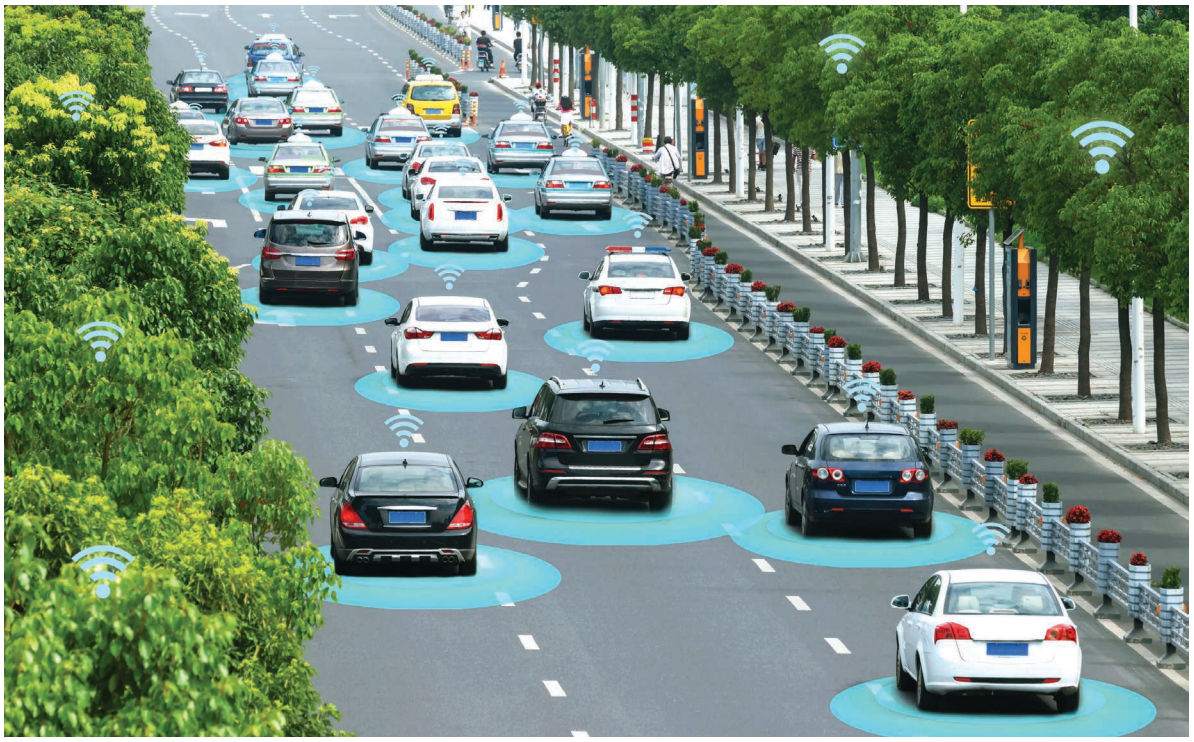

©SHUTTERSTOCK.COM/ZAPP2PHOTO

On-Road Object Detection and Tracking Based on Radar and Vision Fusion: A Review

Xiaolin Tang and Zhiqiang Zhang
Are with the College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, 400044, China. Email: tangxl0923@cqu.edu.cn; 20193202026t@cqu.edu.cn.

Yechen Qin*
Is with the School of Mechanical Engineering, Beijing Institute of Technology, Beijing, 100081, China, and the Institute of Advanced Technology, Beijing Institute of Technology, Jinan, 250101, China. Email: qinyechenbit@gmail.com.

Digital Object Identifier 10.1109/MITS.2021.3093379
Date of current version: 4 August 2021

*Corresponding author

***Abstract*—Environment perception, one of the most fundamental and challenging problems of autonomous vehicles (AVs), has been widely studied in recent decades. Due to inferior fault tolerance and the insufficient information caused by a single autonomous sensor (e.g., radar, lidar, or camera), multisensor fusion plays a significant role in environment perception systems, and its performance directly defines the safety of AVs. Due to good performance and low cost, radar-vision (RV) fusion has become popular and widely applied in the mass production of AVs. However, there have been a few generalizations about RV fusion, and in that context, this article presents a comprehensive review on RV fusion for both object detection and object tracking by RV fusion. With respect to the input data and fusion framework, this article categorizes the existing fusion frameworks into two categories, providing a detailed overview of each: object detection and tracking by RV fusion. Also, the state-of-the-art detectors and trackers based on deep learning are introduced, along with an analysis of their advantages and limitations. Finally, challenges and improvements are summarized to facilitate future research in the RV fusion field.**

According to the agenda of the World Health Organization, every year, all around the world, approximately 1.35 million people die in vehicle crash accidents while 50 million people are injured [1]. Road traffic accidents continue to impose unacceptable costs on humanity in terms of enormous human suffering and significant economic losses. To reduce the number of traffic accidents, Stockholm Declaration released a vision for the future: a 50% reduction in casualties over the next decade on the way to Vision Zero by 2050 [2].

Obviously, the hardware reliability of a vehicle is the basis of safe driving, thus it has been widely studied in recent years [3]. However, most road accidents are still attributed to driver inexperience, declining abilities, and inattention, rather than the hardware reliability of a vehicle [4]. In addition, fuel consumption caused by traffic congestion has a negative effect on environmental pollution [5]. Consequently, developing autonomous vehicles (AVs) to reduce collision, congestion, and pollution has become a research hot spot, both in academia and the automobile industry. Robust and reliable environment perception that includes multiple sensing modalities is the first step to achieve safe-driving AVs. Sensors can simulate the sensory systems of humans and other animals in the aspects of distance, speed, and vision [6], so they can be used to improve driving safety and reduce traffic accidents caused by drivers. Further, the fusion of multisensor information can provide more precise and robust results than any of the sensors used individually [7]. Three sensor types (e.g., radar, camera, and lidar) have been extensively used in AVs. These sensors are commonly combined, and the most widely used sensor combinations include radar vision (RV), vision lidar (VL), and RV lidar (RVL) [8]. The performances of different sensors and their combinations are presented in Figure 1.

Due to the current trend of advanced driver assistance systems (ADASs), RV has been the mainstream in both academia and industry [9]. Although both RV and RVL can address the challenges introduced by a complex road environment, considering the cost, RV fusion has priority over RVL. Therefore, a comprehensive analysis of RV fusion is necessary. On-road RV environment perception is presented in Figure 2. The advantages of RV fusion is summarized as follows.

- Vision can realize object detection and a semantic understanding that lidar cannot perform, such as traffic light recognition, traffic sign recognition, and pedestrian gesture recognition [10]. However, radar can provide information on vehicle position and velocity, which vision lacks. Consequently, RV fusion can achieve the same or even better performance than the other fusion approaches.
- Lidar has the disadvantages of high cost and low reliability [11]. Even if lidar's cost decreases dramatically in the future, its application to AVs will still be limited by ambient light, adverse weather, similar-frequency interference, and the poor durability of mechanical lidar [12]. Moreover, lidar has to be installed at a certain height above a vehicle, affecting its air drag and aesthetics.
- RV fusion has been widely used in ADASs, and it has already been applied to mass production in the automotive industry, such as Tesla [13]. Also, many vision companies, including Mobileye [14], Minieye [15], and MM Solutions [16], have already matured the perceptual technology and road-testing experience. Apollo Lite used a vision-based vehicle framework to achieve level autonomy [17].

Related Work

To the best of the authors' knowledge, there has not been a comprehensive review on RV fusion for vehicle detection and tracking. However, there have been many notable related surveys on both object detection and tracking, which can be grouped into three following categories:

1) *Vision-based object detection*: Many articles have studied the detection approaches of a single-class object, such as vehicle [18], [19], motorcycle [20], and pedestrian

Performance	Radar	Vision	Lidar	RV	VL	RVL
Sensing Distance	✓	Fair	✓	✓	✓	✓
Object Detection	✓	✓	✓	✓	✓	✓
Lane Detection	×	✓	✓	✓	✓	✓
Object Classification	Fair	✓	×	✓	✓	✓
Location Estimation	✓	×	✓	✓	✓	✓
Velocity Estimation	✓	×	✓	✓	✓	✓
Traffic Lights/Signs Recognition	×	✓	×	✓	✓	✓
Gestures (Human) Recognition	×	✓	×	✓	✓	✓
Adverse Weather Performance	✓	Fair	Fair	✓	Fair	✓
Low-Light Performance	✓	Fair	✓	✓	✓	✓

FIG 1 A performance comparison of the sensors and their combinations. Note that "✓" represents excellent performance and "×" signifies limited performance.

detection [21]. In [22] and [23], various object detection methods were discussed in which most objects were unlikely to be on the road. The listed articles are focused mainly on detection techniques and algorithms.

2) *Vision-based object tracking*: The visual object tracking approaches based on object appearance [22] and deep learning (DL) [24] have been commonly reviewed. The reviews consider mainly the specific object appearances, algorithms, and network structures.
3) *Other approaches of sensor-based object detection and tracking*: Granström et al. [25] illustrated the extended object tracking methods using several different sensors and object types. Walia and Kapoor [26] reviewed the object tracking technique where data were generated from single and multiple sensors. Object detection based on radar, lidar, and camera was briefly introduced in [27]; however, only the tracking cues were introduced, instead of a detailed description of the fusion approaches and their performances.

FIG 2 On-road RV environment perception. The relative position information and labels of the RV results are represented by yellow text and boxes, while the green boxes show the objects detected by a camera.

Contributions and Organization

On-road object detection and tracking is helpful to perceive environmental comprehensiveness and improve driving safety. Therefore, the aim of this review is to provide an overview of the RV fusion for on-road environment perception. The main contributions of this review can be summarized as follows.

- The object detection and tracking hot spots are analyzed, and the state-of-the-art object detection and tracking methods published in different core journals are investigated.
- The RV fusion perspectives are discussed in detail. With respect to the input data and fusion framework, the RV fusion methods that differ from traditional data fusion can be classified into two types: object detection and object tracking by RV fusion.

- The real-time performances of object detection and tracking on the public data set are summarized.

Object Detection

As the most challenging task in AVs, on-road object detection aims to determine the position, velocity, label, and bounding box (BBox) of various objects on the road to improve driving safety. This section reviews mainly the object detection approaches that use the data obtained by radar and camera. On the one hand, radar-based object detection uses millimeter-wave (mm-wave) to determine the distance, velocity, and angle to a particular object based on the Doppler effect. On the other hand, vision-based object detection uses a camera to capture the ambient light and then employs computer vision (CV)-based procedures to locate and recognize the object. The mapping radar measurement data and image information obtained by perspective projection and an RV fusion strategy can be used for robust object detection. A comparison of the interactions between sensors and objects is presented in Figure 3(a)–(c).

Vision-Based Object Detection

Vision-based object detection approaches can be divided into two major groups: appearance [28] and DL based [23]. Appearance-based object detection methods use a priori knowledge to recognize on-road objects, and they are based on subjective assumptions about vehicles. These strategies are suitable for situations with obvious appearances and a simple background; however, in practice, the background and the object are complex and changeable, so it is difficult to detect objects based on the a priori appearance. Moreover, the prior appearances of objects in various complex and changeable situations cannot be summarized. Conversely, DL-based approaches are data driven and use large volumes and rich data to train detection models to automatically extract the abstract features that the computer understands, rather than the appearance of the manual design. As illustrated in Figure 4, the research in this field has evolved from appearance (e.g., shadow and Haar-like) to DL. Specifically, DL can easily migrate to new object classes and improve the recognition accuracy of occluded objects. Moreover, DL is more robust to light changes and can cope with both dark and bright environments. With a strong generalization ability, DL-based object detection strategies can be easily applied to actual dynamic scenes, which is why they have become popular recently.

