimation was applied to detect vehicle accidents. To integrate vehicle detection results, Huang *et al.* [135] proposed a two-stream convolutional network architecture to jointly detect vehicle accidents in real time. The spatial stream network uses appearance features to detect accident vehicles. The temporal stream network utilizes the temporal features to generate vehicle trajectories by MOT. In addition, LSTM can carry out follow-up processing of vehicle trajectories to realize behavior recognition. In [203], after vehicle trajectories were modeled by tracking a vehicle in the whole video frame sequence, the bidirectional long short-term memory network (BiLSTM) was used to recognize vehicle behaviors, including left turns, right turns, straight motion, etc.

Unlike tracking the trajectory of each vehicle to capture vehicle motion, we can also preset behavior rules based on the normal trajectories. When a vehicle trajectory deviates from the rule, the vehicle is considered abnormal. Makhmutova *et al.* [204] used the kernelized correlation filter (KCF) tracking algorithm [205] to track vehicles and obtain vehicle trajectories compared with normal trajectories. However, this strategy assumes the spline approximation of frequent trajectories as a vehicle motion pattern. To solve this problem, Xu *et al.* [158] proposed using both static and dynamic vehicle motion patterns. In the static motion pattern, abnormal stationary vehicles are detected by the Faster-RCNN object detector. The dynamic motion pattern actively extracts the trajectories of all moving vehicles on the road and clusters the trajectories to find the main motion mode. Afterward, any vehicle motion differing from this movement mode is identified as abnormal. In deep-learning-based methods, global scene information is employed for feature learning to improve the quality of trajectory extraction. Therefore, their performance is better than that of statistical-learning-based methods in adapting to different traffic scenes.

4) Quantitative and Qualitative Results: From 2018 to 2021, AI CITY CHALLENGE included traffic anomaly detection tasks every year. Traffic anomalies include single or multiple vehicle crashes and stalling. The 2021 challenge, for example, included 100 training videos and 150 test videos. There were 18 anomalies in the 100 training videos. The evaluation metrics of this task are the F_1 score and the $NRMSE_t$.

F_1 is the harmonic mean value of the anomaly detection accuracy and recall rate, and $NRMSEt$ is the $RMSE$ between the real anomaly start time and the anomaly start time detected by the model. In Table VII, we list the representative works in every year.

In short, most anomaly detection work is based on preprocessing methods. First, background modeling is carried out to retain stationary vehicles, and then, a road mask is modeled to remove the influence of off-road vehicles. Finally, the trajectories of the abnormal vehicles are tracked back to determine the start time. The accuracy of these methods depends on the performance of the vehicle detection and tracking algorithms. Among them, the detection and tracking of distant small vehicles are particularly critical.

In the abnormal behavior dataset [213], the duration of most videos is 1–5 minutes, and the total duration of the dataset is 128 hours. Abnormal behaviors include abuse, arrest, and traffic accidents. Agrawal *et al.* [169] selected 32 videos for the experiment, including 16 accident videos and 16 nonaccident videos. The relevant experimental results are listed in Table VIII.

5) Discussion: In vehicle behavior recognition, methods based on motion feature extraction extract vehicle motion features from traffic videos. In traffic monitoring scenes, most work can avoid vehicle accidents through anomaly detection, trajectory analysis, lane change detection, and so on. Complex vehicle behavior classification and recognition should be combined with trajectory analysis. Methods based on video frame object classification combine video spatial and temporal features and are thus suited only for simple vehicle behavior recognition in sparse traffic scenes. Currently, tracking-trajectory-based methods are the mainstream methods for vehicle behavior recognition and have wide applicability.

B. Traffic Agent Interaction Behavior Recognition

In addition to recognizing the behavior of a single vehicle, the complex spatiotemporal interaction behavior between vehicles and other vehicles, as well as between vehicles and other traffic agents, should be considered. We classify the related works into the following two categories. Methods based

TABLE VII
REPRESENTATIVE WORKS OF AI CITY CHALLENGE TRAFFIC ANOMALY DETECTION

Feature Extraction Approaches	Classification	Methods	Score	Year
Deep learning based	Track vehicle trajectory	Zhao et al.[206]	0.9355	2021
		Wu et al.[207]	0.9220	2021
		Chen et al.[208]	0.9197	2021
Background separation	Extract background region	Doshi et al.[209]	0.8597	2021
Deep learning based	Track vehicle trajectory	Li et al.[210]	0.9695	2020
Background separation	Extract background region	Doshi et al.[211]	0.5763	2020
	Generate road mask region	Shine et al.[212]	0.5438	2020
Background separation	Generate road mask region	Bai et al.[165]	0.9534	2019
Deep learning based	Track vehicle trajectory	Wang et al.[201]	0.9362	2019
		Chang et al.[180]	0.6997	2019
Background separation	Generate road mask region	Zhao et al.[163]	0.6585	2019
		Nguyen et al.[164]	0.6129	2019

TABLE VIII
THE EXPERIMENTAL RESULTS BASED ON THE ABNORMAL BEHAVIOR DATASET

Category	Classification	Methods	Average Accuracy	Year
Motion Feature Extraction Based	Dense optical flow	Maaloul et al.[133]	91.00%	2017
Frame Object Classification Based	CNN	Long et al.[214]	94.00%	2018
		Chen et al.[215]	81.20%	2019
	CNN-LSTM	Ghosh et al.[216]	92.38%	2019
	CNN	Shareef et al.[217]	86.00%	2020
		Agrawal et al.[169]	94.14%	2020

on interactive dynamic features capture the real-time vehicle status (speed, acceleration, yaw rate, etc.). Then, statistical learning or deep learning methods encode and extract the motion features associated with vehicle status. Methods based on the interaction model utilize the dynamic spatiotemporal dependency relationship between the vehicles and other traffic agents to build the interaction model. In Table XI, we summarize the popular methods used for interaction behavior recognition.

1) Approaches Based on Interactive Dynamic Features: By considering the vehicle motion state change between driving normally and interacting with other traffic agents, we can realize multiple traffic agent interaction behavior recognition. Combined with the geometric, kinematic, and behavioral constraints of heterogeneous traffic agents, Luo *et al.* [218] proposed integrating the three constraints into a constraint geometry optimization framework. Therefore, the trajectory change of a traffic agent caused by the interaction behavior can be deduced. Yu *et al.* [219] suggested projecting vehicles into 3D Euclidean space to calculate the real vehicle velocity and acceleration change through vehicle trajectories. Then, vehicle accidents due to vehicle interactions can be detected by motion changes. Similarly, Ijjina *et al.* [220] utilized the change in three traffic motion parameters, i.e., acceleration, trajectory, and angle, to observe substantial changes in the vehicle velocity and position at the vehicle interaction moment. Constrained by a high-density traffic flow, this strategy generalizes poorly. In [149], traffic scene motion information was used to generate the MIF to obtain the strength and direction of vehicle interactions. This algorithm observes the overall spatiotemporal state change of a traffic video frame and is simple to implement but also limited by the scene density.

At the same time, the trajectory features [198] extracted from the spatiotemporal video volumes and the features extracted from the 3DCNN [213] are scalable to interaction behavior recognition. Recent works [186], [188], [221] encoded visual information into the CNN-LSTM network to learn the motion state of each traffic agent. For example, Ramanishka *et al.* [221] proposed extracting and integrating the vehicle dynamic status (speed, accelerator, yaw rate, etc.) into the LSTM network to better understand the multiple vehicle behaviors (e.g., left turn, right turn, crosswalk passing, and lane change).

2) Approaches Based on the Interaction Model: In traffic scenes, vehicle behavior is easily affected by nearby traffic agents, e.g., the change in distance and speed between two vehicles will make a vehicle stop or take a turn. In this case, the vehicle interaction state can be represented by the vehicle's relationship with nearby traffic agents. With the development of vehicle-tracking-related and trajectory-analysis-related

TABLE IX
SUMMARY OF VEHICLE BEHAVIOR RECOGNITION METHODS

Methods	Feature Extraction Approaches	Specific Implementation Methods	Applications
Motion Feature Extraction Based	Optical Flow	Sparse optical flow[149, 151–153]	Anomaly detection, traffic detection
		Dense optical flow[154, 156, 157]	
	Background separation	Extract background region[158–160]	
		Generate a road mask region[163–165]	
Frame Object Classification Based	Feature dimension reduction of video data	PCA[170]	Lane change
		Manifold learning[171]	
		Low-rank matrix[172]	Traffic accident
		Perceptual video summarization[173]	Anomaly detection
	Feature ectraction of video frame image	Sparse autoencoder[174]	Lane change
		topic model[175]	Anomaly detection
		CNN[159, 164, 176, 178, 180, 182, 184]	Traffic violations, lane change, traffic accidents
	Temporal feature extraction of continuous frame	CNN-LSTM[186–188]	Lane change, take a turn
Tracking Trajectory Based	Statistical learning methods	DPMMs[193, 194]	Traffic accident, anomaly detection, take a turn, stop
		Sparse reconstruction[195]	
		HMMs[196]	
		Sparse topic model[197]	
		AutoEncoder[198]	
	Deep learning based	Vehicle trajectory[135, 136, 201, 203]	
		Pre-set normal behavior rules[158, 204, 205]	

works, this method can be implemented to analyze vehicle interaction behaviors by modeling the relationship between vehicles. In [222], Roy *et al.* designed the Siamese interaction long short-term memory network (SILSTM) to learn vehicle interaction trajectories and combined it with a collision energy model to detect the vehicle interaction result.

GNNs can also be employed to understand the interaction behaviors by taking the traffic agents as the nodes and the interaction relationships as the edges. In [223], spatiotemporal action graph networks (STAGs) were used to detect vehicle accidents based on the motion relationship between vehicles. The network model consists of two parts: the spatial structure models the relationship between traffic agents, and the temporal structure aggregates the traffic video temporal context. In contrast to [223], which modeled spatial and temporal structures separately, Li *et al.* [224] applied two GCN networks (Ego-Thing Graph and Ego-Stuff Graph). The overall interaction scene was divided into the interaction between vehicles and traffic agents (vehicles, pedestrians, etc.) and the interaction between vehicles and road infrastructure (lane markers, signal lights, etc.). The Ego-Thing network and the Ego-Stuff network jointly focus on recognizing vehicle behaviors (lane change, left turn, right turn, stop for crossing pedestrians, etc.). By modeling to comprehensively utilize the spatiotemporal features between vehicles, the vehicle interaction behavior can be fully understood.