Appearance-Based Object Detection Approaches

Appearance-based object detection approaches are based on shallow (e.g., symmetry, shadow, and edge) and local features [e.g., histogram of oriented gradients (HOG) and Haar-like] [28]. Shallow features are used mainly for vehicle detection, while local features can be used in the detection of various types of objects, such as pedestrians, vehicles, and two-wheel vehicles. Many appearance-based vehicle detection approaches search for vehicles in the entire image [18], [19]. This section briefly introduces the appearance-based perspectives studied in the RV fusion-related literature, and most of them are applied to the regions of interest

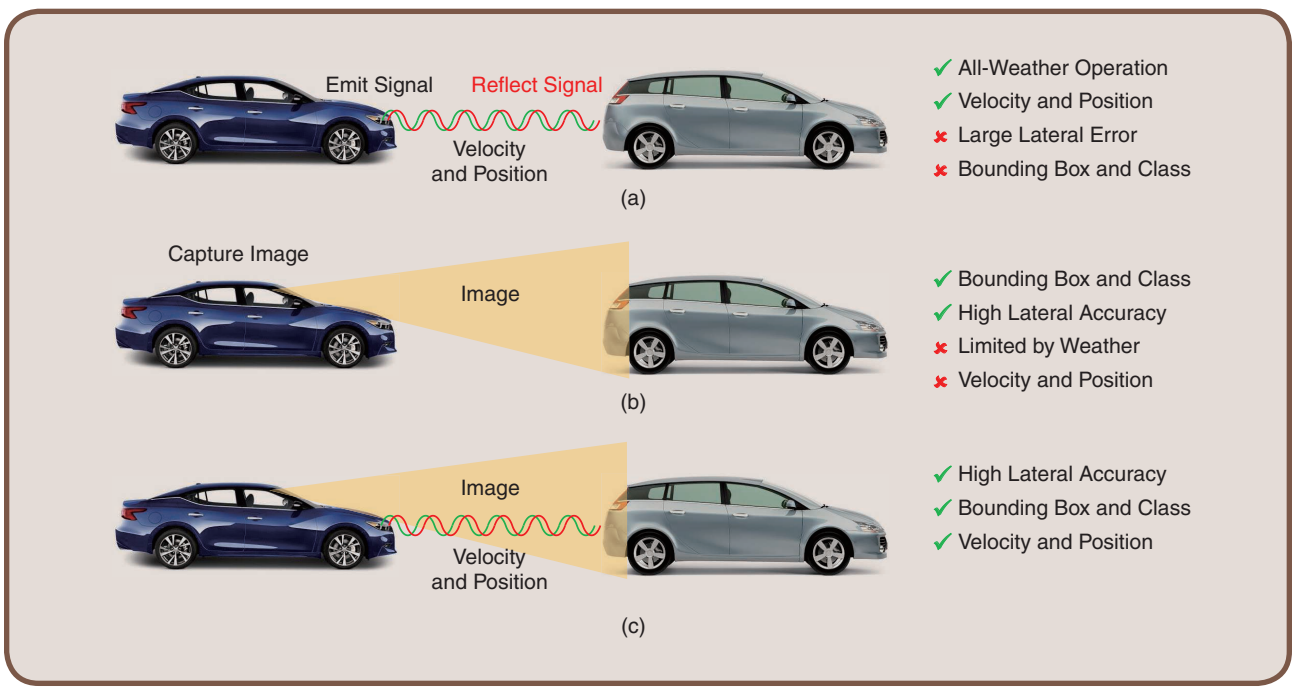

FIG 3 A comparison of the interactions between sensors and objects: (a) radar for object detection, (b) camera for object detection, and (c) RV object detection. Their main pros and cons are listed on the right side.

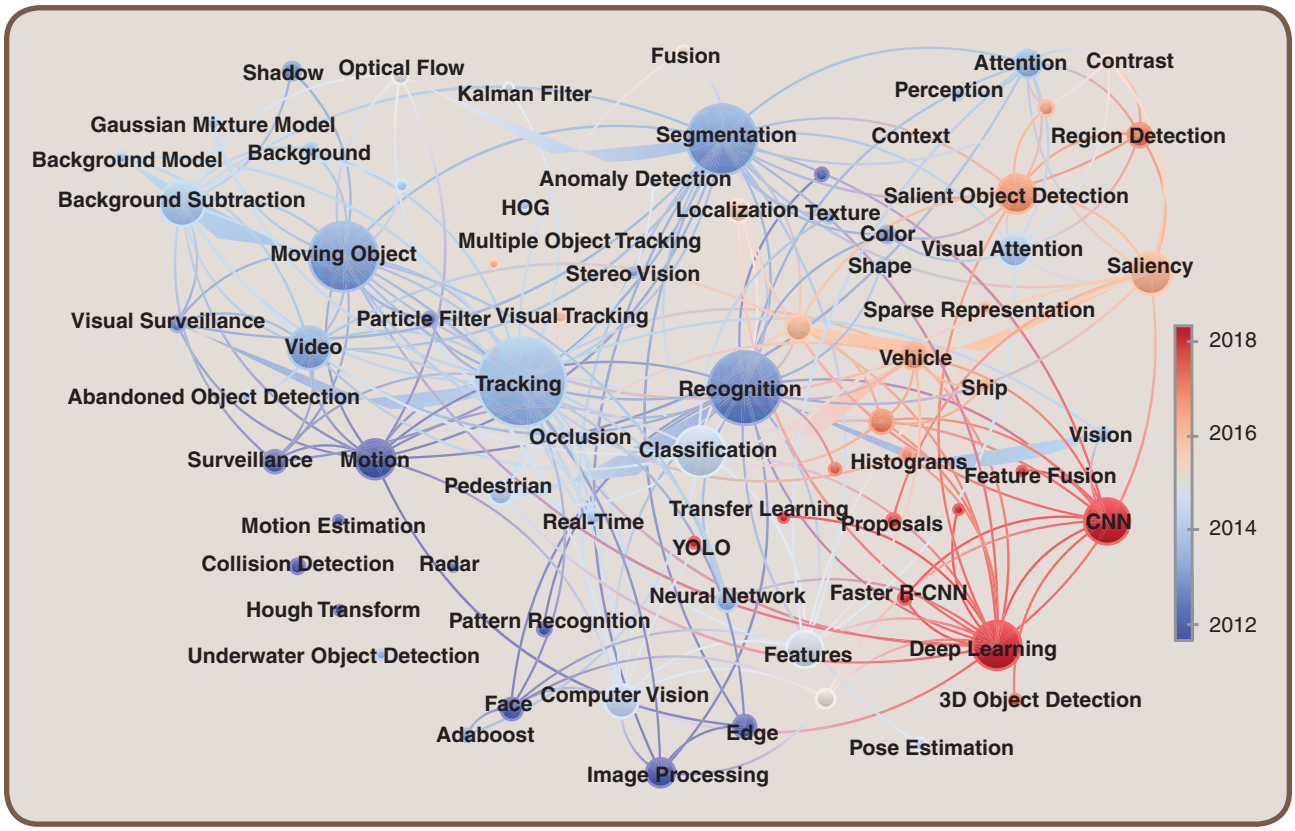

FIG 4 The keyword co-occurrence map of object detection in the Web of Science papers. The color bar in the right corner indicates the keywords' changes of research focus from blue (previous literature) to red (latest literature). The circle size is defined by the frequency of a particular keyword. HOG: histogram of oriented gradient.

(ROI) generated from radar points. The shallow and local features are shown in Figure 5, where the dotted rectangle in (a) illustrates the shallow features of the car and (b) depicts the local features.

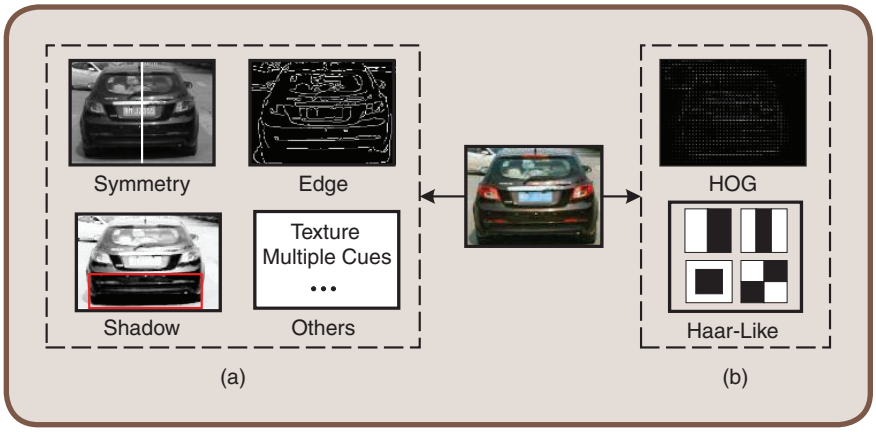

FIG 5 (a) The shallow and (b) local features of the appearance-based object detection approaches. RGB: red, green, blue.

Shallow Features

Shallow features are convenient to represent the information on a vehicle, and they can be extracted fast. These features have been widely used in on-road vehicle detection. The shallow features are presented in the following.

Symmetry

The images of many vehicles captured from the front or rear views are symmetrical over a horizontal or vertical center line, respectively. Therefore, a possible vehicle location in an image is estimated based on the highly symmetric cue [29]–[33]. The summed normalized cross correlation is used to calculate the symmetry feature needed for object detection [34]. Burlet and Dalla Fontana [35] proposed histogram search techniques to calculate the likelihood of an object being a vehicle and to evaluate the symmetry of the region in it. To reduce the computing cost, in [30]–[33], the edge was extracted to calculate the symmetry. However, symmetry is suitable to detect a vehicle from only a specific perspective, and it is highly sensitive to noise, which makes it a less-used feature in vehicle detection approaches.

Shadow

On an asphalt-paved road, the region underneath a vehicle is obviously darker than the other regions [18]. Consequently, shadow information was a significant sign pattern for vehicle detection in early research [9], [36], [37]. As one of the local illumination changes, shadow may bring many problems, such as merging, shape distortion, and a loss of objects [28]. Meanwhile, image illumination highly depends on weather conditions, which increases the difficulty of shadow segmentation. In addition, not all the shadows in an image originate from vehicles. As a result, shadow can be used to locate other possible vehicle regions [38]. With the help of shadow, the ROI of a vehicle can be located in an image, and then other vehicle detection features, such as symmetry, edge, and symmetry, can be employed in the ROI.

Edge

The different views of most vehicles, especially the rear or front views, can show horizontal and vertical edges, which are beneficial to generate the ROI. Due to low computational cost, edge detection can meet the real-time requirement of AVs. The edge-based vehicle detection approaches often use Canny [39] and Sobel [30], [40] operators to generate the edge map. The vertical and horizontal edges were used in [41] and [42], respectively, while the edge-based constraint filters were employed to segmentize vehicles from the background region. Additionally, in [43] and [44], the authors utilized a camera to retrieve contour features to detect vehicles. The edge is often combined with other shallow features in vehicle detection schemes.

Multiple Cues

Multiple features have been combined for object detection [9], [29], [31], [33], [38], [45]. In [45], four different image processing algorithms using shadow, rear lights, and line symmetry were applied to extract different features and generate the many candidates of a possible object position automatically. A belief network was designed for each object, and local approximation was used in the Bayesian inference algorithm to determine the actual position of the object. However, using multiple features can significantly increase computational complexity.

Other Cues

Unlike road surface and background objects, many vehicles have a congeneric color (that is, body and signal light color). Therefore, optical flow [46], motion stereo approach [47], [48], corners [49], and head and tail lights [50] can also be used for on-road object detection. Also, in [51], the bearing angle measurement of radar was used to detect the ambiguous BBox of the object, and a trained blob detector was employed to update the BBox in the image.

Local Feature Descriptors

Over the past two decades, there has been a transition from using simple shallow features to utilizing local features for object detection. The HOG and Haar-like features are commonly used in modern object detection approaches. The local features are shown in the right of Figure 5(b).

HOG

The HOG feature is constituted by calculating the statistical histogram of the gradient direction of the local area of an image for object detection. Even with partial occlusion, the contours of the human body and vehicle could be extracted using the HOG feature, which is not sensitive to light. The HOG features have been used to detect pedestrians and vehicles [52]–[56]. Meanwhile, the support vector machine has been used to classify HOG features.

Haar-Like

Haar-like [57] used three types (e.g., vertical, horizontal, and diagonal) of 2D nonstandard Haar wavelets to learn from example images, relying on neither priori knowledge nor motion-based segmentation. Lienhart and Maydt [58] extended the Haar-like feature prototypes for line, edge, and center-surround features. Meanwhile, by using the entire image, fast feature extraction was achieved, allowing for real-time object detection on a common CPU. An AdaBoost classifier was used to classify Haar-like features; this approach has been popular in both face and vehicle detection. Chang and Cho [59] employed online boosting, which is an online AdaBoost scheme combined with a cascade of strong classifiers. The cascade of strong classifiers can be trained online to adapt to changing traffic environments. Analogously, a number of related studies [60]–[62] have used AdaBoost to classify Haar-like features.

DL-Based Object Detection Approaches

With the tremendous evolution of DL-based image classification, object detection techniques have been improved significantly, especially real-time object detection techniques. Common neural networks (NNs) have been employed for object detection in early RV fusion research. For instance, Ji and Prokhorov [63] adopted a multilayer in-place learning network, an all-purpose classification, and a regression network to divide the research objects into two classes: vehicles and nonvehicles. A recurrent NN was used in [64] to predicate pedestrian-crossing intention in one or more time moments in the future. Königshof and Stiller [65] estimated the exact dimensions of a 3D box with the help of a lightweight convolutional NN (CNN). A three-layer NN was used in [66] to learn deep information from the areas extracted by a radar, which allowed for incremental and online learning. A multiscale CNN was employed in [67] for object detection.

ImageNet, with deep CNNs [68], achieved record-breaking results in image classification. Since then, deeper CNNs have led to significant progress in object detection, resulting in a remarkable region-based CNN (R-CNN) [69]. Thanks to contributions in DL, many real-time object detection frameworks and codes have been developed, and they could be used in combination with the perceptual systems of AVs in the future.

This section focuses on the real-time performance of object detection approaches. As shown in Figure 6, DL-based object detection frameworks can be generally divided into two main categories: two- and one-stage detectors [70]. Two-stage detectors, which are also called *region-based detectors*, generate object proposals before classification and thus reduce the time expense. One-stage detectors, also called *region proposal-free detectors*, do not preprocess detection proposal.

Two-Stage Detectors

Many two-stage detectors have been developed in recent years, including R-CNN [69], Fast R-CNN [71], Faster R-CNN [72], R-FCN [73], and Mask R-CNN [74]. Dimitrievski et al. [75] used the Faster R-CNN [72] to detect people in an image. However, two-stage approaches are computationally expensive for AVs, which have limited storage and computational capability. Although advanced object detection methods achieve higher accuracy, they are too slow to be employed in real-time on-road scenarios.

One-Stage Detectors

Instead of attempting to improve the individual components of an intricate region-based pipeline, researchers have begun to develop one-stage detection strategies. In particular, a one-stage detector directly predicts the BBox offsets and class probabilities from the entire image without performing feature resampling and area proposal generation, investigating all of the computations in the same network. As the whole pipeline is a single network, the detection performance can be directly optimized in an end-to-end manner, such as in you only look once (YOLO) [100], YOLO9000 [101], YOLOv3 [96], YOLOv4 [97], single shot multibox detector (SSD) [93], CornerNet [77], CenterNet [102], Efficientdet [70], and TTFNet [95]. These models can be trained by the on-road image data and deployed in AVs in the future.

A comparison of the aforementioned detectors and state-of-the-art approaches is given in Table 1. The detection speed and accuracy are both influenced by backbones and schemes. The two-stage detectors are more flexible and more accurate than one-stage detectors, which tend to be simpler and more efficient because they leverage predefined anchors [103]. The corresponding evaluations are presented in Table 2.

Radar-Based Object Detection

Radar, which is one of the most fundamental and important sensors in ADASs, can be integrated into many AVs. Radar-based object detection uses mm-waves to determine relative information about the detection object (e.g., distance, velocity, and angle) based on the Doppler effect. The built-in algorithms equipped in radar are used to parse the range, direction, and Doppler velocity of objects. The all-weather mm-wave radar sensor described in [104] used frequency-modulated continuous wave radar to detect and track obstacles in the front field. Two consecutive fast Fourier transforms (FFTs) were adopted in [106] to obtain the time and space dimensions of a signal. Further, the guardrail and steel plates on the road have large radar cross sections, which can influence radar-returned signals and make essential

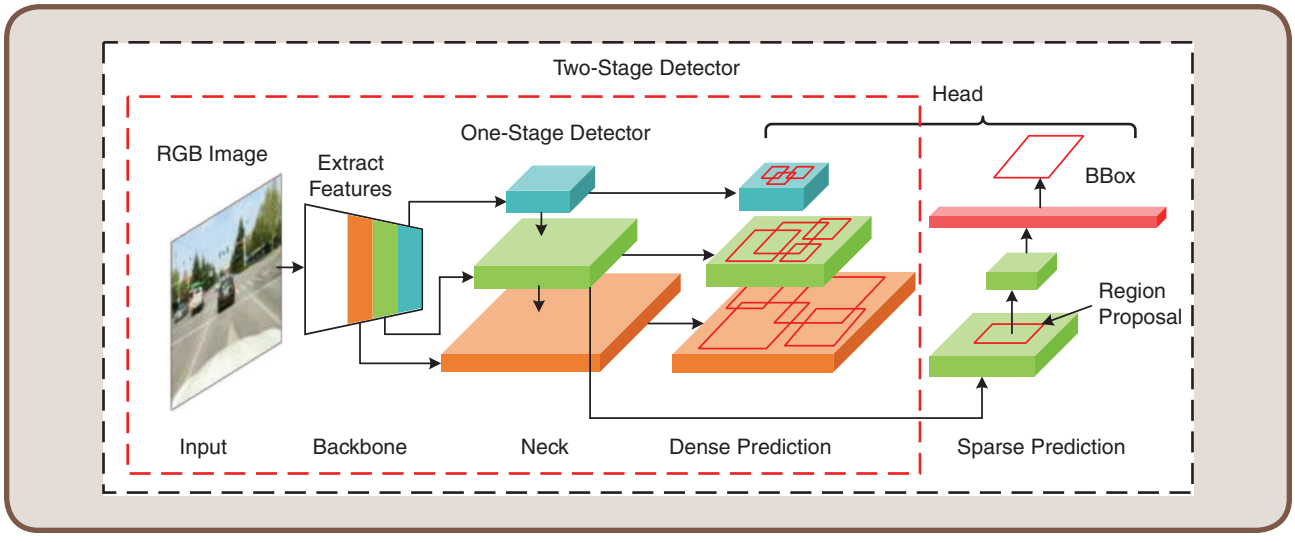

FIG 6 A deep learning object detector consisting of a one- and two-stage detector.

Table 1. The performances of the public real-time object detection approaches on the Common Objects in Context (COCO) data set [more than 30 frames per second (fps)].

	Approach	Backbone	Resolution	Fps	Average Precision (%)	AP50 (%)
One stage	CenterMask-ite [76]	VoVNetV2-39-FPN [76]	600 × 600	35.7	40.7	—
	CornerNet-Squeeze [77]	Hourglass-54 [77]	255 × 255	33	34.4	—
	CenterNet [78]	ResNet101 [79]	512 × 512	45	34.6	53
	DAFS [80]	VGG16 [81]	512 × 512	35	33.8	52.9
	EFM (SAM) [82]	CSPPeleeNet [83]	512 × 512	109	27.6	50.4
	EFGRNet [84]	VGG16 [81]	320 × 320	48	33.2	53.4
	EfficientDet-D2 [70]	Efficient-B2 [70]	768 × 768	56.5	43	62.3
	HSD [85]	VGG16 [81]	320 × 320	40	33.5	53.2
	LRF [86]	ResNet101 [79]	512 × 512	31	37.3	58.5
	PANet (SPP) [87]	CSPResNeXt50 [83]	608 × 608	35	38.4	60.6
	Pelee [88]	PeleeNet [88]	304 × 304	106	22.4	38.3
	PFPNet (R) [89]	VGG16 [81]	320 × 320	33	31.8	52.9
	PP-YOLO [90]	ResNet50-vd-dcn [90]	608 × 608	72.9	45.2	65.2
	PRN [82]	PeleeNet [88]	416 × 416	145	23.3	45
	RefineDet [91]	VGG16 [81]	320 × 320	40	29.4	49.2
	RFBNet [92]	VGG16 [81]	512 × 512	35	33.8	54.2
	SSD [93]	HarDNet85 [94]	512 × 512	32	35.1	54.8
	TTFNet [95]	DarkNet53 [96]	512 × 512	54	35.1	52.5
	YOLOv3 [96]	DarkNet53 [96]	608 × 608	30	33	57.9
	YOLOv3 (SPP) [96]	DarkNet53 [96]	608 × 608	30	36.2	60.6
	YOLOv3 (tiny) [96]	DarkNet [96]	416 × 416	330	—	33.1
	YOLOv4 [97]	CSPDarknet53 [83]	608 × 608	33	43.5	65.7
	YOLOv4 (tiny) [97]	CSPDarknet53 [83]	416 × 416	75	—	40.2
	Refine YOLOv4 [98]	CSPDarknet53 [83]	960 × 540	38	—	67.7
Two stage	ThunderNet [99]	SNet535 [99]	320 × 320	214	28	46.2

The table is expanded and improved from CSPNet [83] and focuses mainly on fps and AP50. All of the approaches are developed on GTX 1080ti (batch = 1), except for Refine YOLOv4 [98] (Tesla K80), EfficientDet-D2 (Tesla V100), and PP-YOLO (Tesla V100). All of the results are obtained using the COCO test-dev set, except for those of the TTFNet [95] (minval5k) and Refine YOLOv4 [98], where the UA-DETRAC data set is used. ResNet: residual neural network. DAFS: dynamic anchor feature selection; SAM: Spatial Attention Module; CSPNet: Cross Stage Partial Network; EFGRNet: enriched feature guided refinement network; HSD: hierarchical shot detector; VGG: Visual Geometry Group; LRF-NET: Learning Rich Features; PANet: Path Aggregation Network; SPP: spatial pyramid pooling; PFPNet: parallel feature pyramid network; PP-YOLO: PaddlePaddle–You Only Look Once; PRN: partial residual networks; RFB: Receptive Field Block Net; TTFNet: training time friendly network; SNet: ShuffleNetV2.

Table 2. The evaluation metric of object detection.

Metric	Perfect (%)	Better	Description
Average precision (AP)	100%	↑	Mean AP [the AP averaged more than 10 intersections over union (IOU) :0.5:0.05:0.95]
AP50	100%	↑	Mean AP (the AP at the IOU is equal to 0.5)
(Fps)	Inf.	↑	Processing speed on the benchmark (fps)

"↑" indicates that a higher number denotes a better performance, while "perfect" means the score of the best performance, and "better" means a better performance. Inf: infinite number +∞.

objects such as pedestrians and motorcycles invisible. An asymptotic and full-physics electromagnetics solver was used in [107] to reduce the radar cross section of steel plates and guardrails. By using these built-in algorithms, the relative position and velocity of in-front objects can be determined.

However, radar often makes a false detection and regards an irrelevant target as a relevant one [107]. False detection can cause emergency braking, which can affect driving comfort and even cause traffic accidents. To reduce the number of false detections and redundant detections, additional filter and cluster approaches are commonly used to correct radar measurement data; for instance,

these measurements are processed by the algorithms built into a radar. In this way, radar-based object detection becomes more reliable and stable. In [108], an mm-wave radar system was designed; this system used an FFT and a Kalman filter (KF) to estimate the relative distance and velocity, respectively, of an object. Similar to [106] and [108], Botha et al. [48] processed the radar measurement data by a 2D Fourier analysis and used a nonlinear probability hypothesis density (PHD) filter to estimate the radar's state. Hsu et al. [62] used the density-based spatial clustering of applications with the noise (DBSCAN) to cluster true objects and remove false objects. Consequently, by using additional filter and cluster approaches, many false detections can be eliminated.

Recently, deep NNs (DNNs) have been playing a predominant role in the radar field [114]. DNNs include CNNs [115], RNNs [116], and long short-term memory (LSTM) networks [117]. Many lidar-based DNN techniques (e.g., PointNet [118] and PointNet++ [119]) have been improved for radar-based object detection. In [118], the PointNets were adjusted for radar data, which performed 2D object classification using segmentation and 2D BBox regression. Also, many methods have been motivated by image-processing CNN-based techniques, such as YOLO [100]. YOLO was successfully trained to classify and localize objects in the radar domain [120].

Object Tracking

According to the number of tracking objects, tracking algorithms can be divided into single- and multiobject tracking (MOT) algorithms. A MOT algorithm is a complex form of single-object tracking. The main solution paradigms for MOT include multiple hypothesis tracking, joint probabilistic data association, and random finite sets (RFSs) [121]. Due to the complex and crowded road environment, MOT algorithms underpin crucial application importance in AVs. MOT algorithms aim to construct and predict the trajectories of multiple on-road objects in video sequences and radar measurement data while conserving object identities. These algorithms rely mainly on features and states (i.e., position and velocity) of multiple objects between consecutive frames or radar points. This section reviews the conventional object tracking approaches that have been reported in RV-related literature. In recent years, DL and CNNs have been widely used in MOT techniques, and visual tracking has become the mainstream research direction, as illustrated in Figure 7.

For this reason, the following sections introduce state-of-the-art, DL-based MOT strategies. These approaches have been evaluated on several benchmark data sets, including MOT [110] and KITTI [122], and their performances are

Table 3. The performances of the public real-time online object tracking approaches on the KITTI data set [105] [more than 30 frames per second (fps)].

Tracker	MOTA (%)	MOTP (%)	MT (%)	ML (%)	IDs	Frag	Fps
Car							
ExtraCK	79.99	82.46	62.15	5.54	343	938	33.3
RMOT*	65.83	75.42	40.15	9.69	209	727	50
IWNCC	86.86	85.39	75.38	2.92	130	521	100
MASS	85.04	85.53	74.31	2.77	301	744	100
TDFL*	84.3	85.63	71.69	2.92	328	815	100
mbodSSP*	72.69	78.75	48.77	8.77	114	858	100
mbodSSP	56.03	77.52	23.23	27.23	0	699	100
RMOT	52.42	75.18	21.69	31.85	50	376	100
HM	43.85	78.34	12.46	39.54	12	571	100
Pedestrian							
HWFD	67.27	74	44.67	22.34	116	918	33.3
RMOT*	43.77	71.02	19.59	41.24	153	748	50
RMOT	34.54	68.06	14.43	47.42	81	685	100

"*" indicates regionlets that were used as detections; MOTA: multiobject tracking accuracy; MOTP: multiobject tracking precision; MT: mostly tracked; ML: mostly lost; IDs: the number of identity switches; frag: the number of times a trajectory is fragmented or interrupted during tracking; ExtraCK: fast MO extrapolation tracker; RMOT: Bayesian multi-object tracking using motion context from multiple objects; MASS: multiple object tracking with attention to appearance, structure, motion, and size; HM: Hungarian algorithm.

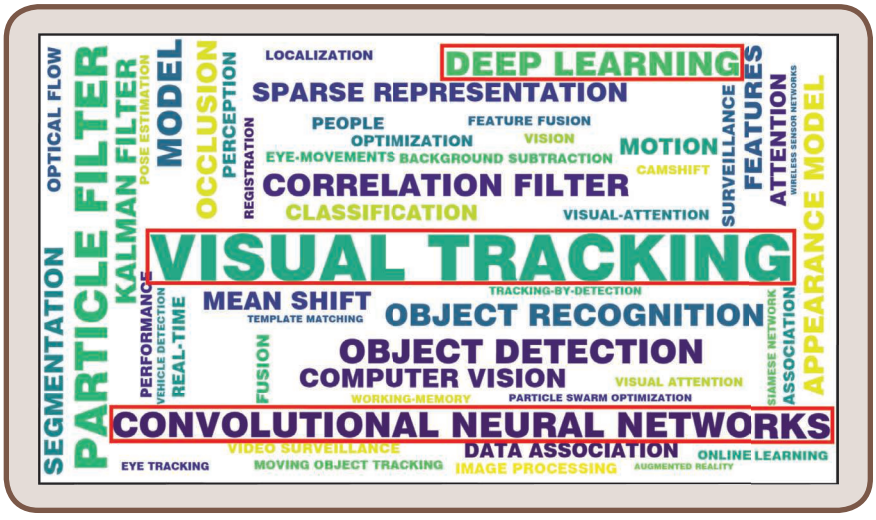

FIG 7 The most frequent keywords in the Web of Science papers in the object tracking field published from 2015 to 2020. DL and CNNs have gained more attention in recent years. The larger the font, the more studies on that topic have been presented in this work.

presented in Tables 3 and 4, while the corresponding evaluations are listed in Table 5.