3) Quantitative and Qualitative Results: The HDD dataset [221] contains approximately 104 hours of vehicle first-person-view driving videos. The videos contain different driving scenes, such as urban, suburban, and highway scenes. In these scenes, there is a complex interaction between vehicles and other vehicles and between vehicles and pedestrians on the road, e.g., the vehicles perform right turns, left turns, lane changes, intersection passing, crosswalk passing and other behaviors. By detecting these behaviors, the effectiveness of interactive behavior recognition models can be evaluated.

Table X compares some methods, with the data taken from [224]. Compared with C3D and I3D, which use 3D CNNs to expand the temporal dimension and extract the spatiotemporal features, GCN methods model the interaction between traffic agents in a scene to achieve better behavior recognition.

4) Discussion: Methods based on interactive dynamic features are simple to implement and require extracting only the status of each vehicle in a traffic video and observing the vehicle motion changes. However, the real-time vehicle state

TABLE X
TRAFFIC AGENT INTERACTION BEHAVIOR RECOGNITION RESULTS BASED ON THE HDD DATASET

Classification	Methods	Intersection passing	Left turn	Right turn	Left lane change	Right lane change	Crosswalk passing	Average mAP
Vehicle Motion Learning	CNN[221]	53.4	47.3	39.4	23.8	17.9	4.8	20.7
	CNN-LSTM[221]	65.7	57.3	54.4	27.8	26.1	16.0	26.9
	ED[188]	63.1	54.2	55.1	28.3	35.9	7.1	27.2
	TRN[188]	63.5	57.0	57.3	28.4	37.8	11.0	29.7
	DEPSEG-LSTM[225]	70.9	63.4	63.6	48.0	40.9	16.1	33.7
3D CNN	C3D[226]	82.4	77.4	80.7	67.9	56.9	17.4	45.5
	I3D[227]	85.6	79.1	78.9	74.0	62.4	29.8	49.5
GNN	GCN[224]	85.5	77.9	79.1	76.0	62.0	29.6	51.1

is uncertain. The motion information cannot be represented accurately and can change at any time, resulting in the inability to accurately recognize vehicle behaviors. The methods based on the interaction model consider the relationship between vehicles, and they perform well for heterogeneous traffic scenes. In fact, uncertain multimodal interactive behavior is a major challenge to study. The static infrastructure of a traffic scene, which plays a guiding role in vehicle interactions (e.g., signal lights and sidewalk), should also be considered.

C. Complex Temporal Reasoning of Vehicle Behavior

Temporal reasoning refers to the short-term or long-term prediction of vehicle behavior, e.g., lane change, parking, turning, and vehicle accidents [228]. The future behavior of vehicles can be predicted through a series of changes in vehicle appearance and motion features in temporal sequences [229], [230]. We classify the temporal reasoning approaches into physical-model-based, traffic-agent-model-based, and decision-model-based approaches. In Table XIV, we list the representative works of complex temporal reasoning in each category.

1) Approaches Based on a Physical Model: Based on a vehicle kinematics model, the vehicle motion state (speed, acceleration, yaw rate, etc.) can be predicted in the physical model methods, and then, the vehicle behaviors can be predicted. Deo *et al.* [231] proposed utilizing a Bayesian filter to integrate constant velocity, constant acceleration, and constant turn rate into three motion models to obtain motion state information. Then, the spatiotemporal motion information change is captured based on the HMM to predict vehicle motion. In [218], vehicle motion prediction was transformed into a constrained geometric optimization problem in velocity space by calculating the instantaneous vehicle velocity. However, these methods are easily affected by different traffic scenes. They cannot accurately predict all vehicle motion in a traffic scene. To solve this problem, many works are based on classifiers such as those in SVMs [232], [233], LSTMs [234] and Bayesian networks [235]. However, due to the uncertainty of the current vehicle motion state, the prediction of the future vehicle motion state and trajectory is not accurate, especially for a long duration.

2) Approaches Based on a Traffic Agent Model: Traffic agent models build interaction relationships to predict multiple future traffic agent behaviors. These models depend mainly on LSTM and GNN methods.

LSTM-based methods model the traffic agent interaction relationship by capturing the spatiotemporal information. Because a future traffic agent is affected by the surrounding traffic agents and its nearby static obstacles, Gupta *et al.* [236] introduced the adversarial feature learning of the GAN into the behavior prediction task. They also proposed adopting a pooling mechanism to aggregate the surrounding information to reduce the computing requirement of the LSTM network. The attention mechanism can learn the part that needs to focus on learning in behavior prediction. Based on the relative importance of each pedestrian in a crowd, Vemula *et al.* [237] predicted the future position of each pedestrian by considering mechanical attention. In heterogeneous traffic scenes, Chandra *et al.* [238] utilized the CNN-LSTM network to predict the traffic agent location combined with the differences in appearance, dynamic motion information, and behavior of different traffic agents. Since static objects can also affect traffic agent trajectories, Haddad *et al.* [239] proposed a spatiotemporal graph network based on an LSTM network. This network predicts interactive trajectories with the influence of static objects and dynamic objects. However, these LSTM-based methods perform poorly in large-scale heterogeneous traffic scenes.

GNNs can effectively describe the topology of traffic agents. Chandra *et al.* [240] used a dynamic weighted traffic graph to represent the interaction between traffic agents. By modeling a dual-stream graph-LSTM network, they predicted the long-term trajectory and vehicle behaviors (such as speeding, normal driving, and braking). In the second stream, the spectral clustering method was applied to predict vehicle behavior with motion feature vectors. In [241], the distance and interaction between traffic agents, time series, and high-level categorization features were constructed into a 4D graph. The previous state of the traffic agents was learned to predict the trajectory. The performance of graph-based methods is better than that of LSTM-based methods because graph-based methods can exploit the spatiotemporal features and interactions of traffic agents for in-depth modeling.

Similar to the interaction model, the traffic agent model also takes the vehicle motion state as the input. The difference is that traffic-agent-model-based methods learn from past and current vehicle behaviors and predict future behaviors.

3) Approaches Based on a Decision Model: Unlike the traffic agent model, the decision model simplifies the temporal reasoning task for sequence prediction based on RNNs [242]. The past and current motion states of a single traffic agent are

TABLE XI
SUMMARY OF INTERACTION BEHAVIOR RECOGNITION METHODS

Methods	**Classifiers**	**Descriptions**	Applications
Dynamic Features Based	Change in Motion Parameters[218–220]	The change in the motion state between interaction vehicles is detectd	Traffic accident
	Motion Interaction Field[149]	The MIF is used to obtain the strength and direction in vehicle interactions	Traffic accident
	Vehicle Motion Learning[186, 188, 221]	The CNN-LSTM network learns the motion state of each vehicle based on its previous motion state	Left turn, right turn, crosswalk passing, or lane change
Interaction Model Based	LSTM[222]	The interaction trajectories of vehicles are analyzed by temporal modeling	Traffic accident, vehicle stop or make a turn
	GNN[223, 224]	The spatiotemporal interaction features of vehicles and other traffic agents in traffic scenes are aggregated	Traffic accident, lane change, left turn, right turn, stop for crossing pedestrians

used as the training data to extract the sequence features. Many works propose hybrid network architectures based on RNNs to achieve vehicle behavior prediction. For example, after obtaining prediction samples by a conditional variational autoencoder (CVAE), Lee *et al.* [243] proposed extracting semantic scene context and historical vehicle motion state information based on RNNs for trajectory prediction. The common strategy mostly used in present works is the CNN-LSTM architecture. Combined with the nonlinear motion model and the instance segmentation algorithm, the CNN-LSTM architecture was used to predict vehicle trajectories in [244].

In addition, some works use the decision model to predict vehicle accidents. Yao *et al.* [245] proposed extracting spatiotemporal features through an RNN encoder-decoder model based on future object localization (FOL) [246]. The position of a future traffic agent can be output to predict the occurrence of vehicle accidents. With the attention in RNNs, the method can focus on learning the motion features of the traffic agent. Chan *et al.* [247] applied a dynamic spatial attention recurrent neural network (DSA-RNN) to identify the motion changes in vehicle behavior before and after the occurrence of accidents. By using dynamic spatial attention LSTM to allocate attention, Shah *et al.* [248] made a more accurate prediction than is possible based on [247]. Instead of using attention to train RNNs, Suzuki *et al.* [249] designed an adaptive loss for early anticipation (AdaLEA) method to dynamically change the loss value during the whole vehicle motion process.

4) Quantitative and Qualitative Results: The Dashcam Accident Dataset (DAD) [247] contains multiple traffic videos at 720p resolution and 20 frames per second captured in Taiwan. This dataset includes 1,130 normal videos and 620 vehicle accident videos. Each video is made up of 100 frames (5 seconds), and vehicle accidents always occur in the last 10 frames. The perceived ability of the model to recognize traffic accident risk is represented by the average precision (AP), and the ability to predict an accident as early as possible is represented by the mean time to accident (mTTA). Table 13 shows the evaluation results of works on the DAD.

The ETH [253] and UCY [254] datasets use a bird's eye view to observe the pedestrian interaction scene to predict a pedestrian's movement and location. Typically, the first 3.2 seconds of each video are used as the input data, and the future trajectory after 4.8 seconds is predicted. The evaluation metrics include the average displacement error (ADE) to measure the accuracy of the entire trajectory and the final displacement error (FDE) to measure the accuracy of the endpoint of the trajectory.

As shown in Table XIV, these recent works focus on the temporal and spatial associations of agent motion states. By integrating the attention mechanism into the RNN and GCN, the target motion state and surrounding target motion state can be captured better, improving the trajectory prediction accuracy.

5) Discussion: Complex temporal reasoning can be used to predict vehicle trajectories and future behaviors. Methods based on the physical model can use vehicle motion features to predict future motion. Compared with the traffic agent model, these methods lack interactions with other traffic agents. When the movement of a vehicle changes (brake, slow down, turn, etc.), prediction error is likely to occur. The traffic agent model considers the global interaction so that it can predict the trajectory over a long period. However, modeling the interactive relationship between multiple traffic agents leads to a complex calculation. Furthermore, approaches based on decision models are often used to predict vehicle trajectories and accidents in sparse traffic scenes.

D. Conclusion

We introduce vehicle behavior recognition methods that cover most behaviors in the process of vehicle driving. Only

TABLE XII
TRAFFIC ACCIDENT PREDICTION BASED ON THE DAD

Category	Methods	Description	AP(%)	mTTA(s)	Year
Decision Model Based	DSA[247]	Identify vehicle motion changes before and after an accident	48.1	1.34	2016
Traffic Agent Model Based	L-RAI[250]	Model spatial and appearance interactions between traffic agents using RNNs	51.4	3.01	2017
Decision Model Based	AdaLEA[249]	Predict accidents early by considering the loss function adaptive weight distribution	52.3	3.43	2018
Traffic Agent Model Based	GCRNN[251]	Use GCNs and RNNs to learn the relationship of feature representations, and use the Bayesian neural network to further focus on their intrinsic variability	53.7	3.53	2020
Decision Model Based	DSTA[252]	Use dynamic temporal attention to select the time period of an accident, and use dynamic spatial attention to learn the spatial region in each frame	56.1	3.66	2021

TABLE XIII
PEDESTRIAN TRAJECTORY PREDICTION WORKS USING THE ETH AND UCY DATASET

Classification	Methods	Description	ADE↓	FDE↓	Year
LSTM	Social-LSTM[242]	Pool the LSTM hidden states of surrounding agents to extract interactive information	0.72	1.54	2016
	Social-GAN[236]	Use the pooling module to learn global pooling features	0.74	1.54	2018
	MATF[255]	Model the agent interactions and constraints based on multiagent tensor fusion	0.64	1.26	2019
	FvTraj[256]	By rendering first-person-view images, capture the relations between historical motion states and visual features	0.50	1.04	2020
GNN	STAR-D[257]	Use the transformer-based graph convolution mechanism to model intragraph crowd interactions	0.41	0.87	2020
	Trajection++[258]	Use a graph-structured recurrent model combining agent dynamics and heterogeneous data	0.37	0.91	2020
Encoder-Decoder	SGNET[259]	Use multiple temporal scales to capture historical information and predict the target trajectory based on n encoder and a decoder module	0.35	0.83	2021

simple vehicle behavior (lane change, vehicle accidents, etc.) can be recognized based on a single traffic video frame. Combined with the temporal features in a continuous traffic video frame, we can realize the accurate classification and recognition of various vehicle behaviors. In addition, to analyze the interaction behavior, it is necessary to combine the dynamic motion information and topological relationship between vehicles. In complex temporal reasoning, accurate vehicle behavior prediction can be achieved by comprehensively considering the spatiotemporal motion information of traffic agents.

IV. DISCUSSION OF FUTURE DEVELOPMENT

In the previous sections, we reviewed the representative works of traffic semantic understanding. However, there are still many unsolved problems. Here, we offer some perspectives on the future development of traffic semantic understanding.

A. Depth Information Missing in Traffic Scenes

Although traffic videos have a wide field of view, they can offer a variety of traffic parameters. However, because they contain only two-dimensional data, the depth information of traffic scenes is missing. As a result, the accuracy is reduced due to occlusions, shadows and other problems in vehicle detection and tracking algorithms. This problem can be solved in two ways:

1) Vision-Based Depth Estimation Methods: Missing depth information occurs because a monocular camera generates only two-dimensional information [260]. Monocular 3D vehicle detection methods [261], [262] can use the geometric constraint relationship between 2D and 3D bounding boxes, multiple coordinate point regression, and rotation angle regression based on 2D vehicle detection. However, the detection accuracy is not as good as that of the stereo vision detection methods. Due to the large detection range of a traffic scene, an active light cannot be cast onto an entire road section. However, stereo vision detection methods based on natural light can be used [263], [264]. In general, solutions based on stereo vision can be divided into two kinds: point cloud models and road projection models.

- Point cloud model needs to reconstruct the traffic scene [265] through point cloud data and then obtain the three-dimensional coordinates of the traffic agents [263] through the 3D object detector.

TABLE XIV
MAIN METHODS OF COMPLEX TEMPORAL REASONING

Models	Classifiers	Descriptions
Physical Model Based	HMMs[231]	Use a combination of temporal and spatial motion information to predict vehicle trajectories
	SVMs[232, 233]	Predict future vehicle motion based on the SVM classifier
	Bayesian network[235]	Based on the prior motion information, obtain reliable vehicle motion prediction results
Traffic Agent Model Based	LSTM[236–239]	Model the spatiotemporal interaction of traffic agents for sequence prediction
	GNN[240, 241]	Build traffic graphs through the topology of traffic agents to predict the vehicle trajectory and behavior
Decision Model Based	Vehicle motion state prediction[242–244]	Extract vehicle motion features by using the CNN-LSTM network for prediction
	Traffic accident prediction[245–249]	Predict the motion state and combine with the attention mechanism to identify and locate the vehicle involved in the accident

- Road projection model projects point cloud data [266] or disparity map data [267] onto the reference plane and detects the front view or bird's eye view. Road projection methods can reduce the three-dimensional detection model to two-dimensional space, and analysis and detection can be realized through the two-dimensional image method. This approach can meet the 3D detector requirement of real-time detection.

2) Multisensor Joint Detection: The drawbacks of using only visual data are that monocular vision cannot provide reliable 3D geometric information. Although binocular vision can provide reliable 3D geometric information, its performance is not good under a high degree of occlusion and poor illumination conditions [268]. To efficiently obtain reliable depth data, multisensor data fusion among cameras, radar, LiDAR, etc. needs to be realized. Van *et al.* [269] combined RGB images and LiDAR point clouds to generate dense depth maps of traffic scenes from sparse and irregular point clouds. Long *et al.* [270] associated radar and camera pixels to learn mapping relationships between radar and pixels to achieve depth completion. Experimental results show that it is better to obtain traffic scene depth based on multisensor fusion than using only cameras.

B. Traffic Semantic Analysis and Prediction of a Complex Road Network

Traffic monitoring information is employed to analyze the current road congestion situation and predict future congestion by the properties of the traffic flow, which can provide a basis for traffic scheduling [271], [272]. However, the traffic analysis of a single intersection or section cannot be extended to the overall road network.

1) Road Network Traffic Flow Analysis Method: At present, traffic flow prediction algorithms based on vision methods focus only on a single road in the road network. However, for the whole urban traffic system, all roads influence each other in the road network. Most traffic prediction methods use traffic sensor data, e.g., GPS, and divide the road network into traffic grids or graphs. Guo *et al.* [273] proposed regarding the road network as a grid and automatically capturing the correlation of each section of the road network based on a 3D CNN for traffic flow prediction. Unlike [273], the road network is defined as an undirected graph in [274]. Graph convolution and standard convolution are jointly used together to obtain the spatiotemporal information of the road network and predict the traffic flow. To the best of our knowledge, no work has used visual information to predict road network traffic flow. The main reason is that there are fewer vision datasets for road network traffic flow research. At present, road network works based on visual information mostly use self-built datasets. The STREETS dataset [275] includes more than 4 million images of different road sections as well as information on the road topology. Directed traffic graphs can be determined by using these data, and spatiotemporal information can be obtained for traffic flow prediction.

2) Relationship Between Macro Traffic Flow and Micro Vehicle Behavior: Vehicle model differences and multimodal road composition cause many problems in traffic flow operation. For example, the traffic volume is still far less than the designed capacity of the road, but traffic congestion is frequent in actual operation. The main reason for these problems is that there are various groups with different characteristics in the traffic flow. The mutual interference between the road groups makes the traffic flow a mixed flow with internal differences and inhomogeneity. The congestion degree of traffic flow will affect individual road vehicle behaviors [276], [277], while micro vehicle behavior will affect the overall traffic flow state [278]. However, there are few studies on the relationship between macro traffic flow and micro

vehicle behavior obtained by vision-based methods. For this reason, it is necessary to comprehensively consider the feedback relationship between micro vehicle behavior and macro traffic flow in the future. Therefore, the analysis and prediction model can more accurately reflect the dynamic features of traffic flow.

C. Limitations of Traffic Semantic Training Samples

Unlike other computer vision tasks (object detection, tracking, semantic segmentation, etc.), vehicle behaviors vary. However, the relevant datasets can contain only one or several related behaviors, e.g., vehicle crashes, lane changes, and turns. At the same time, the existing algorithms cannot perform a detailed analysis of all vehicle behaviors in an entire traffic scene. For this reason, we propose using several methods to break through the constraint on the training samples:

1) Unsupervised and Semisupervised Learning: Unsupervised learning methods [71], [165], [204] do not need to label the dataset but learn according to the distribution rules of abnormal vehicle behaviors on highways or intersections [191], [211]. To this end, the strategy automatically determines the rules of normal behaviors, and all behaviors deviating from these rules are defined as abnormal. Semisupervised learning methods [136], [194], [197] utilize large amounts of unlabeled traffic videos as well as less labeled vehicle behavior data to extract behavior features.

2) Few-Shot Learning: Few-shot learning can reduce the difficulty of collecting training samples and has become a new direction in object detection. In [279], infrared cameras were used to detect the forward traffic environment. Due to the difficulty of collecting infrared samples, few-shot learning was adopted. Specifically, this method first trains the network by mature visible images and then fine-tunes the network parameters by using all infrared samples. In addition to few-shot learning, zero-shot learning can also be applied to vehicle behavior recognition. Yu *et al.* [101] adopted the zero-shot learning method to regress the vehicle trajectory to a predefined vehicle driving trajectory without labeling the vehicle trajectory. As it is usually difficult to acquire and label vehicle behaviors in traffic scenes, few-shot learning, one-shot learning, and zero-shot learning can be used to promote the research progress of vehicle behavior recognition.

3) Focus on Different Traffic Agents by the Attention Mechanism: Because the types and quantities of traffic agents in sparse and dense traffic scenes are very different, the vehicle behavior analysis algorithm cannot be applied to various traffic scenes (such as busy intersections, freeways, and urban roads) at the same time. Moreover, the interaction between traffic agents in heterogeneous traffic scenes should be considered. The attention mechanism can capture the dynamic spatial correlation between the positions of the traffic agent, as well as the temporal correlations between motion states [237]. Currently, algorithms [178], [179] that add an attention mechanism are mostly based on LSTM or GNNs.