Conventional MOT

Generally, on-road objects represent different temporal and appearance information; therefore, valuable features should be extracted before tracking. The existing visual methods depend mainly on the representations of histogram, area, edge or contour, and model, and the radar tracking approaches are based mainly on velocity and position. The objects can be tracked from frame to frame based on the similarity in visual appearance or the kinematical characteristics between consecutive radar measurement data. As a result, object tracking approaches can be divided into two classes: prior knowledge and probability based.

Prior-Knowledge-Based Approaches

Prior-knowledge-based approaches perform data matching according to features' similarity between consecutive frames to realize object tracking. These features include the shallow and local features, as explained in the "Object Detection" section. The representations of histogram [6], [9], edge [40], and region [6] are used to match objects and tracks in consecutive frames. Image histogram has been widely used in various image processing applications, especially color-based image retrieval and image classification, due to its low computational cost and image invariability of translation, rotation, and zoom. In addition to the histogram, an edge or a contour can also be used for object tracking. Srinivasa et al. [40] used the three-component matching metric, including the Euclidean distance criterion, edge density, and sum-of-square-of-difference in the intensity of pixels to track detected vehicles.

Moreover, in [123], Haar-like features were used to match two consecutive frames for data association, and a constant-velocity KF was used for tracking. Chavez-Garcia and Aycard [53] modified the model-based object detection approach, where the Markov chain Monte Carlo sampling process was used to find the best trajectories. In addition, in [124], a vehicle light tracking method based on temporal trajectory analysis was proposed to track the objects in an image.

Table 4. The performances of the public real-time online object tracking approaches on the MOT data set [109] [more than 30 frames per second (fps)].

Tracker	MOTA (%)	MOTP (%)	MT (%)	ML (%)	IDs	Frag	Fps
2D MOT 2015 [110]							
FairMOT [111]	59	76.8	45.6	11.5	521	1,747	30.5
EDA_GNN	21.8 (±13.8)	70.5	9	40.2	1,488	1,851	56.4
GMMA_intp	27.3 (±12.1)	70.9	6.5	43.1	987	1,848	132.5
RNN_LSTM	19 (±18.1)	71	5.5	45.6	1,490	2,081	165.2
GMPHD_OGM	30.7 (±12.6)	71.6	11.5	38.1	1,034	1,351	169.5
DEEPDA_MOT	22.5 (±0)	70.9	6.4	62	1,159	1,538	172.8
SORT [112]	33.4 (±0)	72.1	11.7	30.9	1,001	1,764	260
AdTobKF	24.8 (±12.4)	70.8	4	52	666	1,300	206.5
C++SORT	21.7 (±11.8)	71.2	3.7	49.1	1,231	2,005	1,112.1
MOT 2016 [113]							
GMPHD_Re-ID	40.4 (±9.3)	75.2	11.2	43.3	792	2,529	31.6
C++SORT	31.5 (±9)	77.3	4.3	59.9	1,587	2,239	687.1
EAGS16	47.4 (±8.6)	75.9	17.3	42.7	575	913	197.3
MOT 2017 [113]							
GMPHD_Rd17	46.8 (±14.7)	76.4	19.7	33.3	3,865	8,097	30.8
GM_PHD	36.4 (±14)	76.2	4.1	57.3	4,607	11,317	38.4
EDA_GNN	45.5 (±13.8)	76.3	15.6	40.6	4,091	5,579	39.3
SORT17	43.1 (±13.3)	77.8	12.5	42.3	4,852	7,127	143.3

MOTA: multiobject tracking accuracy; MOTP: multiobject tracking precision; MT: mostly tracked; ML: mostly lost; IDs: the number of identity switches; frag: the number of times a trajectory is fragmented or interrupted during tracking. EDA_GNN: Neural Network for Online Multiple-Object Tracking; GMMA: GMPHD Filter with Group Management and Relative Motion Analysis; RNN_LSTM: Recurrent Neural Network Long Short Term Memory; GMPHD_OGM: GMPHD Filter and Occlusion Group Management; DEEPDA_MOT: data association for multi-object tracking via deep neural networks; AdTobKF: Adaptive Tobit Kalman-based tracking; GMPHD_Rd17: occlusion-robust online multi-object visual tracking using a GM-PHD filter with a CNN-based re-identification; GMPHD: Gaussian mixture probability hypothesis density filter; SORT17: simple online and realtime tracking.

Probability-Based Approaches

Probability-based approaches typically represent the states of objects as a distribution with a specific uncertainty. The goal of these tracking algorithms is to estimate the probabilistic distribution of the object's state using a variety of probability-reasoning techniques based on the existing observations, such as past and present ones [125]. Various probability-based strategies have been applied to object tracking, including KFs, extended KFs (EKFs), particle filters (PFs), Bernoulli filters (BFs), and many others. In the conventional tracking paradigm, after detection, the measurements must be associated with tracks, and the state of a target of interest is updated by a KF or EKF. However, it is difficult to associate measurements and tracks in complex scenes because there are many objects to track, e.g., guardrail vehicles and pedestrians. A PHD filter using RFSs is a promising solution for MOT in complex scenes [126].

KFs

As a linear quadratic estimation, a KF uses a strand of measurements observed over time, including noise and other inaccuracies, to estimate unknown variables [127]. In general, these estimations are more accurate than those based on a single measurement. Moreover, KFs can take advantage of the historical information and decrease the search space of an image, thus markedly improving processing speed. Wang et al. [128] employed a single-pedestrian tracking approach based on a KF, and multipedestrian tracking using a maximum a posteriori probability-based association was used in [129]. KFs have been widely employed in the object tracking field [9], [29], [32], [38], [47], [130]. Under the linear Gaussian model, a KF can update the optimal estimation of the object's state and thus achieve better tracking accuracy [131]. To deal with a nonlinear filtering problem, an EKF expands a nonlinear function of a system by a first-order Taylor to obtain the linearized system equation so as to complete the filtering estimation and other processing of the object's data. In [35], [43], [132], and [133], the EKF was used to track the objects detected by camera or radar.

PFs

A PF, which represents a Bayesian sequential-importance sampling technique, uses a finite set of weighted samples to approximate the posterior distribution recursively [134]. A PF can be easily implemented and provides an effective solution for the analysis of nonlinear dynamic systems, which is why it has attracted wide attention in the object tracking field [62], [75]. The core ideology of a PF is to take advantage of a column of random samples with associated weights and make an estimation based on them and then to represent the posterior probability density.

BFs

A simple Bayesian formulation based on RFSs is provided in the RFS theory. A dominant trend in an RFS-based MOT is to develop multiobject conjugate priors. A BF is an RFS-based filter for non-Gaussian, nonlinear dynamic systems, which is especially suitable for solving the problems of objects appearing and disappearing during object tracking. BFs include the multi-Bernoulli mixture (MBM) [121], Poisson MBM (PMBM) [135], [136], and two-trajectory Poisson MB (PMB) filters [137]. To solve the problem of performance degradation when object plotting arises, a PMB filter with spawning was proposed in [138]. An MBM describes the distribution of detected objects, while a PMBM filter provides a closed-form solution for multiple extended object filtering with standard models. The PMBM conjugate priors are versatile for point objects and extended objects and have been applied to the data obtained from lidar, radar, or camera. They have also been successfully applied to the tracking of moving objects and the mapping of stationary objects [139]. Therefore, PMBM filters are suitable for object tracking.

Other Filters

There are a number of object tracking schemes that are not widely used but are still effective, such as an extended RFS-based multiobject target tracking approach [140] and a Gaussian mixture of integrated probability data association [51] and Gaussian-mixture PHD filters [141], [142]. Scheel and Dietmayer [143] presented a new variational radar model for vehicle tracking based on radar detection data. This model was trained by actual radar data using variational Gaussian mixtures and avoided excessive manual engineering. The Kanade–Lucas–Tomasi feature was used in [144] to track the preceding vehicle in an image. Additionally, a joint-integrated probabilistic data association filter was used to track pedestrians in the front view and blind regions [52].

DL-Based MOT

In the past decades, MOT has been extensively studied in the CV field, following tremendous breakthroughs in DL. The following sections focus mainly on vision- and radar-based MOT. In addition, some laboratories, whose research

Table 5. The evaluation metrics used in object tracking.

Metric	Perfect (%)	Better	Description
MOTA	100	↑	Multiobject tracking accuracy, combining three error sources: false positives, missed targets, and identity switches (IDs).
MOTP	100	↑	Multiobject tracking precision; the overlap between the annotated and the predicted BBoxs.
MT	100	↑	Mostly tracked objects; the percentage of the ground-truth trajectories that are covered by the tracker hypothesis for at least 80% of their length.
ML	0	↓	Mostly lost objects; the percentage of the ground-truth trajectories that are covered by the tracker hypothesis for, at most, 20% of their length.
IDs	0	↓	The number of IDs; the total number of times that a trajectory changes its matched ground-truth identity.
Frag	0	↓	The number of times that a trajectory is fragmented or interrupted during tracking.
Frames per second (fps)	Inf.	↑	The processing speed on the benchmark; the fps, excluding the detector.

"↑" indicates that a higher number denotes a better performance, and ↓ means the opposite; while "perfect" means the score of the best performance, and "better" means a better performance.

interests include object detection and/or object tracking, are listed in Table 6.

Vision-Based MOT

Vision-based MOT can be divided into two-step and one-shot approaches [111]. Two-step methods, also named *tracking by-detection approaches*, divide the MOT process into two steps: object detection and tracking, as shown in Figure 8(a) and (b). The two-step MOT outlooks use separate detectors and re-identification (Re-ID), where Re-ID extracts features of every detection for data association, whereas one-shot MOT approaches integrate detectors and Re-ID into a unified NN.

Two-Step Approaches

The two-step approaches have a higher accuracy than the one-shot techniques. These schemes include TrackletNet Tracker [145], near-online multi-target tracking [146], and online multi-object tracking with dual matching attention networks [147]. In the two-step approaches, first, the CNN-based detectors, such as the YOLO and R-CNN series, are employed to localize the BBoxs of all the objects in an image. Object detection can be conducted using accurate detectors, such as EfficientDet [70], R-CNN [69], Fast-R-CNN [71], Faster-R-CNN [72], Mask-R-CNN [74], and RFCN [73], or fast detectors, such as YOLO [100], YOLO9000 [101], YOLOv3 [96], YOLOv4 [97], and SSD [93]. Next, the images in the BBoxs are cropped and fed to the embedding model to extract the Re-ID features, and the BBoxs are then assigned a suitable existing track. Deep SORT [148], an object tracker, integrates deep appearance representations to improve the performance of the SORT [112]. Kapania et al. [149] used the YOLOv3 RetinaNet detector and Deep SORT tracker to achieve object detection and tracking, respectively.

Table 6. The summary of laboratory research object detection and/or tracking.

Laboratory/University	Research	Leader	URL
Karlsruher Institut für Technologie	Detection	Christoph Stiller	https://www.mrt.kit.edu/mitarbeiter_stiller.php
The Chinese University of Hong Kong MMLab	Detection	Dahua Lin	http://www.dahua.me/
University of Washington and Allen Institute for AI and Facebook AI Research	Detection	Joseph Redmon	http://pjreddie.com/yolo/
Visual Computing Group of Microsoft Research and Facebook AI Research	Detection	Kaiming He	http://kaiminghe.com/
Google Research, Brain Team	Detection	—	https://opensource.google/
Signal Processing group of Chalmers University of Technology	Tracking	Karl Granström	https://www.chalmers.se/en/staff/Pages/karl-granstrom.aspx
und Regelungstechnik	Tracking	Klaus Dietmayer	https://www.uni-ulm.de/in/mrm
Department of Electrical and Computer Engineering Curtin University	Tracking	Ba Tuong Vo	http://ba-tuong.vo-au.com/
Shanghai Jiao Tong University and ZTE Corp	Tracking	Weixiao Lin	https://motchallenge.net/approach/MOT=756&chl=5
The Australian National University	Tracking	Hongdong Li	http://users.cecs.anu.edu.au/~hongdong/
Visual Computing Lab, Hong Kong Polytechnic University	Tracking	Lei Zhang	http://www4.comp.polyu.edu.hk/~cslzhang/
Chinese Academy of Sciences, National Laboratory of Pattern Recognition	Tracking	Tianzhu Zhang	http://nlpr-web.ia.ac.cn/mmc/homepage/tzzhang/index.html
Vision and Learning Lab, University of California, Merced	Tracking	Ming-Hsuan Yang	https://faculty.ucmerced.edu/mhyang/
Visual Geometry Group, University of Oxford	Tracking	João F. Henriques	https://www.robots.ox.ac.uk/~joao/
Torr Vision Group, University of Oxford	Tracking	Philip Torr	http://www.robots.ox.ac.uk/~phst/
Tencent AI Lab	Tracking	Yibing Song	https://ybsong00.github.io/
E.T.	Tracking	—	https://www.taobao.com/markets/cnwww/lab-xiao-g
Alibaba DAMO Academy	Detection and Tracking	Gang Wang	https://damo.alibaba.com/labs/intelligent-transportation
Mobileye	Detection and Tracking	—	https://www.mobileye.com/
MINIEYE	Detection and Tracking	—	https://www.minieye.cc/
Megvii	Detection and Tracking	Gang Yu	http://www.skicyyu.org/
SenseTime	Detection and Tracking	Xiaogang Wang	http://www.ee.cuhk.edu.hk/%7Exgwang/
Visual Cognitive Systems Laboratory, University of Ljubljana	Detection and Tracking	Matej Kristan	http://www.vicos.si/matejk

DAMO: Discovery, Adventure, Momentum, and Outlook.