D. Optimal Computational Efficiency

Theoretically, the ITSs algorithm needs to be used in real-time to recognize various vehicle behaviors efficiently. However, system cannot run in real time due to the time required for data acquisition, data transmission, and model operation. Next, we mainly analyze the model operational efficiency:

In macro traffic flow, the methods of AI CITY CHALLENGE 2021 Task 1 [94]–[96] can count vehicle turns in real time on IoT devices. In addition, we believe that speed, flow volume and density should be analyzed in real time. However, computational efficiency analysis is lacking in most of the works discussed in this paper.

In micro vehicle behavior analysis, it is necessary to analyze and predict the behavior of each vehicle based on vehicle detection and tracking, which may also involve modeling the entire traffic scene as a graph for analysis.

E. Future Work

For missing depth information, by learning geometric information in 3D space, the 3D features of real traffic agents can be represented. However, due to the restriction of the receptive field, the current 3D object detectors cannot learn local features at different scales very well [280], [281]. Even though the multisensor fusion method can also obtain the depth information and dynamic motion information of traffic scenes, the fusion of multisource heterogeneous data is a bottleneck hindering the adoption of the multisensor joint detection method. Therefore, using multisensor fusion to quickly and effectively obtain the depth information of road scenes is of great significance to achieve more accurate quantification of traffic semantics.

For traffic semantic analysis and prediction of a complex road network, cameras are widely distributed on the road network. In past works, vision-based methods have made some progress in addressing traffic behavior understanding at a single road intersection or section. In future work, this will be an important research direction. By constructing a dataset of road network videos and using the road topology and spatiotemporal information, we can analyze and predict the traffic semantics of a road network.

Due to the limitations of traffic semantic training samples, current research works on vehicle behavior recognition are limited by the types of vehicle behaviors labeled in the datasets. Due to the complexity of traffic scenes (different road types, weather conditions, etc.), it is impossible to understand the various behaviors fully and accurately. Few-shot learning can reduce the difficulty of collecting vehicle behavior samples such that a few vehicle behavior samples can be used to achieve comprehensive vehicle behavior recognition. The attention mechanism can focus on traffic agent behavior and realize an efficient understanding of various traffic agent behaviors. The unsupervised and semisupervised methods can not only reduce labeling costs but also improve the generalization ability of recognizing vehicle behaviors in different traffic scenes.

For optimal computational efficiency: We suggest that future deep-learning-based works optimize the model architecture by using compression techniques, learning techniques, automation (such as hyperparameter optimization and architecture

searches) and efficient architectures (parameter sharing, attention mechanism, etc.) to improve the computational efficiency [282].

V. Conclusion

Vehicle surveillance based on road traffic semantic understanding is an important part of ITSs. Based on the appearance and motion information of a vehicle, the behavior is analyzed and predicted so that the overall road driving environments can be understood in time. In this paper, we reviewed the latest literature. First, we classified road behavior into two parts for learning. (1) Macro spatiotemporal feature quantification estimates and understands the three most important features of the overall traffic flow. (2) Micro road behavior recognition jointly studies the vehicle behavior of traffic agents and the complex temporal characteristics of roads. Finally, we elaborate on the challenges for vehicle semantic understanding and introduce future research priorities. We hope that further work will promote vehicle behavior understanding based on road visual information to provide a solid foundation for the development of ITSs.

References

[1] M. Naphade *et al.*, "The 4th AI city challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 626–627.

[2] M. Naphade *et al.*, "The 5th AI city challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4263–4273.

[3] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," in *Proc. AAAI*, vol. 34, no. 1, 2020, pp. 914–921.

[4] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Deep learning on traffic prediction: Methods, analysis and future directions," 2020, *arXiv:2004.08555*.

[5] Y. Wang, Z. Jiang, X. Gao, J.-N. Hwang, G. Xing, and H. Liu, "RODNet: Radar object detection using cross-modal supervision," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 504–513.

[6] X. Weng, Y. Wang, Y. Man, and K. M. Kitani, "GNN3DMOT: Graph neural network for 3D multi-object tracking with 2D-3D multi-feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6499–6508.

[7] R. Nabati and H. Qi, "CenterFusion: Center-based radar and camera fusion for 3D object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1527–1536.

[8] C. Fu, B. Li, F. Ding, F. Lin, and G. Lu, "Correlation filters for unmanned aerial vehicle-based aerial tracking: A review and experimental evaluation," 2020, *arXiv:2010.06255*.

[9] D. Feng, A. Harakeh, S. Waslander, and K. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," 2020, *arXiv:2011.10671*.

[10] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3152–3168, Aug. 2020.

[11] S. Felici-Castell, M. García-Pineda, J. Segura-Garcia, R. Fayos-Jordan, and J. Lopez-Ballester, "Adaptive live video streaming on low-cost wireless multihop networks for road traffic surveillance in smart cities," *Future Gener. Comput. Syst.*, vol. 115, pp. 741–755, Feb. 2021.

[12] A. A. Brincat, F. Pacifici, S. Martinaglia, and F. Mazzola, "The Internet of Things for intelligent transportation systems in real smart cities scenarios," in *Proc. IEEE 5th World Forum Internet Things (WF-IoT)*, Apr. 2019, pp. 128–132.

[13] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, "CNN-based density estimation and crowd counting: A survey," 2020, *arXiv:2003.12783*.

[14] D. F. Llorca, A. H. Martínez, and I. G. Daza, "Vision-based vehicle speed estimation: A survey," *IET Intell. Transp. Syst.*, vol. 15, no. 8, pp. 987–1005, Aug. 2021.

[15] K. K. Santhosh, D. P. Dogra, and P. P. Roy, "Anomaly detection in road traffic using visual surveillance: A survey," *ACM Comput. Surv.*, vol. 53, no. 6, pp. 1–26, Nov. 2021.

[16] N. Buch, S. A. Velastin, and J. Orwell, "A review of computer vision techniques for the analysis of urban traffic," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 3, pp. 920–939, Mar. 2011.

[17] Y. Liu, B. Tian, S. Chen, F. Zhu, and K. Wang, "A survey of vision-based vehicle detection and tracking techniques in ITS," in *Proc. IEEE Int. Conf. Veh. Electron. Saf.*, Jul. 2013, pp. 72–77.

[18] L. Chen, S. Lin, X. Lu, D. Cao, and F. Y. Wang, "Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3234–3246, Jun. 2021.

[19] Z. Wang, J. Zhan, C. Duan, X. Guan, P. Lu, and K. Yang, "A review of vehicle detection techniques for intelligent vehicles," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 5, 2022, doi: 10.1109/TNNLS.2021.3128968.

[20] X. Ma, W. Ouyang, A. Simonelli, and E. Ricci, "3D object detection from images for autonomous driving: A survey," 2022, *arXiv:2202.02980*.

[21] J. Wang *et al.*, "A survey on driver behavior analysis from in-vehicle cameras," *IEEE Trans. Intell. Transp. Syst.*, early access, Nov. 17, 2021, doi: 10.1109/TITS.2021.3126231.

[22] J. Li, Z. Xu, L. Fu, X. Zhou, and H. Yu, "Domain adaptation from daytime to nighttime: A situation-sensitive vehicle detection and traffic flow parameter estimation framework," *Transp. Res. C, Emerg. Technol.*, vol. 124, Mar. 2021, Art. no. 102946.

[23] Z. Shi, Y. Chen, and P. Ma, "Video data based traffic state prediction at intersection," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–6.

[24] P. Burnos, J. Gajda, P. Piwowar, R. Sroka, M. Stencel, and T. Zeglen, "Measurements of road traffic parameters using inductive loops and piezoelectric sensors," *Metrol. Meas. Syst.*, vol. 14, no. 2, pp. 187–203, 2007.

[25] P.-E. Mazaré, O.-P. Tossavainen, A. Bayen, and D. Work, "Trade-offs between inductive loops and GPS probe vehicles for travel time estimation: A mobile century case study," in *Proc. Transp. Res. Board 91st Annu. Meeting (TRB)*, vol. 349, 2012, pp. 1–20.

[26] X.-Y. Lu, P. Varaiya, R. Horowitz, Z. Guo, and J. Palen, "Estimating traffic speed with single inductive loop event data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2308, no. 1, pp. 157–166, Jan. 2012.

[27] J. Zhang, W. Xiao, B. Coifman, and J. P. Mills, "Vehicle tracking and speed estimation from roadside lidar," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5597–5608, 2020.

[28] Z. Zhang, J. Zheng, H. Xu, and X. Wang, "Vehicle detection and tracking in complex traffic circumstances with roadside LiDAR," *Transp. Res. Rec.*, vol. 2673, no. 9, pp. 62–71, 2019.

[29] M. Mandava, R. S. Gammenthaler, and S. F. Hocker, "Vehicle speed enforcement using absolute speed handheld lidar," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–5.

[30] L. Du, Q. Sun, C. Cai, J. Bai, Z. Fan, and Y. Zhang, "A vehicular mobile standard instrument for field verification of traffic speed meters based on dual-antenna Doppler radar sensor," *Sensors*, vol. 18, no. 4, p. 1099, Apr. 2018.

[31] A. Kumar, P. Khorramshahi, W.-A. Lin, P. Dhar, J.-C. Chen, and R. Chellappa, "A semi-automatic 2D solution for vehicle speed estimation from monocular videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 137–144.

[32] J. Sochor, R. Juránek, and A. Herout, "Traffic surveillance camera calibration by 3D model bounding box alignment for accurate vehicle speed measurement," *Comput. Vis. Image Understand.*, vol. 161, pp. 87–98, Aug. 2017.

[33] N. Balamuralidhar, S. Tilon, and F. Nex, "MultEYE: Monitoring system for real-time vehicle detection, tracking and speed estimation from UAV imagery on edge-computing platforms," *Remote Sens.*, vol. 13, no. 4, p. 573, Feb. 2021.

[34] D. W. Wicaksono and B. Setiyono, "Speed estimation on moving vehicle based on digital image processing," *IJCSAM Int. J. Comput. Sci. Appl. Math.*, vol. 3, no. 1, pp. 21–26, 2017.

[35] A. Nurhadiyatna *et al.*, "Improved vehicle speed estimation using Gaussian mixture model and hole filling algorithm," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, Sep. 2013, pp. 451–456.

[36] R. Krishnapuram, S. Shorewala, and P. Rao, "Link speed estimation for traffic flow modelling based on video feeds from monocular cameras," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–6.

[37] D. C. Luvizon, B. T. Nassu, and R. Minetto, "Vehicle speed estimation by license plate detection and tracking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 6563–6567.

[38] W. Wu, V. Kozitsky, M. E. Hoover, R. Loce, and D. T. Jackson, "Vehicle speed estimation using a monocular camera," *Proc. SPIE*, vol. 9407, Mar. 2015, Art. no. 940704.

[39] W.-C. Ma, S. Wang, R. Hu, Y. Xiong, and R. Urtasun, "Deep rigid instance scene flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3614–3622.

[40] Z. Song, J. Lu, T. Zhang, and H. Li, "End-to-end learning for inter-vehicle distance and relative velocity estimation in ADAS with a monocular camera," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 11081–11087.

[41] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.

[42] X. He and N. C. Yung, "A novel algorithm for estimating vehicle speed from two consecutive images," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Feb. 2007, p. 12.

[43] D. J. Dailey, F. W. Cathey, and S. Pumrin, "An algorithm to estimate mean traffic speed using uncalibrated cameras," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 2, pp. 98–107, Jun. 2000.

[44] M.-T. Tran *et al.*, "Traffic flow analysis with multiple adaptive vehicle detectors and velocity estimation with landmark-based scanlines," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 100–107.

[45] M. Kampelmühler, M. G. Müller, and C. Feichtenhofer, "Camera-based vehicle velocity estimation from monocular video," 2018, *arXiv:1802.07094*.

[46] M. Naphade *et al.*, "The 2018 NVIDIA AI city challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 53–60.

[47] J. Sochor *et al.*, "Brnocompspeed: Review of traffic camera calibration and comprehensive dataset for monocular speed measurement," 2017, *arXiv:1702.06441*.

[48] Z. Tang, G. Wang, H. Xiao, A. Zheng, and J.-N. Hwang, "Single-camera and inter-camera vehicle tracking and 3D speed estimation based on fusion of visual and semantic features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 108–115.

[49] H. Shi, Z. Wang, Y. Zhang, X. Wang, and T. Huang, "Geometry-aware traffic flow analysis by detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 116–120.

[50] T. Mao *et al.*, "AIC2018 report: Traffic surveillance research," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 85–92.

[51] W. Feng, D. Ji, Y. Wang, S. Chang, H. Ren, and W. Gan, "Challenges on large scale surveillance video analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 69–76.

[52] J. Sochor, J. Spanhel, R. Juranek, P. Dobes, and A. Herout, "Graph@FIT submission to the NVIDIA AI city challenge 2018," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 77–84.

[53] T. Huang, "Traffic speed estimation from surveillance video data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 161–165, 2018.

[54] M. Dubska, A. Herout, and J. Sochor, "Automatic camera calibration for traffic understanding," in *Proc. Brit. Mach. Vis. Conf.*, 2014, p. 8.

[55] L. Yu, D. Zhang, X. Chen, and A. Hauptmann, "Traffic danger recognition with surveillance cameras without training data," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.

[56] H. Dong, M. Wen, and Z. Yang, "Vehicle speed estimation based on 3D ConvNets and non-local blocks," *Future Internet*, vol. 11, no. 6, p. 123, May 2019.

[57] V. Kocur, "Perspective transformation for accurate detection of 3D bounding boxes of vehicles in traffic surveillance," in *Proc. 24th Comput. Vis. Winter Workshop*, vol. 2, 2019, pp. 33–41.

[58] V. Kocur and M. Ftáčnik, "Detection of 3D bounding boxes of vehicles using perspective transformation for accurate speed measurement," *Mach. Vis. Appl.*, vol. 31, nos. 7–8, pp. 1–15, Nov. 2020.

[59] M. Zhu, S. Zhang, Y. Zhong, P. Lu, H. Peng, and J. Lenneman, "Monocular 3D vehicle detection using uncalibrated traffic cameras through homography," 2021, *arXiv:2103.15293*.

[60] R. Ke, Z. Li, J. Tang, Z. Pan, and Y. Wang, "Real-time traffic flow parameter estimation from UAV video based on ensemble classifier and optical flow," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 54–64, Jan. 2019.

[61] G. Antonini and J. P. Thiran, "Counting pedestrians in video sequences using trajectory clustering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 8, pp. 1008–1020, Aug. 2016.

[62] H. Song, X. Wang, C. Hua, W. Wang, Q. Guan, and Z. Zhang, "Vehicle trajectory clustering based on 3D information via a coarse-to-fine strategy," *Soft Comput.*, vol. 22, no. 5, pp. 1433–1444, Mar. 2018.

[63] R. Zhao and X. Wang, "Counting vehicles from semantic regions," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 1016–1022, Jun. 2013.

[64] X. Liu, Z. Wang, J. Feng, and H. Xi, "Highway vehicle counting in compressed domain," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3016–3024.

[65] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 833–841.

[66] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2467–2474.

[67] C. Zeng and H. Ma, "Robust head-shoulder detection by PCA-based multilevel HOG-LBP detector for people counting," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2069–2072.

[68] L. Del Pizzo, P. Foggia, A. Greco, G. Percannella, and M. Vento, "Counting people by RGB or depth overhead cameras," *Pattern Recognit. Lett.*, vol. 81, pp. 41–50, Oct. 2016.

[69] F. Y. Shih and X. Zhong, "Automated counting and tracking of vehicles," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 12, Dec. 2017, Art. no. 1750038.

[70] H. Yang and S. Qu, "Real-time vehicle detection and counting in complex traffic scenes using background subtraction model with low-rank decomposition," *IET Intell. Transp. Syst.*, vol. 12, no. 1, pp. 75–85, Nov. 2017.

[71] P. Wei, H. Shi, J. Yang, J. Qian, Y. Ji, and X. Jiang, "City-scale vehicle tracking and traffic flow estimation using low frame-rate traffic cameras," in *Proc. Adjunct Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, Sep. 2019, pp. 602–610.

[72] C. Liu, D. Q. Huynh, Y. Sun, M. Reynolds, and S. Atkinson, "A vision-based pipeline for vehicle counting, speed estimation, and classification," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7547–7560, Dec. 2021.

[73] M. Lei, D. Lefloch, P. Gouton, and K. Madani, "A video-based real-time vehicle counting system using adaptive background method," in *Proc. IEEE Int. Conf. Signal Image Technol. Internet Based Syst.*, Nov. 2008, pp. 523–528.

[74] Z. Moutakki, I. M. Ouloul, K. Afdel, and A. Amghar, "Real-time system based on feature extraction for vehicle detection and classification," *Transp. Telecommun. J.*, vol. 19, no. 2, pp. 93–102, Jun. 2018.

[75] F. Liu, Z. Zeng, and R. Jiang, "A video-based real-time adaptive vehicle-counting system for urban roads," *PLoS ONE*, vol. 12, no. 11, Nov. 2017, Art. no. e0186098.

[76] M. Vasu, N. Abreu, R. Vásquez, and C. López, "Vehicle-counting with automatic region-of-interest and driving-trajectory detection," 2021, *arXiv:2108.07135*.

[77] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2006, pp. 705–711.

[78] A. Ospina and F. Torres, "Countor: Count without bells and whistles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 600–601.

[79] K. Dijkstra, J. van de Loosdrecht, W. A. Atsma, L. R. B. Schomaker, and M. A. Wiering, "CentroidNetV2: A hybrid deep neural network for small-object segmentation and counting," *Neurocomputing*, vol. 423, pp. 490–505, Jan. 2021.

[80] M. A. Abdelwahab, "Accurate vehicle counting approach based on deep neural networks," in *Proc. Int. Conf. Innov. Trends Comput. Eng. (ITCE)*, Feb. 2019, pp. 1–5.

[81] Q. Wu, Y. Zhou, X. Wu, G. Liang, Y. Ou, and T. Sun, "Real-time running detection system for UAV imagery based on optical flow and deep convolutional networks," *IET Intell. Transp. Syst.*, vol. 14, no. 5, pp. 278–287, May 2020.

[82] Z. Tu *et al.*, "A survey of variational and CNN-based optical flow techniques," *Signal Process., Image Commun.*, vol. 72, pp. 9–24, Mar. 2019.

[83] A. Gomaa, M. M. Abdelwahab, M. Abo-Zahhad, T. Minematsu, and R.-I. Taniguchi, "Robust vehicle detection and counting algorithm employing a convolution neural network and optical flow," *Sensors*, vol. 19, no. 20, p. 4588, Oct. 2019.

[84] T.-H. Chen, Y.-F. Lin, and T.-Y. Chen, "Intelligent vehicle counting method based on blob analysis in traffic surveillance," in *Proc. 2nd Int. Conf. Innov. Comput., Informatio Control (ICICIC)*, Sep. 2007, p. 238.

[85] T.-H. Chen, J.-L. Chen, and C.-H. Chen, "Vehicle detection and counting by using headlight information in the dark environment," in *Proc. 3rd Int. Conf. Intell. Inf. Hiding Multimedia Signal Process. (IIH-MSP)*, Nov. 2007, pp. 519–522.

[86] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 589–597.

[87] D. Kang and A. Chan, "Crowd counting by adaptively fusing predictions from an image pyramid," 2018, *arXiv:1805.06115*.

[88] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "CrowdNet: A deep convolutional network for dense crowd counting," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 640–644.

[89] Z. Zhao, H. Li, R. Zhao, and X. Wang, "Crossing-line crowd counting with two-phase deep neural networks," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 712–726.

[90] S. Zhang, G. Wu, J. P. Costeira, and J. M. F. Moura, "FCN-rLSTM: Deep spatio-temporal neural networks for vehicle counting in city cameras," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3667–3676.

[91] Y. Fang, B. Zhan, W. Cai, S. Gao, and B. Hu, "Locality-constrained spatial transformer network for video crowd counting," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2019, pp. 814–819.

[92] F. Xiong, X. Shi, and D.-Y. Yeung, "Spatiotemporal modeling for crowd counting in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5151–5159.