In summation, different detectors and trackers can be combined to achieve the desired performance.

One-Shot Approaches

The one-shot approaches are faster than the two-step ones [111]. The key principle of the one-shot perspective is to integrate object detection and identity embedding (Re-ID features) into a united network to reduce the inference time by sharing the same set of low-level features. In the joint detection and embedding (JDE) strategy [165], a shared model was used for object detection and feature embedding. Their combination results in a near-real-time MOT approach, with a fast speed of 18.8–24.1 frames per second (fps), which depends on the input's image resolution. Similar to JDE, Track-R-CNN [166] also jointly executed object detection and identified feature embedding. More specifically, it [166] employed Re-ID as the head of the Mask-R-CNN, and the BBox and Re-ID features were regressed for every proposal. The DeepMOT-Tractor [167] achieved a multiple object tracking accuracy of 54.8 and 53.7% on MOT16 and MOT17, respectively.

However, the tracking accuracy of one-shot approaches is usually lower than that of the two-step ones. To address this problem, a simple anchor-free scheme for a one-shot MOT was proposed in [111], which outperformed the previous state-of-the-art procedures on several public data sets at 30 fps.

Radar-Based MOT

In recent years, DL-based MOT has become popular. DL is useful to reduce noise influence stochastically. As a result, scholars have been researching the possibility of environmental recognition by machine learning for mm-wave radar. Ebert et al. [168] proposed a DNN-based data-driven approach, which can be implemented into the existing tracking frameworks, such as KFs or EKFs. A DNN architecture (Radar TrackNet) was proposed in [119]; this architecture used radar point clouds to detect road users and calculate their tracking information. Akita and Mita [169] used an LSTM for the classification and tracking of target objects and achieved a peak accuracy rate of 98.67%. They concluded that a bidirectional LSTM could achieve better accuracy than the original LSTM when only the data in the forward time direction was used. In addition, an LSTM-based, data-driven DNN (DeepDA) was designed to learn the measurement-to-track association probability from the radar's noisy measurement data and the existing tracks [170]. Compared with other algorithms, DeepDA has less time consumption but higher computational efficiency.

RV Fusion

Sensor fusion integrates the information obtained by multiple homogeneous or heterogeneous sensors to avoid the perceptual uncertainties and limitations of individual

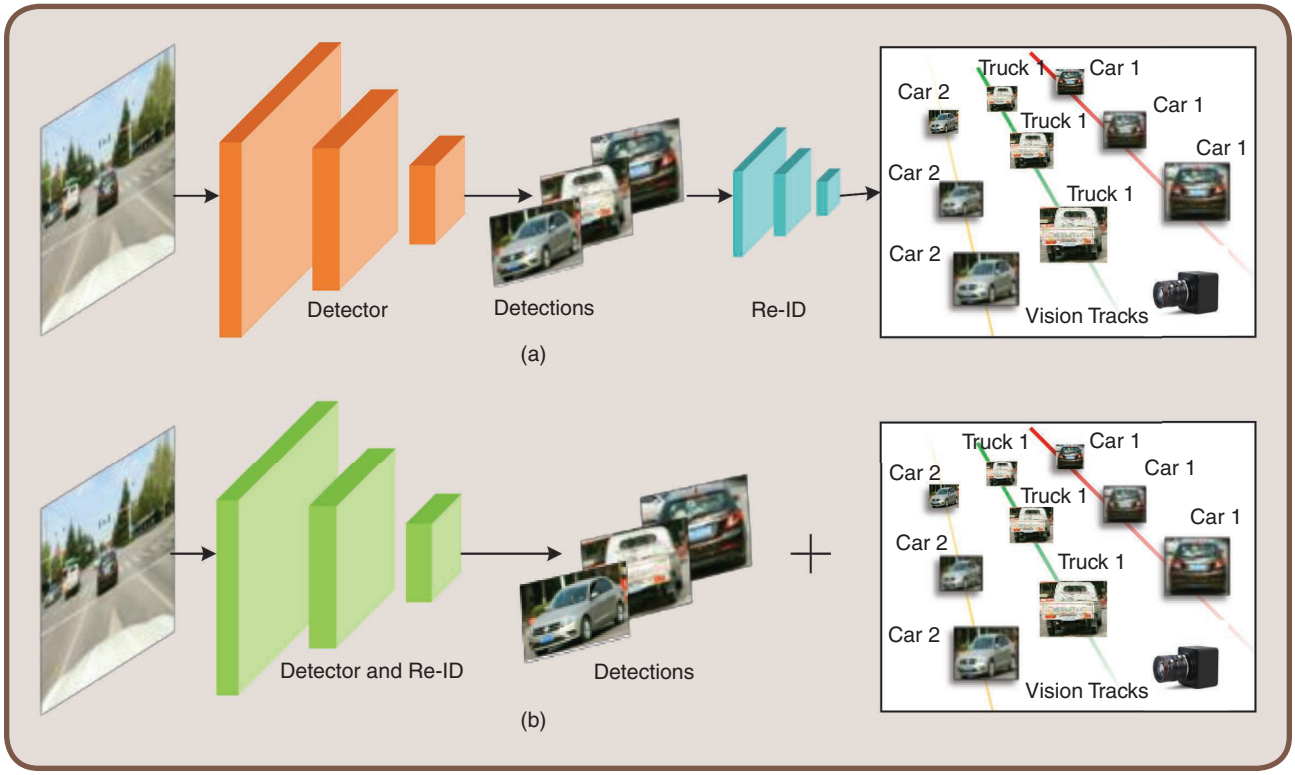

FIG 8 The architecture of the (a) two-stage and (b) one-shot MOT approaches. Re-ID: re-identification.

A region's proposal network is a bottleneck in most visual object detection networks, increasing the processing time and resulting in slow networks.

sensors [171], [172]. The data obtained from sensors are equally and heavily relied on, regardless of the sensor type. RV fusion systems require the use of a decision framework to produce reliable outputs when a sensor fails [158]. As long as an object is detected by at least one sensor, the fusion decision framework can track the object with reliable accuracy, but the detection accuracy is higher when the object is detected by two sensors. In this way, environment perception and object recognition become more comprehensive, and a vehicle's external perception ability is improved. With high precision, reliability, and robustness to uncertainty, sensor fusion offers extended spatial and temporal coverage, which are essential in the safety systems of AVs.

Meanwhile, due to the different sampling rates and fields of view of sensors, it is necessary to align these sensors in time and space. Homography was used in [51] to map radar and camera. Ren et al. [130] used Zhang's calibration algorithm [173] to match the centroid coordinates of both video and radar objects. However, high-accuracy radar and vision-based projection algorithms need further research.

This section comprehensively introduces the RV fusion frameworks of the object detection and tracking approaches. A comparison of the various RV fusion methods in terms of their fps values is presented in Table 7, where the detector, tracker, and fusion strategies of each of the schemes are listed. The advantages and disadvantages of the common RV fusion schemes are summarized in Table 8.

Object Detection By RV Fusion

With respect to the input data and fusion frameworks, there are different RV fusion strategies for object detection. The fusion approaches presented in this article include two types: ROI based and end to end. However, the reliability and accuracy of RV fusion systems heavily depend on the fusion framework utilized. ROI-based fusion strategies may be able to detect objects based on the proposed ROI, using only one of the sensors. When a sensor fails, end-to-end fusion strategies may not be available if the network was trained on both inputs.

ROI-Based Fusion Strategies

ROI-based fusion refers to the fusion process in which the ROI of a distinguishable object are first detected in an image obtained by a radar or camera, with the information (e.g., the velocity, position, and features) obtained after fusion further is tracked. Generally, ROI can be generated from the radar points and BBoxs of the objects detected in an input image. Therefore, ROI-based fusion strategies can be divided into two classes: radar ROI and image BBox.

Radar ROI

Modern high-resolution radar sensors generate multiple radar targets per object, which makes them particularly suitable for 2D object detection tasks [118]. Radar data can help significantly in developing more robust detection networks and achieving performance improvement. To reduce the number of anchor searches, radar position coordinates are converted to pixel coordinates, which are considered ROI. These represent the input of both the DL- and appearance-based frameworks used for detection or classification, respectively. The various object detection methods introduced in the "Object Detection" section can be applied to the ROI obtained from the radar measurement data [53]–[56], [66], [128], [150], such as HOG [55], [56], [150] and symmetry [31], [33]. A simple illustration of the radar ROI fusion framework is displayed in Figure 9(a), where radar ROI is generated from a rectangular box projected onto the red, green, blue (RGB) image by the radar points. Image patches are fed to the input of the object classification model, such as the DL-based approach, which is presented in Figure 10, or the appearance-based technique, presented in Figure 4. Then, the classified objects are mapped with the corresponding radar measurement data.

It is worth noting that the size of the rectangular box depends not only on the object's type but also on the distance between the ego vehicle and the front object. According to the focal perspective of the lens, a fixed-size rectangular box in the world coordinate system is automatically resized to obtain an ROI in an image, based on the distance. The attention selection provided in [66] generates a fixed-size rectangular ROI of 3.0 m × 3.8 m for every valid radar object. The fixed-size ROI of 5 m × 4 m was used in [60] and [61]. In contrast, Kim and Song [47] suggested a gating algorithm for data association, where the gate size increased when the tracking objects were positioned relatively far apart. Another problem is that not all ROI are correct because many false positives can occur due to the presence of other objects in an image. Therefore, a filter should be used to remove false-positive ROI. To the authors' knowledge, this scenario also happens when the same object returns several radar points, so an additional step is required to combine similar ROI and remove redundant ones. When the ROI of the same radar overlap, the largest box is preferred [174], and then, an

average ROI is computed from two ROI with similar sizes and the same bottoms.

A region's proposal network is a bottleneck in most visual object detection networks, increasing the processing time and resulting in slow networks. Radar point-based region proposal algorithms for object detection, such as radar PointNets [118], RANet[156], BIRANet[156], and RRPN a radar region proposal network (RRPN) [153], can meet the

Table 7. A comparison of the various RV fusion approaches.

Year	Reference	Detection Approaches	Fusion	Fps	Device
		Object Detection By RV Fusion			
2002	[45]	Vehicle: radar, symmetry, shadow, and rear light	Causal structure	25	—
2007	[30]	Vehicle: symmetry Guard rail: line searching	Radar ROI	40	Electronic control unit (ECU)
2008	[31]	Vehicle: vertical symmetry and shadow Pedestrian: vertical symmetry and motion stereo	Radar ROI	7	Pentium 4 at 3 GHz
2010	[66]	Vehicle: three-layer NN	Radar ROI	15.12	2.4-GHz Intel Core 2, 4 GB of memory
2011	[36]	Vehicle: edge and shadow	Radar ROI	> 10	1.8-GHz Core Duo CPU
2011	[37]	Vehicle: shadow	Visual BBox and radar ROI	15	AMD Athlon 64 3000+
2012	[46]	Vehicle: optical flow	Radar ROI	10	ECU
2014	[51]	Pedestrian: homography matrix	Radar ROI	500	Intel Core 2 Duo, 2.67 GHz
2014	[6]	Vehicle: symmetry, boundaries, and active contour detection	Radar ROI	60	3-GHz Core i7 CPU
2015	[150]	Vehicle: random decision forest classifier	Radar ROI	10	Core i7 processor, 8 GB of RAM
2016	[9]	Vehicle: active contour detection	Radar ROI	60	Intel i7 3-GHz CPU
2016	[54]	Vehicle/pedestrian/two wheels/traffic cones: four DPMs trained by ISVM and HOG	Radar ROI	29	—
2018	[56]	Pedestrian: HOG and SVM	Radar ROI	40	—
2019	[151]	Bus/bicycle/car/pedestrian/motorcycle: YOLOv2	Radar ROI	15	—
2019	[152]	Car: CNN network	End to end	—	Two Nvidia Titan V GPUs
2019	[153]	Bus/bicycle/car/pedestrian/motorcycle: Fast R-CNN	Radar ROI	—	—
2020	[62]	Car/pedestrian/motorcycle: Haar-like + AdaBoost, PCA + HOG + SVM	Radial-basis function NN	16.7	Intel i7 2.6 GHz
2020	[154]	Vehicle: feature pyramid+ Improved YOLOV3	End to end	—	Geforce GTX 1080 Ti
2020	[155]	Vehicle: spatial attention	End to end	—	8 GTX 1080 Ti GPUs
2020	[156]	Multiple object: RANet/BIRANet	Radar ROI	—	TITAN Pascal GPU
		Object Tracking By RV Fusion			
2018	[124]	Tracking: temporal trajectory analysis	Fusion before tracking	30–40	GIGABYTE Mini-PC
2019	[157]	Tracking: IMM	Fusion before tracking	—	—
2019	[158]	Tracking: LSTM	Fusion before tracking	—	Tesla K80 GPU
2020	[159]	Tracking: gate + EKF	Fusion after tracking	—	—
2020	[160]	Tracking: JPDAF underlying a KF	Fusion before and after tracking	—	—
2020	[161]	Tracking: UKF	Fusion after tracking	10	Nvidia Jetson Xavier
2020	[162]	Tracking: IMM	Fusion before tracking	—	—
2021	[163]	Tracking: Improved GM-PHD	Fusion before tracking	—	—
2021	[164]	Tracking: JPDA	Fusion before tracking	20	Intel dual core, 4 G of memory

DPM: deformable part model; ISVM: incremental support vector machine; SVM: support vector machine; JPDA: joint probabilistic data association; GM-PHD: Gaussian mixture PHD; RAM: random-access memory; IMM: interacting multiple model; JPDAF: joint probability data association filter; UKF: unscented Kalman filter.