[93] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," 2015, *arXiv:1506.04214*.

[94] J. Lu *et al.*, "Robust and online vehicle counting at crowded intersections," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4002–4008.

[95] S. V.-U. Ha, N. M. Chung, T.-C. Nguyen, and H. N. Phan, "Tiny-PIRATE: A tiny model with parallelized intelligence for real-time analysis as a traffic countEr," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4119–4128.

[96] D. N.-N. Tran, L. H. Pham, H.-H. Nguyen, T. H.-P. Tran, H.-J. Jeon, and J. W. Jeon, "A region-and-trajectory movement matching for multiple turn-counts at road intersection on edge device," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4087–4094.

[97] V.-H. Tran *et al.*, "Real-time and robust system for counting movement-specific vehicle at crowded intersections," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4228–4235.

[98] D. Gloudemans and D. B. Work, "Fast vehicle turning-movement counting using localization-based tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4155–4164.

[99] V. Kocur and M. Ftacnik, "Multi-class multi-movement vehicle counting based on CenterTrack," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4009–4015.

[100] Z. Liu *et al.*, "Robust movement-specific vehicle counting at crowded intersections," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 614–615.

[101] L. Yu, Q. Feng, Y. Qian, W. Liu, and A. G. Hauptmann, "Zero-VIRUS: Zero-shot vehicle route understanding system for intelligent transportation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 594–595.

[102] J. Folenta, J. Spanhel, V. Bartl, and A. Herout, "Determining vehicle turn counts at multiple intersections by separated vehicle classes using CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 596–597.

[103] K.-H.-N. Bui, H. Yi, and J. Cho, "A vehicle counts by class framework using distinguished regions tracking at multiple intersections," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 578–579.

[104] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[105] S. An, W. Liu, and S. Venkatesh, "Face recognition using kernel ridge regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.

[106] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. Brit. Mach. Vis. Conf.*, 2012, p. 3.

[107] S. Zhang, G. Wu, J. P. Costeira, and J. M. F. Moura, "Understanding traffic density from large-scale web camera data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5898–5907.

[108] C. Liu *et al.*, "City-scale multi-camera vehicle tracking guided by crossroad zones," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4129–4137.

[109] J. Ye *et al.*, "A robust MTMC tracking system for AI-city challenge 2021," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4044–4053.

[110] M. Wu, Y. Qian, C. Wang, and M. Yang, "A multi-camera vehicle tracking system based on city-scale vehicle re-ID and spatial-temporal information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4077–4086.

[111] X. Ke, L. Shi, W. Guo, and D. Chen, "Multi-dimensional traffic congestion detection based on fusion of visual features and convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2157–2170, Jun. 2019.

[112] C. Yeshwanth, P. S. A. Sooraj, V. Sudhakaran, and V. Raveendran, "Estimation of intersection traffic density on decentralized architectures with deep networks," in *Proc. Int. Smart Cities Conf. (ISC)*, Sep. 2017, pp. 1–6.

[113] J. Nubert, N. G. Truong, A. Lim, H. I. Tanujaya, L. Lim, and M. A. Vu, "Traffic density estimation using a convolutional neural network," 2018, *arXiv:1809.01564*.

[114] C.-T. Lam, H. Gao, and B. Ng, "A real-time traffic congestion detection system using on-line images," in *Proc. IEEE 17th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2017, pp. 1548–1552.

[115] C.-T. Lam, B. Ng, and C.-W. Chan, "Real-time traffic status detection from on-line images using generic object detection system with deep learning," in *Proc. IEEE 19th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2019, pp. 1506–1510.

[116] B. Qi, W. Zhao, H. Zhang, Z. Jin, X. Wang, and T. Runge, "Automated traffic volume analytics at road intersections using computer vision techniques," in *Proc. 5th Int. Conf. Transp. Inf. Saf. (ICTIS)*, Jul. 2019, pp. 161–169.

[117] L. Jiang, Y. Wang, and Y. Zhao, "Real-time traffic congestion detection with SIGHTA regression network," in *Proc. IEEE 9th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2019, pp. 45–50.

[118] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[119] A. Sobral, L. Oliveira, L. Schnitman, and F. D. Souza, "Highway traffic congestion classification using holistic properties," in *Proc. Comput. Graph. Imag. Signal Process., Pattern Recognit. Appl.*, 2013, pp. 1–8.

[120] K. Garg, S.-K. Lam, T. Srikanthan, and V. Agarwal, "Real-time road traffic density estimation using block variance," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[121] F. Porikli and X. Li, "Traffic congestion estimation using HMM models without vehicle tracking," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2004, pp. 188–193.

[122] Z. Luo, P.-M. Jodoin, S.-Z. Li, and S.-Z. Su, "Traffic analysis without motion features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3290–3294.

[123] Z. Luo, P.-M. Jodoin, S.-Z. Su, S.-Z. Li, and H. Larochelle, "Traffic analytics with low-frame-rate videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 4, pp. 878–891, Apr. 2018.

[124] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, and D. Onoro-Rubio, "Extremely overlapping vehicle counting," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.*, Cham, Switzerland: Springer, 2015, pp. 423–431.

[125] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23, 2010, pp. 1324–1332.

[126] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 615–629.

[127] S. Surya, "TraCount: A deep convolutional neural network for highly overlapping vehicle counting," in *Proc. 10th Indian Conf. Comput. Vis., Graph. Image Process.*, 2016, pp. 1–6.

[128] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt, "Where are the blobs: Counting by localization with point supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 547–562.

[129] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.

[130] Y. Liu, M. Shi, Q. Zhao, and X. Wang, "Point in, box out: Beyond counting persons in crowds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6469–6478.

[131] J. Wan, Q. Wang, and A. B. Chan, "Kernel-based density map generation for dense object counting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1357–1370, Mar. 2022.

[132] L. Ciampi, C. Santiago, J. P. Costeira, C. Gennaro, and G. Amato, "Domain adaptation for traffic density estimation," in *Proc. VISIGRAPP (VISAPP)*, 2021, pp. 185–195.

[133] B. Maaloul, A. Taleb-Ahmed, S. Niar, N. Harb, and C. Valderrama, "Adaptive video-based algorithm for accident detection on highways," in *Proc. 12th IEEE Int. Symp. Ind. Embedded Syst. (SIES)*, Jun. 2017, pp. 1–6.

[134] Z. Chen *et al.*, "Dangerous driving behavior detection using video-extracted vehicle trajectory histograms," *J. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 409–421, Sep. 2017.

[135] X. Huang, P. He, A. Rangarajan, and S. Ranka, "Intelligent intersection: Two-stream convolutional networks for real-time near-accident detection in traffic video," *ACM Trans. Spatial Algorithms Syst. (TSAS)*, vol. 6, no. 2, pp. 1–28, 2020.

[136] P. Chakraborty, A. Sharma, and C. Hegde, "Freeway traffic incident detection from cameras: A semi-supervised learning approach," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 1840–1845.

[137] H. Kim, S. Park, and J. Paik, "Pre-activated 3D CNN and feature pyramid network for traffic accident detection," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2020, pp. 1–3.

[138] H.-L. Ooi, G.-A. Bilodeau, and N. Saunier, "Tracking in urban traffic scenes from background subtraction and object detection," in *Proc. Int. Conf. Image Anal. Recognit.*, Cham, Switzerland: Springer, 2019, pp. 195–206.

[139] Y. Yang and G.-A. Bilodeau, "Multiple object tracking with kernelized correlation filters in urban mixed traffic," in *Proc. 14th Conf. Comput. Robot Vis. (CRV)*, May 2017, pp. 209–216.

[140] D.-A. Beaupré, G.-A. Bilodeau, and N. Saunier, "Improving multiple object tracking with optical flow and edge preprocessing," 2018, *arXiv:1801.09646*.

[141] A. Fuentes, S. Yoon, and D. S. Park, "Spatial multilevel optical flow architecture-based dynamic motion estimation in vehicular traffic scenarios," *KSII Trans. Internet Inf. Syst. (TIIS)*, vol. 12, no. 12, pp. 5978–5999, 2018.

[142] J. Athanesious, V. Srinivasan, V. Vijayakumar, S. Christobel, and S. C. Sethuraman, "Detecting abnormal events in traffic video surveillance using superorientation optical flow feature," *IET Image Process.*, vol. 14, no. 9, pp. 1881–1891, Jul. 2020.

[143] J. Zhou and C. Kwan, "Anomaly detection in low quality traffic monitoring videos using optical flow," *Proc. SPIE*, vol. 10649, Apr. 2018, Art. no. 106490F.

[144] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," 2015, *arXiv:1510.01553*.

[145] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[146] D.-Y. Chen and P.-C. Huang, "Motion-based unusual event detection in human crowds," *J. Vis. Commun. Image Represent.*, vol. 22, no. 2, pp. 178–186, 2011.

[147] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 56–62.

[148] S. Wu, H.-S. Wong, and Z. Yu, "A Bayesian model for crowd escape behavior detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 85–98, Jan. 2014.

[149] K. Yun, H. Jeong, K. M. Yi, S. W. Kim, and J. Y. Choi, "Motion interaction field for accident detection in traffic surveillance video," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 3062–3067.

[150] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets horn/schunck: Combining local and global optic flow methods," *Int. J. Comput. Vis.*, vol. 61, no. 3, pp. 211–231, 2005.

[151] P. Ahmadi, E. P. Moradian, and I. Gholampour, "Sequential topic modeling for efficient analysis of traffic scenes," in *Proc. 9th Int. Symp. Telecommun. (IST)*, Dec. 2018, pp. 559–564.

[152] P. Giannakeris, V. Kaltsa, K. Avgerinakis, A. Briassouli, S. Vrochidis, and I. Kompatsiaris, "Speed estimation and abnormality detection from surveillance cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 93–99.

[153] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[154] H. Ullah, M. Ullah, H. Afridi, N. Conci, and F. G. B. D. Natale, "Traffic accident detection through a hydrodynamic lens," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2470–2474.

[155] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scandin. Conf. Image Anal.*, Cham, Switzerland: Springer, 2003, pp. 363–370.

[156] M. Bortnikov, A. Khan, A. M. Khattak, and M. Ahmad, "Accident recognition via 3D CNNs for automated traffic monitoring in smart cities," in *Proc. Sci. Inf. Conf.*, Cham, Switzerland: Springer, 2019, pp. 256–264.

[157] E. Batanina, I. E. I. Bekkouch, Y. Youssry, A. Khan, A. M. Khattak, and M. Bortnikov, "Domain adaptation for car accident detection in videos," in *Proc. 9th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2019, pp. 1–6.

[158] Y. Xu *et al.*, "Dual-mode vehicle motion pattern learning for high performance road traffic anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 145–152.

[159] L. Shine, A. Edison, and C. Jiji, "A comparative study of faster R-CNN models for anomaly detection in 2019 ai city challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 306–314.

[160] J. Wei, J. Zhao, Y. Zhao, and Z. Zhao, "Unsupervised anomaly detection for traffic surveillance based on background modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 129–136.

[161] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. 17th Int. Conf. Pattern Recognit.*, vol. 2, Aug. 2004, pp. 28–31.

[162] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[163] J. Zhao *et al.*, "Unsupervised traffic anomaly detection using trajectories," in *Proc. CVPR Workshops*, Jan. 2019, pp. 133–140.

[164] K.-T. Nguyen *et al.*, "Vehicle re-identification with learned representation and spatial verification and abnormality detection with multi-adaptive vehicle detectors for traffic video analysis," in *Proc. CVPR Workshops*, Jan. 2019, pp. 363–372.

[165] S. Bai *et al.*, "Traffic anomaly detection via perspective map based on spatial-temporal information matrix," in *Proc. CVPR Workshops*, Jan. 2019, pp. 117–124.

[166] K. M. Biradar, A. Gupta, M. Mandal, and S. K. Vipparthi, "Challenges in time-stamp aware anomaly detection in traffic videos," 2019, *arXiv:1906.04574*.

[167] M. Biparva, D. Fernández-Llorca, R. Izquierdo-Gonzalo, and J. K. Tsotsos, "Video action recognition for lane-change classification and prediction of surrounding vehicles," 2021, *arXiv:2101.05043*.

[168] G. Madhumitha, R. Senthilnathan, K. M. Ayaz, J. Vignesh, and K. Madhu, "Estimation of collision priority on traffic videos using deep learning," in *Proc. IEEE Int. Conf. Mach. Learn. Appl. Netw. Technol. (ICMLANT)*, Dec. 2020, pp. 1–6.

[169] A. K. Agrawal *et al.*, "Automatic traffic accident detection system using ResNet and SVM," in *Proc. 5th Int. Conf. Res. Comput. Intell. Commun. Netw. (ICRCICN)*, Nov. 2020, pp. 71–76.

[170] I. Baek and M. He, "Vehicles lane-changing behavior detection," 2018, *arXiv:1808.07518*.

[171] J. Li, C. Lu, Y. Xu, Z. Zhang, J. Gong, and H. Di, "Manifold learning for lane-changing behavior recognition in urban traffic," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 3663–3668.

[172] S. Xia, J. Xiong, Y. Liu, and G. Li, "Vision-based traffic accident detection using matrix approximation," in *Proc. 10th Asian Control Conf. (ASCC)*, May 2015, pp. 1–5.

[173] S. S. Thomas, S. Gupta, and V. K. Subramanian, "Event detection on roads using perceptual video summarization," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 9, pp. 2944–2954, Dec. 2018.

[174] Z. Wang, X. Wang, L. Zhao, and G. Zhang, "Vision-based lane departure detection using a stacked sparse autoencoder," *Math. Problems Eng.*, vol. 2018, pp. 1–15, Sep. 2018.

[175] O. Isupova, D. Kuzin, and L. Mihaylova, "Learning methods for dynamic topic modeling in automated behavior analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 3980–3993, Sep. 2018.

[176] J. Spanhel, J. Sochor, and A. Makarov, "Detection of traffic violations of road users based on convolutional neural networks," in *Proc. 14th Symp. Neural Netw. Appl. (NEUREL)*, Nov. 2018, pp. 1–6.

[177] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 21–37.

[178] Z. Wei, C. Wang, P. Hao, and M. J. Barth, "Vision-based lane-changing behavior detection using deep residual neural network," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 3108–3113.

[179] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[180] M.-C. Chang *et al.*, "AI city challenge 2019-city-scale video analytics for smart transportation," in *Proc. CVPR Workshops*, Jan. 2019, pp. 99–108.

[181] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[182] J. Leng, Y. Liu, D. Du, T. Zhang, and P. Quan, "Robust obstacle detection and recognition for driver assistance systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1560–1571, Apr. 2020.

[183] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[184] K. Gupta, S. A. Javed, V. Gandhi, and K. M. Krishna, "MergeNet: A deep net architecture for small obstacle discovery," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 5856–5862.

[185] O. Scheel, N. S. Nagaraja, L. Schwarz, N. Navab, and F. Tombari, "Attention-based lane change prediction," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8655–8661.

[186] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2174–2182.

[187] E. Yurtsever *et al.*, "Risky action recognition in lane change video clips using deep spatiotemporal networks with segmentation mask transfer," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 3100–3107.

[188] M. Xu, M. Gao, Y.-T. Chen, L. Davis, and D. Crandall, "Temporal recurrent networks for online action detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5532–5541.

[189] G.-P. Corcoran and J. Clark, "Traffic risk assessment: A two-stream approach using dynamic-attention," in *Proc. 16th Conf. Comput. Robot Vis. (CRV)*, May 2019, pp. 166–173.

[190] X. Ren, D. Wang, M. Laskey, and K. Goldberg, "Learning traffic behaviors by extracting vehicle trajectories from online video streams," in *Proc. IEEE 14th Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2018, pp. 1276–1283.

[191] M. Umer, N. Ahmed, and M. Shabbar, "Unsupervised video surveillance for anomaly detection of street traffic," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 12, pp. 270–275, 2017.

[192] C.-E. Wu, W.-Y. Yang, H.-C. Ting, and J.-S. Wang, "Traffic pattern modeling, trajectory classification and vehicle tracking within urban intersections," in *Proc. Int. Smart Cities Conf. (ISC)*, Sep. 2017, pp. 1–6.

[193] K. K. Santhosh, D. P. Dogra, P. P. Roy, and B. B. Chaudhuri, "Trajectory-based scene understanding using Dirichlet process mixture model," *IEEE Trans. Cybern.*, vol. 51, no. 8, pp. 4148–4161, Aug. 2021.

[194] S. K. Kumaran, D. P. Dogra, P. P. Roy, and A. Mitra, "Video trajectory classification and anomaly detection using hybrid CNN-VAE," 2018, *arXiv:1812.07203*.

[195] C. Li, Z. Han, Q. Ye, and J. Jiao, "Visual abnormal behavior detection based on trajectory sparse reconstruction analysis," *Neurocomputing*, vol. 119, pp. 94–100, Nov. 2013.

[196] Ö. Aköz and M. E. Karsligil, "Traffic event classification at intersections based on the severity of abnormality," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 613–632, Dec. 2014.

[197] J. Wang, L. Xia, X. Hu, and Y. Xiao, "Abnormal event detection with semi-supervised sparse topic model," *Neural Comput. Appl.*, vol. 31, no. 5, pp. 1607–1617, May 2019.

[198] D. Singh and C. K. Mohan, "Deep spatio-temporal representation for detection of road accidents using stacked autoencoder," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 879–887, Mar. 2018.

[199] J. Zhu *et al.*, "Bidirectional long short-term memory network for vehicle behavior recognition," *Remote Sens.*, vol. 10, no. 6, p. 887, Jun. 2018.

[200] S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Trajectory-based surveillance analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 1985–1997, Jul. 2019.

[201] G. Wang, X. Yuan, A. Zheng, H.-M. Hsu, and J.-N. Hwang, "Anomaly candidate identification and starting time estimation of vehicles from traffic videos," in *Proc. CVPR Workshops*, Jan. 2019, pp. 382–390.

[202] 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*, Oct. 2019, pp. 482–490.

[203] J. Zhu, W. Lin, K. Sun, X. Hou, B. Liu, and G. Qiu, "Behavior recognition of moving objects using deep neural networks," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Oct. 2018, pp. 45–52.

[204] A. Makhmutova, R. Minnikhanov, M. Dagaeva, I. Anikin, T. Bolshakov, and I. Khuziakhmetov, "Intelligent detection of Object's anomalies for road surveilance cameras," in *Proc. Int. Multi-Conf. Eng., Comput. Inf. Sci. (SIBIRCON)*, Oct. 2019, pp. 762–767.

[205] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[206] Y. Zhao, W. Wu, Y. He, Y. Li, X. Tan, and S. Chen, "Good practices and a strong baseline for traffic anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 3993–4001.

[207] J. Wu, X. Wang, X. Xiao, and Y. Wang, "Box-level tube tracking and refinement for vehicles anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4112–4118.

[208] J. Chen *et al.*, "Dual-modality vehicle anomaly detection via bilateral trajectory tracing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4016–4025.

[209] K. Doshi and Y. Yilmaz, "An efficient approach for anomaly detection in traffic videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4236–4244.

[210] Y. Li *et al.*, "Multi-granularity tracking with modularlized components for unsupervised vehicles anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 586–587.

[211] K. Doshi and Y. Yilmaz, "Fast unsupervised anomaly detection in traffic videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 624–625.

[212] L. Shine, M. A. Vaishnav, and C. V. Jiji, "Fractional data distillation model for anomaly detection in traffic videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 606–607.

[213] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[214] X. Long, C. Gan, G. de Melo, J. Wu, X. Liu, and S. Wen, "Attention clusters: Purely attention based local feature integration for video classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7834–7843.

[215] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," 2019, *arXiv:1905.00546*.

[216] S. Ghosh, S. J. Sunny, and R. Roney, "Accident detection using convolutional neural networks," in *Proc. Int. Conf. Data Sci. Commun. (IconDSC)*, Mar. 2019, pp. 1–6.

[217] S. M. Shareef *et al.*, "Optimized frame detection technique in vehicle accident using deep learning," *Zanco J. Pure Appl. Sci.*, vol. 32, no. 4, pp. 38–47, 2020.