Table 8. The advantages and disadvantages of the various RV fusion approaches.

Approach	Strategy	Advantages	Disadvantages
Object detection by RV fusion	Radar ROI	✓ Less runtime of detection ✓ Low computational cost	× Low fault tolerance × Undetected and false-detection errors
	Visual ROI	✓ Fewer false-detection errors	× Higher computational cost × Undetected errors
	End to end	✓ Real-time fusion ✓ High efficiency	× Higher computational cost
Object tracking by RV fusion	Fusion before tracking	✓ Better fault tolerance ✓ Real-time fusion	× Higher computational cost × Undetected and false-detection errors
	Fusion after tracking	✓ More robust and stable ✓ Fewer detected errors	× Higher computational cost × Fusion hysteresis

real-time requirements of AVs. The RRPN [153] generates object proposals by mapping radar detection data to the image coordinate system and generating predefined anchor boxes for each of the mapped radar detection points, which results in faster and more accurate detections. In [156], radar information was fused in the feature extractor network, and after that, radar points were used to generate guided anchors.

In addition, modern radar can generate multiple detections per object, and the received radar point

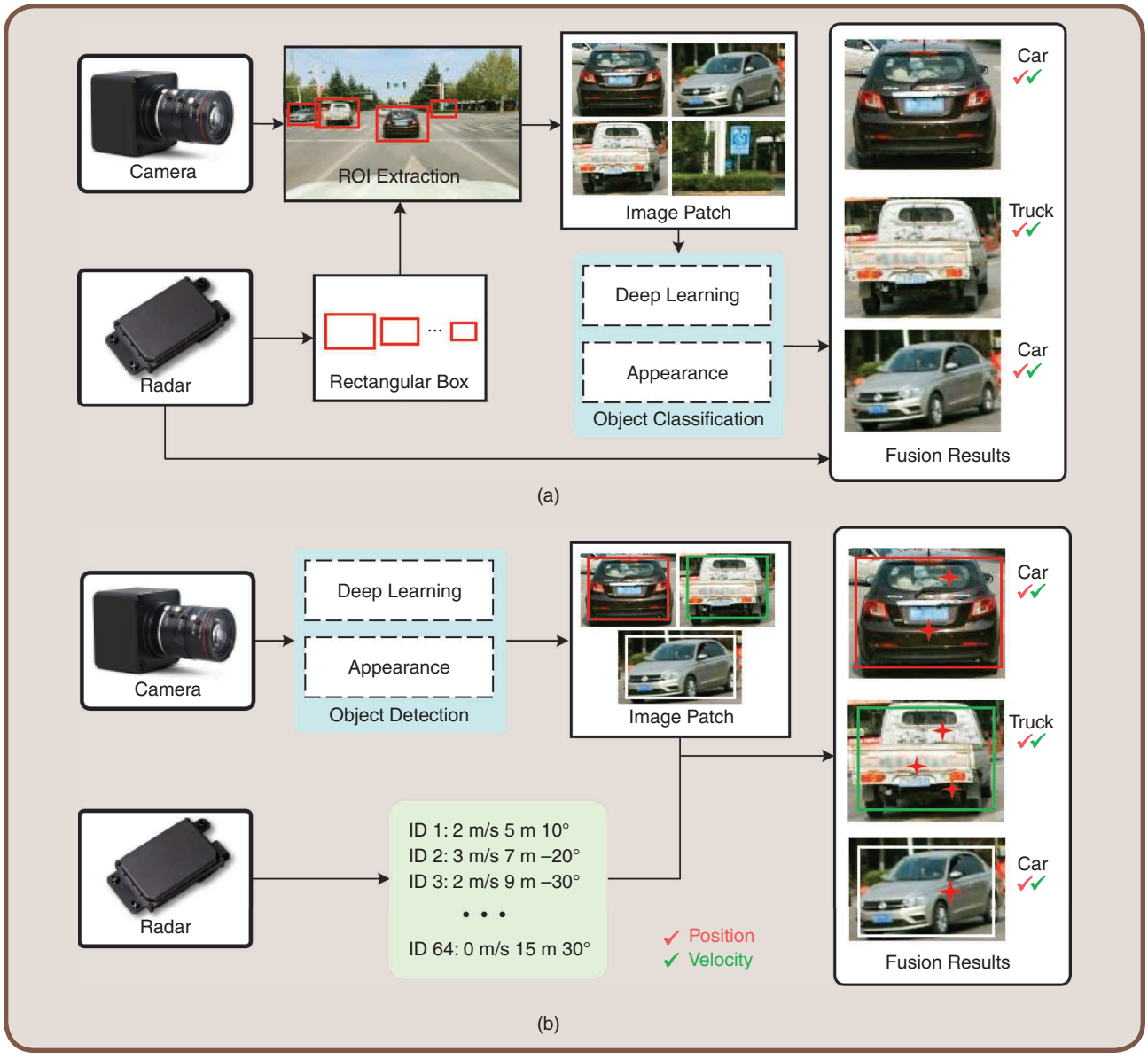

FIG 9 An ROI-based RV fusion scheme. (a) A radar ROI fusion strategy and (b) a visual BBox fusion strategy.

clouds are extremely sparse compared to the lidar point clouds or captured images. Thus, it is a challenging task to recognize different objects using only radar data. Many studies have performed either object classification or BBox estimation for objects. Danzer et al. [118] facilitated a classification and BBox estimation of objects using only radar.

Visual BBox

Visual BBox uses the BBox from visual detection to match the radar points inside it. In the past decades, visual BBox (ROI) were generated using symmetry, edge, contour, HOG, Haar-like, and other features. In Figure 9(b), red asterisks denote the radar points, and rectangular boxes indicate the BBox. BBox is used to match radar measurement data [44]. Wedel and Franke [175] generated a BBox in the image that was subdivided into differences and tracked each vertical slice individually. Recently, DL-based object detection approaches have been used to generate BBox. In [124], the BBox of an object obtained by the YOLO9000 was fine-tuned to the correct position using radar points; the YOLOv3 was used to detect object BBox, which included many vertical slices used to detect the histogram of probability mapped with radar position and velocity [176].

However, vision-based object detection can experience false detection and undetected defects. A cross-verification scheme was used in [37], where the BBox from the image was used to match radar points, and then the vision-based vehicle detection approach was used in the ROI generated from unmatched radar points. In [67], the authors used a circle that took the middle point of the BBox's bottom to match the radar points, and the nearest point to the BBox's center as the optimal point. The circle radius represented the maximal values of the length and width of the BBox.

End-to-End Fusion Strategies

In many related works, high detection accuracy was achieved by simply fusing multiple sensing modes with simple operations, such as addition and average mean, without considering the network's robustness [153]. Compared to other fusion procedures, the end-to-end fusion strategy is more robust; it can also perceive the objects in a straightforward manner, and the radar and visual features are inputted to the same framework [154], [177]. RVNet [178] used a DL-based sensor fusion framework, which fused radar and camera data in the same framework. A DL approach with generative adversarial networks [179] was presented in [179] to fuse the camera images with the radar images that were obtained from the radar measurement data by the fully unsupervised machine learning algorithm. A radar point cloud and an RGB image were used as the inputs of a CNN network, whose output was the BBox prediction, including the information on position and dimension [152].

Recently, attention mechanisms have become an important part of object detection. Similar to the selective visual attention mechanism of humans, the attention mechanism selects critical objects to obtain detailed information on that object rather than other useless information. In [155], a spatial attention fusion approach was proposed for feature fusion, and an attention weight matrix was generated to fuse vision features.

The framework presented in Figure 11 contains two independent feature-extraction branches: fusion and output. Both the radar and visual features are fed to the fusion branch. The four types of fusion techniques are represented by the yellow rectangle. Finally, an NN is employed in the output branch to output the tracks. In conclusion, to facilitate the realization of AVs, an RV fusion algorithm with high fault tolerance and precision, fast reasoning speed, and low calculation cost needs to be developed as soon as possible.

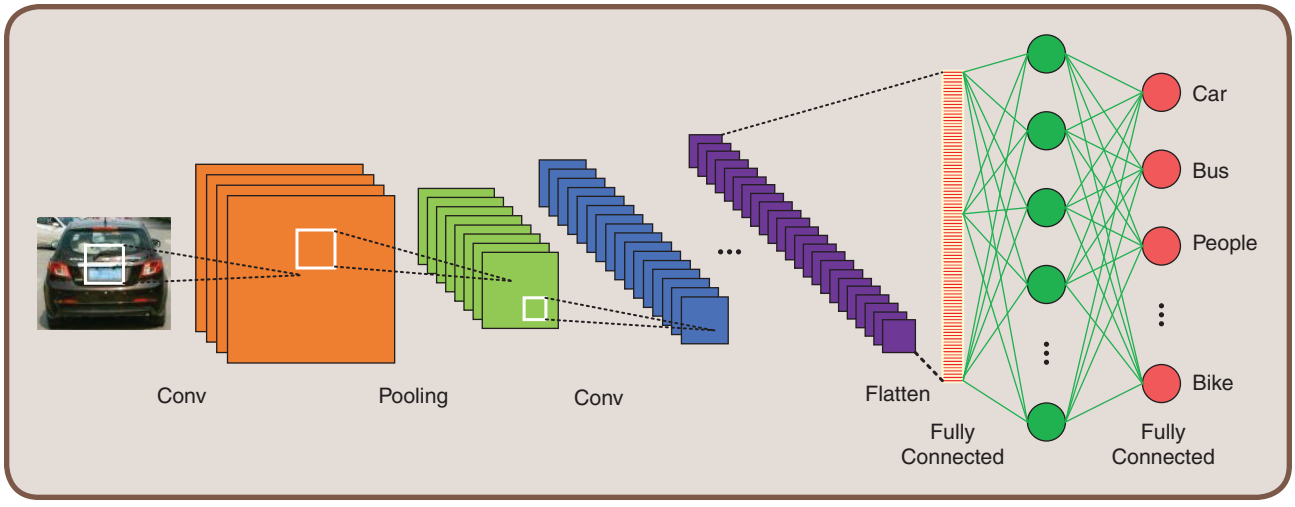

FIG 10 A CNN-based object classification. Conv: convolution.

Object Tracking By RV Fusion

Object tracking fusion strategies can be divided into two classes: fusion before and fusion after tracking. Compared to fusion after tracking, fusion before tracking is less time-consuming but tends to generate more false tracking.

Fusion Before Tracking

Fusion before tracking matches the center coordinates of the image objects with those of the radar objects [43], [130], [157]. This fusion scheme is shown in Figure 12(a), where the yellow circles and red asterisks are matched via data fusion performed by different approaches. Jiang et al. [151] fused the radar and visual objects using the weighted information fusion algorithm. There are some effective matching methods, such as intersection over union [180], Mahalanobis distance [132], global nearest neighbor [133], and means of a likelihood function [35]. LSTM was used in [158] for object tracking after the objects from radar and vision were associated and fused.

Fusion After Tracking

Fusion strategy after tracking, namely, track-to-track fusion (T2TF), matches the tracks of the radar and the camera. In this case, the process noise is not negligible. Due to common process noise from the underlying system, the measurement data of sensors are not conditionally independent [181]. To address this problem, the cross covariance between the local tracks was accounted for in [182], which could decide whether two tracks from different sensors represented the same object. Cross covariance for heterogeneous T2TF was used in [183]. Compared with centralized measurement fusion, T2TF could effectively reduce the communication requirements and was deemed suitable for practical implementation [184], [185]. Therefore, T2TF is an appropriate choice for RV fusion [186]. The visual tracks can be generated by both the two-step and one-shot approaches, and a detailed framework is displayed in Figure 8(a) and (b). In Figure 12(b), the blue block lists object detection strategies, and the pink block presents the object tracking techniques. A unified NN is proposed to perform detections and obtain the Re-ID features, as shown in Figure 12(c). The detailed schematic diagrams of the tracks are shown in Figure 12(d).