[218] Y. Luo, P. Cai, Y. Lee, and D. Hsu, "GAMMA: A general agent motion model for autonomous driving," 2019, *arXiv:1906.01566*.

[219] L. Yu, P. Chen, W. Liu, G. Kang, and A. G. Hauptmann, "Training-free monocular 3D event detection system for traffic surveillance," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 3838–3843.

[220] E. P. Ijjina, D. Chand, S. Gupta, and K. Goutham, "Computer vision-based accident detection in traffic surveillance," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 1–6.

[221] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7699–7707.

[222] D. Roy, T. Ishizaka, C. K. Mohan, and A. Fukuda, "Detection of collision-prone vehicle behavior at intersections using Siamese interaction LSTM," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3137–3147, Apr. 2022.

[223] R. Herzig *et al.*, "Spatio-temporal action graph networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2347–2356.

[224] C. Li, Y. Meng, S. H. Chan, and Y.-T. Chen, "Learning 3D-aware egocentric spatial-temporal interaction via graph convolutional networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 8418–8424.

[225] A. Narayanan, Y.-T. Chen, and S. Malla, "Semi-supervised learning: Fusion of self-supervised, supervised learning, and multimodal cues for tactical driver behavior detection," 2018, *arXiv:1807.00864*.

[226] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[227] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[228] A. Rasouli, "Deep learning for vision-based prediction: A survey," 2020, *arXiv:2007.00095*.

[229] Y. Cai *et al.*, "Pedestrian motion trajectory prediction in intelligent driving from far shot first-person perspective video," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5298–5313, Jun. 2022.

[230] D. Roy, T. Ishizaka, C. K. Mohan, and A. Fukuda, "Vehicle trajectory prediction at intersections using interaction based generative adversarial networks," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 2318–2323.

[231] N. Deo, A. Rangesh, and M. M. Trivedi, "How would surround vehicles move? A unified framework for maneuver classification and motion prediction," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 2, pp. 129–140, Jun. 2018.

[232] H. M. Mandalia and M. D. D. Salvucci, "Using support vector machines for lane-change detection," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, Los Angeles, CA, USA: SAGE Publications, 2005, pp. 1965–1969.

[233] G. S. Aoude, B. D. Luders, K. K. H. Lee, D. S. Levine, and J. P. How, "Threat assessment design for driver assistance system at intersections," in *Proc. 13th Int. IEEE Conf. Intell. Transp. Syst.*, Sep. 2010, pp. 1855–1862.

[234] A. Khosroshahi, E. Ohn-Bar, and M. M. Trivedi, "Surround vehicles trajectory analysis with recurrent neural networks," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 2267–2272.

[235] M. Schreier, V. Willert, and J. Adamy, "Bayesian, maneuver-based, long-term trajectory prediction and criticality assessment for driver assistance systems," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 334–341.

[236] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.

[237] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4601–4607.

[238] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "TraPHic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8483–8492.

[239] S. Haddad, M. Wu, H. Wei, and S. K. Lam, "Situation-aware pedestrian trajectory prediction with spatio-temporal attention model," 2019, *arXiv:1902.05437*.

[240] R. Chandra *et al.*, "Forecasting trajectory and behavior of road-agents using spectral clustering in graph-LSTMs," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4882–4890, Jul. 2020.

[241] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha, "Trafficpredict: Trajectory prediction for heterogeneous traffic-agents," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6120–6127.

[242] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.

[243] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, "DESIRE: Distant future prediction in dynamic scenes with interacting agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 336–345.

[244] R. Chandra, U. Bhattacharya, C. Roncal, A. Bera, and D. Manocha, "RobustTP: End-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs," in *Proc. ACM Comput. Sci. Cars Symp.*, Oct. 2019, pp. 1–9.

[245] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," 2019, *arXiv:1903.00618*.

[246] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush, "Egocentric vision-based future vehicle localization for intelligent driving assistance systems," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 9711–9717.

[247] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Proc. Asian Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 136–153.

[248] A. P. Shah, J.-B. Lamare, T. Nguyen-Anh, and A. Hauptmann, "CADP: A novel dataset for CCTV traffic camera based accident analysis," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–9.

[249] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident DB," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3521–3529.

[250] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. C. Niebles, and M. Sun, "Agent-centric risk assessment: Accident anticipation and risky region localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2222–2230.

[251] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2682–2690.

[252] M. M. Karim, Y. Li, R. Qin, and Z. Yin, "A dynamic spatial-temporal attention network for early anticipation of traffic accidents," 2021, *arXiv:2106.10197*.

[253] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 261–268.

[254] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Comput. Graph. Forum*, vol. 26, no. 3, pp. 655–664, 2007.

[255] T. Zhao *et al.*, "Multi-agent tensor fusion for contextual trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12126–12134.

[256] H. Bi, R. Zhang, T. Mao, Z. Deng, and Z. Wang, "How can I see my future? FvTraj: Using first-person view for pedestrian trajectory prediction," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2020, pp. 576–593.

[257] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 507–523.

[258] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K.: Springer, Aug. 2020, 2020, pp. 683–700.

[259] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise goal-driven networks for trajectory prediction," 2021, *arXiv:2103.14107*.

[260] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7074–7082.

[261] B. Xu and Z. Chen, "Multi-level fusion based 3D object detection from monocular images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2345–2353.

[262] T. He and S. Soatto, "Mono3D++: Monocular 3D vehicle detection with two-scale 3d hypotheses and task priors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8409–8416.

[263] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8445–8453.
[264] P. Li, X. Chen, and S. Shen, "Stereo R-CNN based 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7644–7652.
[265] A. D. Pon, J. Ku, C. Li, and S. L. Waslander, "Object-centric stereo matching for 3D object detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 8383–8389.
[266] Y. Chen, S. Liu, X. Shen, and J. Jia, "DSGN: Deep stereo geometry network for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12536–12545.
[267] J. Chen, W. Xu, H. Xu, F. Lin, Y. Sun, and X. Shi, "Fast vehicle detection using a disparity projection method," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 9, pp. 2801–2813, Sep. 2018.
[268] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3D object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 641–656.
[269] W. Van Gansbeke, D. Neven, B. D. Brabandere, and L. Van Gool, "Sparse and noisy LiDAR completion with RGB guidance and uncertainty," in *Proc. 16th Int. Conf. Mach. Vis. Appl. (MVA)*, May 2019, pp. 1–6.
[270] Y. Long, D. Morris, X. Liu, M. Castro, P. Chakravarty, and P. Narayanan, "Radar-camera pixel depth association for depth completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12507–12516.
[271] M. Akhtar and S. Moridpour, "A review of traffic congestion prediction using artificial intelligence," *J. Adv. Transp.*, vol. 2021, pp. 1–18, Jan. 2021.
[272] P. Chakraborty, Y. O. Adu-Gyamfi, S. Poddar, V. Ahsani, A. Sharma, and S. Sarkar, "Traffic congestion detection from camera images using deep convolution neural networks," *Transp. Res. Rec.*, vol. 2672, no. 45, pp. 222–231, Dec. 2018.
[273] S. Guo, Y. Lin, S. Li, Z. Chen, and H. Wan, "Deep spatial–temporal 3D convolutional neural networks for traffic data forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3913–3926, Oct. 2019.
[274] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 922–929.
[275] C. Snyder and M. N. Do, "Streets: A novel camera network dataset for traffic flow," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
[276] M. Rahman, "Application of parameter estimation and calibration method for car-following models," Ph.D. dissertation, Graduate School Clemson Univ., Clemson, SC, USA, 2013.
[277] N. V. Hung, L. C. Tran, N. H. Dung, T. M. Hoang, and N. T. Dzung, "A traffic monitoring system for a mixed traffic flow via road estimation and analysis," in *Proc. IEEE 6th Int. Conf. Commun. Electron. (ICCE)*, Jul. 2016, pp. 375–378.
[278] J. Tanimoto, S. Kukida, and A. Hagishima, "Social dilemma structures hidden behind traffic flow with lane changes," *J. Stat. Mech., Theory Exp.*, vol. 2014, no. 7, Jul. 2014, Art. no. P07019.
[279] S. Liu, Y. Tang, Y. Tian, and H. Su, "Visual driving assistance system based on few-shot learning," *Multimedia Syst.*, early access, pp. 1–11, Jul. 2021.
[280] B. Li, "3D fully convolutional network for vehicle detection in point cloud," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 1513–1518.
[281] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, "Vote3Deep: Fast object detection in 3D point clouds using efficient convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1355–1361.
[282] G. Menghani, "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," 2021, *arXiv:2106.08962*.

Jing Chen received the Ph.D. degree in mechanical engineering from Zhejiang University, China, in 2010. She is currently an Associate Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, China. Her research interests are in computer vision, machine learning, and urban transportation.

Qichao Wang received the B.S. degree from the Qingdao University of Technology, Qingdao, China, in 2015. He is currently pursuing the M.D. degree in information technology from the School of Computer Science, Hangzhou Dianzi University, Zhejiang, China. His research interests include intelligent transportation systems, pattern recognition, and computer vision.

Harry H. Cheng (Senior Member, IEEE) received the M.S. degree in mathematics and the Ph.D. degree in mechanical engineering from the University of Illinois Chicago, Chicago, in 1986 and 1989, respectively. From 1989 to 1992, he was a Senior Engineer for robotic automation systems with the Research and Development Division, United Parcel Service. He is currently a Professor with the Department of Mechanical and Aerospace Engineering and the Graduate Group in Computer Science, University of California at Davis, where he is also the Director of the Integration Engineering Laboratory. His current research interests include computer-aided engineering, mobile agent-based computing, intelligent mechatronic and embedded systems, and robotics. He is a fellow of the American Society of Mechanical Engineers (ASME). He was the Conference Chair and the Program Chair of the IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications.

Weiming Peng received the Ph.D. degree in computer application technology from the South China University of Technology, Guangzhou, China, in 2013. He is a Lecturer with Hangzhou Dianzi University, Hangzhou, China. His current research interests include quantum computation and data fusion.

Wenqiang Xu received the Ph.D. degree in economics management from Wuhan University, China, in 2015. He is currently a Lecturer with the College of Economics and Management, China Jiliang University. His research interests are in computer vision, machine learning, and urban transportation.