The DBSCAN cluster [48], global association optimization scheme [187], and radial-basis function NN [62] have

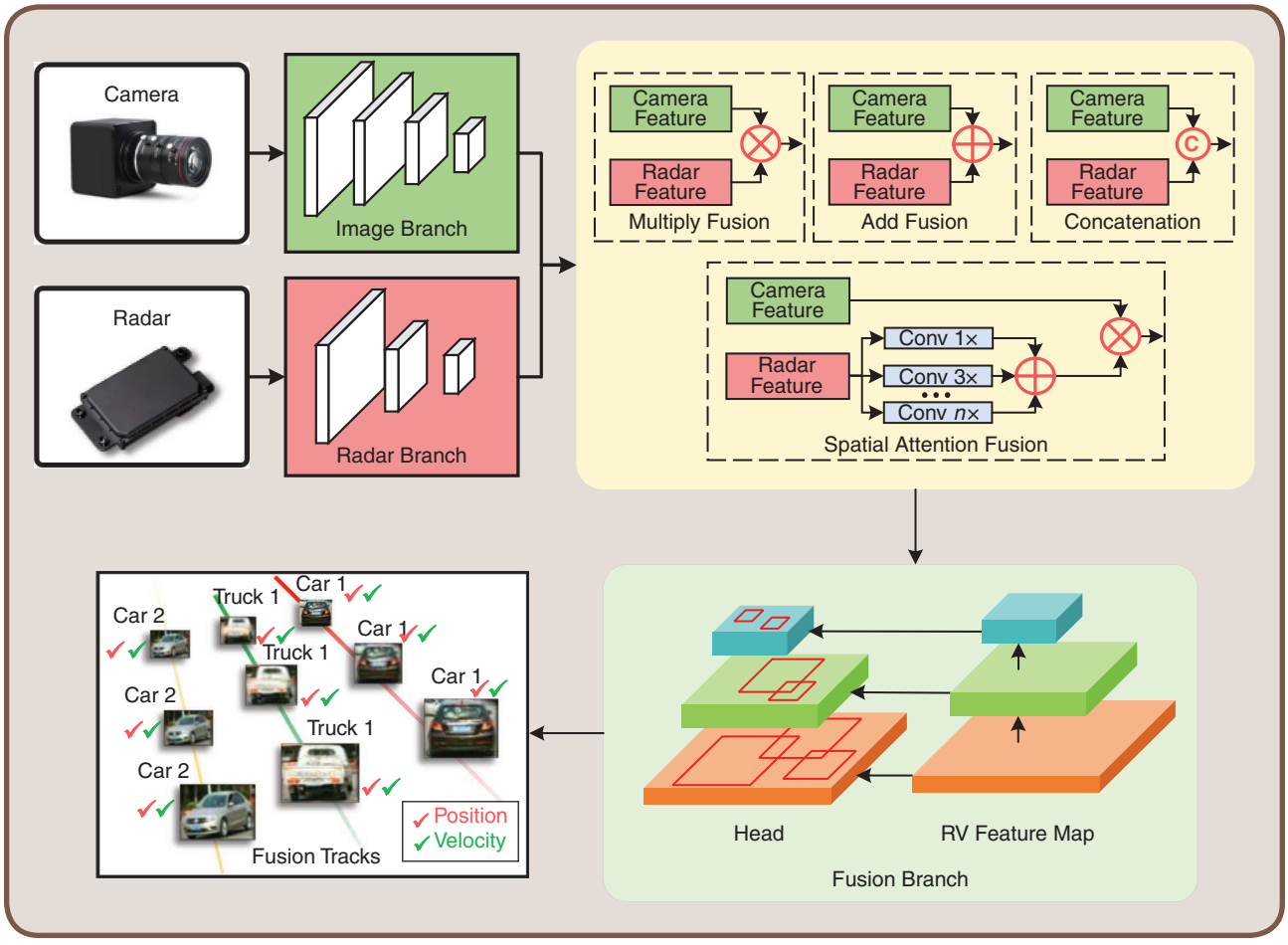

FIG 11 An end-to-end RV fusion scheme. The four dotted boxes represent four fusion algorithms, which select only one while fusing radar and camera.

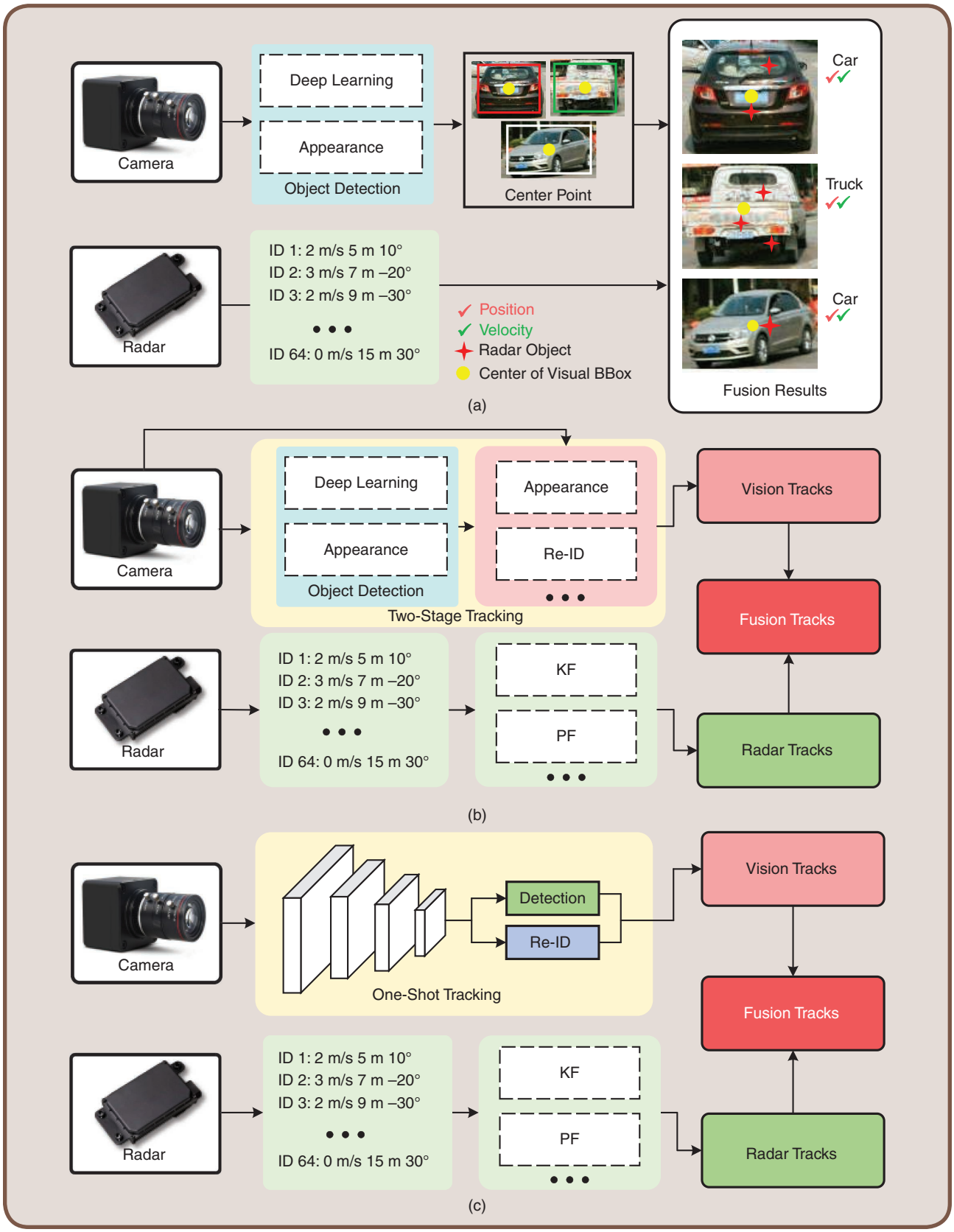

FIG 12 An object-based RV fusion scheme. (a) Fusion before tracking, (b) fusion after tracking by the two-step MOT, and (c) fusion after tracking by the one-shot MOT. *(Continued)*

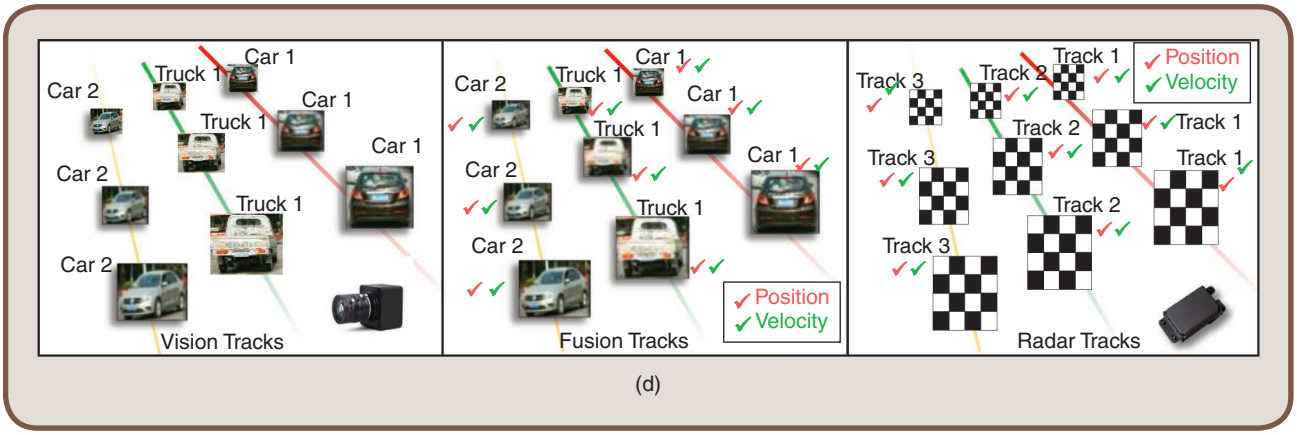

FIG 12 (*Continued*) (d) From left to right are vision tracks, fusion tracks, and radar tracks. The dotted boxes with algorithms are parallel. The RV fusion scheme can be achieved successfully using only one of these dotted boxes.

been proposed to combine both the radar and visual tracks. To improve the robustness of fusion strategy, two fusion approaches were used in [6] and [9]; the ROI obtained from the radar measurement data were used for active contour detection, the trajectories generated from the radar data and input images were combined, and the trajectory error was calculated based on the Euclidean distance. Due to the inconsistencies in sensor sampling frequency, an asynchronous fusing MOT approach that stores a series of images obtained by a camera into a buffer to wait for measurements was proposed in [188], and an interacting multiple model filter with probabilistic data association was used to fuse the radar and camera data.

Due to inaccurate detections, abrupt changes in object motion or appearance, and frequent occlusion by clutter or other objects, MOT still experiences many challenges. Recently, considerable attention has been paid to the fusion of radar and vision into a unified framework. The camera radar fusion network [189] fused the camera and radar sensor's information on vehicles. Bae [190] leveraged a unified framework based on object visual and amplitude model learning and data association approaches for MOT, which had a lower tracking complexity compared to the state-of-the-art DL-based method. A CNN-based architecture was proposed in [191] to detect and track objects of interest in the blind range of radar.

Although DL has been developing rapidly in the CV field, it is still in an early stage in the field of RV MOT because of interdisciplines. Clearly, uncertainty modeling is an important factor for AV safety. The perceptual system shows higher uncertainty under conditions of adverse weather or low-visibility driving environments. The uncertainties need to be propagated to a phase of decision making and motion planning so that AVs can behave accordingly. Dropout was used in [192] to increase the network robustness under fog conditions. Bayesian NNs (BNNs) and unsupervised generative models can help to increase robustness [193]. In addition, sensors' degradation and defects can result in serious accidents. To research the limitations of multisensor fusion systems, safety of the intended functionality (SOTIF) has been introduced to address these problems. However, many articles pay attention to only the high rates of accuracy and fast inference speeds without considering the networks' robustness and SOTIF. The DL-based RV approaches need to be further explored to estimate uncertainty and increase network robustness.

Discussion and Outlook

As mentioned in previous sections, tremendous results have been achieved in the fields of object detection and tracking and RV fusion. However, there are several challenges and problems that have neither drawn much attention nor been considered previously. These problems are discussed in detail in the following.

Future Trends for Detection and Tracking

A complex environment is the major challenge in environmental perception. AV safety can benefit from object detection and MOT. With the recent introduction of in-depth research into the field of detection and tracking, many promising future research directions have arisen.

Detection

In most cases, several features are employed to compensate for the fact that relevant features are unknown priori. It is impossible to use a single feature to suit all conceivable scenarios. In the future, combining multiple cues should be researched in detail to develop more reliable approaches and robust systems for multiobject detection. Further, several well-established DL-based CV schemes that are suitable for on-road object detection have been recently proposed. The state-of-the-art object detection techniques in the International Conference on Computer Vision

(ICCV), European Conference on Computer Vision (ECCV), and the Conference on Computer Vision and Pattern Recognition (CVPR) can make great contributions to AV safety by transfer learning.

Tracking

Most DL-based visual MOT algorithms focus mainly on the pedestrian tracking. However, different on-road objects (e.g., vehicles, motorcycles, and other objects) pose different challenges so possible advances in DL should be investigated. In addition, abundant features and accurate temporal information obtained by camera and radar can be fused into the end-to-end framework so that the errors will not be accumulated from detection to tracking. A large amount of training data, including various scenarios, are needed to train a one-shot architecture. Therefore, over-fitting, data set annotations, and one-shot network design are the primary limitations of the end-to-end tracking framework, which need to be addressed in future research.

Challenging Weather

The quality of the visual image captured by a camera mounted on a moving AV can be notably impaired and distorted under adverse weather conditions, such as rain, snow and fog. The performance of these systems has not been thoroughly researched under adverse weather conditions. Recently, many imaging techniques for image de-raining [194] and video defogging [195] have been studied. Another meaningful research of AV ability to detect objects in a rainy environment was conducted in [196]. However, perception in adverse environments needs to be further explored.

Future Trends for RV Fusion

Different sensors are used in AVs to predict and mitigate collisions. However, employing more effective sensors and fusion approaches in environment perception systems can further improve AV safety considerably. Future trends in RV fusion are as follows.

Radar

Due to good reliability and accuracy, higher-frequency radar systems, such as 76–81-GHz radar, should be used in AVs to achieve better detection performance. The frequency bands higher than 77 GHz are recommended by the European Union Commission to be used for adaptive cruise control and near-range parking [197].

Camera

An enhanced dynamic range enables object detection in the daytime, nighttime, and in backlight scenarios without blooming. The original image can be used in a dark environment because of its richer, dark details as compared to the RGB image. Also, improving the camera resolution provides significant benefits to detection accuracy.

Spatial Calibration

The spatial calibration of the camera and the radar is a significant precondition for on-road object detection based on data fusion. Generally, although the camera calibration process is complex, the calibration results are not accurate enough. Therefore, it is necessary to further research the simplicity, accuracy and instantaneity of spatial calibration of both radar and camera.

Fusion Schemes

Many approaches have been successfully used to fuse the information of radar and camera. The emerging, advanced end-to-end fusion framework based on DL can help to improve the robustness of RV fusion.

Future Trends for Real-Time Performance

A complex road environment requires DL technology with high a high rate of accuracy and strong universality; however, due to multiple parameters and high calculation cost, the real-time performance of DL technology can barely be achieved, even by using high-quality GPUs. Therefore, the inference speed and achievement of real-time performance on a common device should be further studied.

Integrated System

On-road object detection and tracking systems should be modular, extensible, and reconfigurable to enable the location and recognition of various objects. The state-of-the-art research on sensor fusion, object detection, and object tracking approaches can be integrated to overcome all of the problems caused by environmental complexity. Also, by integrating diverse algorithms, scalability, the economical use of resources, and distributed computing can be further improved in the future for achieving better real-time performance.

Model Pruning

To reduce the computational redundancy of a full model, model pruning that can maintain an acceptable accuracy range, such as EagleEye [198], has been researched in recent years. Some pruned models, including YOLOv3-tiny and YOLOv4-tiny, can reduce the hardware resource's budget and provide a higher running speed. However, how best to prune the model parts while maintaining a high rate of accuracy is a difficult question.

Hardware

To process sensor information in real time or close to real time, on-road object detection hardware systems have high computational requirements. Therefore, high-performance embedded supercomputing platforms, such as Nvidia PX2 and Nvidia Jetson TX2, have begun to be developed and employed in intelligent vehicles. In future research, the development of embedded platforms with lower energy resources but higher accuracy and speed will be the dominant trend.

Future Trends for Robustness Performance

Up until now, many scholars have focused mainly their attention on the inference speed and accuracy of object detection and tracking, rarely considering the comprehensiveness and robustness of the AV environment's perception system. Therefore, robustness performance must be further studied.

SOTIF

The situational awareness of the perceptual system and its algorithms can directly affect AV safety in situations when performing anticipated functions. Recently, a post-accident investigation of one of Tesla's vehicles showed that incidents happened not only because of sensor fault but also due to some other, nonsensor fault-related causes. SOTIF has been proposed to address these problems and focuses on the hazards resulting from insufficient, intended functionality or reasonably foreseeable misuse by persons. Consequently, to solve the uncertainties and complexities of the driving environment, SOTIF needs further investigation.

Data Set

The DL-based detection models are trained using a certain set of real data, but these data can be insufficient, noisy, or even contaminated by some irrelevant features. Although the fitting performance during training is good, that does not guarantee that the trained model will have good generalization ability. Thus, a high-quality training data set is crucial for AV development. To further speed up the popularization of AVs, an unsupervised data set can represent a good choice because it aims at modeling the data distribution in an unsupervised way as well as generating new samples with some variations.

Conclusion

This article summarized and analyzed on-road object detection and tracking approaches and focused on RV fusion strategies. The details of the object detection and tracking schemes employed in the RV fusion framework were summarized, and the DL-based, state-of-the-art, real-time object detectors and trackers were listed. The specific RV fusion strategies, which determine the real-time performance and accuracy of environment perception of an AV, were also classified. Furthermore, the contributions and limitations as well as further improvements in object detection and tracking, and RV fusion were discussed. Finally, possible future developments in terms of sensors, real-time performance, SOTIF, and a data set were introduced. This article provided a concise overview of the latest achievements and trends in on-road object detection and tracking based on the fusion of radar and vision, which can help and guide both experienced and new researchers in this rapidly developing AV field.

Acknowledgments

This work was supported, in part, by the National Natural Science Foundation of China under grants 52072051 and 51805028; the Graduate Research and Innovation Foundation of Chongqing of China under grant CYS20019; the State Key Laboratory of Mechanical System and Vibration under grant MSV202016; the Natural Science Foundation of Chongqing of China under grant cstc2020jcyj-msxmX0956; the Beijing Institute of Technology Research Fund Program for Young Scholars; and the Young Elite Scientists Sponsorship Program, funded by the China Society of Automotive Engineers.

About the Authors

Xiaolin Tang (tangxl0923@cqu.edu.cn) earned his Ph.D. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2015. He is currently an associate professor at the State Key Laboratory of Mechanical Transmissions and at the Department of Automotive Engineering, Chongqing University, Chongqing, 400044, China. His research focuses on hybrid electric vehicles, vehicle dynamics, noise and vibration, and transmission control. He is a Member of IEEE.

Zhiqiang Zhang (20193202026t@cqu.edu.cn) earned his B.S. degree in vehicle engineering from Northeast Forestry University, Harbin, China, in 2019. He is currently pursuing his M.S. degree in automotive engineering at Chongqing University, Chongqing, 400044, China. His research interests include multisensor information fusion, computer vision, and machine learning.

Yechen Qin (qinyechenbit@gmail.com) earned his B.Sc. and Ph.D. degrees in mechanical engineering from the Beijing Institute of Technology, in 2010 and 2001, respectively. He is currently an associate professor at the Beijing Institute of Technology, Beijing, 100081, China. From 2013 to 2014, he was a visiting Ph.D. student with Texas A&M University. He was also a postdoctoral research fellow and a visiting scholar with the Beijing Institute of Technology and University of Waterloo. His current research interest is autonomous vehicle dynamics control. His research interests include autonomous vehicle control, road estimation, and in-wheel motor vibration control. He is a Member of IEEE.

References

[1] Y. Qin, E. Hashemi, and K. Amir, "Integrated crash avoidance and mitigation algorithm for autonomous vehicles," *IEEE Trans. Ind. Inf.*, 2021. doi: 10.1109/TII.2021.3058948.

[2] World Health Organization, "Stockholm declaration," Road Safety Sweden, Feb. 2020. [Online]. Available: https://www.roadsafetysweden.com/about-the-conference/stockholm-declaration/

[3] Y. Qin, C. Wei, X. Tang, N. Zhang, M. Dong, and C. Hu, "A novel nonlinear road profile classification approach for controllable suspension system: Simulation and experimental validation," *Mech. Syst. Signal Process.*, vol. 125, pp. 79–98, 2019. doi: 10.1016/j.ymssp.2018.07.015.

[4] C. Hu et al., "RISE-based integrated motion control of autonomous ground vehicles with asymptotic prescribed performance," *IEEE Trans. Syst., Man, Cybern., Syst.*, 2019. doi: 10.1109/TSMC.2019.2950468.

[5] T. Liu, X. Tang, H. Wang, H. Yu, and X. Hu, "Adaptive hierarchical energy management design for a plug-in hybrid electric vehicle," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 11,513–11,522, 2019. doi: 10.1109/TVT.2019.2926733.

[6] X. Wang, L. Xu, H. Sun, J. Xin, and N. Zheng, "Bionic vision inspired on-road obstacle detection and tracking using radar and visual information," in *Proc. 17th Int. IEEE Conf. Intell. Transport. Syst. (ITSC)*, 2014, pp. 39–44.

[7] M. L. Fung, M. Z. Chen, and Y. H. Chen, "Sensor fusion: A review of methods and applications," in *Proc. 29th Chinese Control Decision Conf. (CCDC)*, 2017, pp. 3853–3860. doi: 10.1109/CCDC.2017.7979175.

[8] Z. Wang, Y. Wu, and Q. Niu, "Multi-sensor fusion in automated driving: A survey," *IEEE Access*, vol. 8, pp. 2847–2868, 2019. doi: 10.1109/ACCESS.2019.2962554.

[9] X. Wang, L. Xu, H. Sun, J. Xin, and N. Zheng, "On-road vehicle detection and tracking using MMW radar and monovision fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 2075–2084, 2016. doi: 10.1109/TITS.2016.2533542.

[10] P. Prasher, "LiDAR, radar, or camera? Demystifying the ADAS/AD technology mix," Apr. 2021. [Online]. Available: https://leddartech.com/lidar-radar-camera-demystifying-adas-ad-technology-mix/

[11] E. Brandt, "Lidar vs Radar: Pros and cons of different autonomous driving technologies," The Drive, Dec. 2017. [Online]. Available: https://www.thedrive.com/article/16916/lidar-vs-radar-pros-and-cons-of-different-autonomous-driving-technologies

[12] "Pros and cons of LIDAR," Lidar Radar, Apr. 2021. [Online]. Available: https://lidarradar.com/info/pros-and-cons-of-lidar

[13] S. Crowe, "Researchers back Tesla's non-LiDAR approach to self-driving cars," The Robot Report, Apr. 2019. [Online]. Available: https://www.therobotreport.com

[14] Mobileye. Accessed: Apr. 22, 2021. [Online]. Available: https://www.mobileye.com/

[15] MINIEYE. Accessed: Apr. 22, 2021. [Online]. Available: https://www.minieye.cc/

[16] "MM solutions." Accessed: Apr. 22, 2021. [Online]. Available: https://www.mm-sol.com

[17] K. Wiggers, "Baidu claims its Apollo Lite vision-based vehicle framework achieves level 4 autonomy," Venture Beat, June 2019. [Online]. Available: https://venturebeat.com/2019/06/19/baidu-claims-its-apollo-lite-vision-based-vehicle-framework-achieves-level-4-autonomy/

[18] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 694–711, 2006. doi: 10.1109/TPAMI.2006.104.

[19] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1773–1795, 2013. doi: 10.1109/TITS.2013.2266661.

[20] J. E. Espinosa, S. A. Velastín, and J. W. Branch, "Detection of motorcycles in urban traffic using video analysis: A review," *IEEE Trans. Intell. Transp. Syst.*, 2020.

[21] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018. doi: 10.1016/j.neucom.2018.01.092.

[22] K. A. Joshi and D. G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *Int. J. Soft Comput. Eng.*, vol. 2, no. 3, pp. 44–48, 2012.

[23] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020. doi: 10.1007/s11263-019-01247-4.

[24] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, 2020. doi: 10.1016/j.neucom.2019.11.023.

[25] K. Granstrom, M. Baum, and S. J. Reuter, "Extended object tracking: Introduction, overview and applications," 2016, arXiv:1604.00970.

[26] G. S. Walia and R. Kapoor, "Recent advances on multicue object tracking: A survey," *Artif. Intell. Rev.*, vol. 46, no. 1, pp. 1–39, 2016. doi: 10.1007/s10462-015-9454-6.

[27] K. V. Sakhare, T. Tewari, and V. Vyas, "Review of vehicle detection systems in advanced driver assistant systems," *Arch. Comput. Meth. Eng.*, vol. 27, no. 2, pp. 591–610, 2020. doi: 10.1007/s11831-019-09321-3.

[28] Z. Yang and L. S. Pun-Cheng, "Vehicle detection in intelligent transportation systems and its applications under varying environments: A review," *Image Vis. Comput.*, vol. 69, pp. 143–154, 2018. doi: 10.1016/j.imavis.2017.09.008.

[29] A. Gern, U. Franke, and P. Levi, "Robust vehicle tracking fusing radar and vision," in *Proc. Conf. Documentation Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI 2001)*, 2001, pp. 323–328.

[30] G. Alessandretti, A. Broggi, and P. Cerri, "Vehicle and guard rail detection using radar and vision data fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 1, pp. 95–105, 2007. doi: 10.1109/TITS.2006.888597.

[31] M. Bertozzi et al., "Obstacle detection and classification fusing radar and vision," in *Proc. IEEE Intell. Veh. Symp.*, 2008, pp. 608–613.

[32] S. S. Teoh and T. Bräunl, "Symmetry-based monocular vehicle detection system," *Mach. Vis. Appl.*, vol. 23, no. 5, pp. 831–842, 2012. doi: 10.1007/s00138-011-0355-7.

[33] M. Chunmei, "Obstacles detection based on millimetre-wave radar and image fusion techniques," in *Proc. IET Int. Conf. Intell. Connect. Veh. (ICV 2016)*, 2016, vol. 2016, pp. 1–6.

[34] M. Nishigaki, S. Rebhan, and N. Einecke, "Vision-based lateral position improvement of radar detections," in *Proc. 15th Int. IEEE Conf. Intell. Transp. Syst.*, 2012, pp. 90–97.

[35] J. Burlet and M. Dalla Fontana, "Robust and efficient multi-object detection and tracking for vehicle perception systems using radar and camera sensor fusion," in *Proc. IET and ITS Conf. Road Transp. Inf. Control (RTIC 2012)*, 2012, pp. 1–6. doi: 10.1049/cp.2012.1553.

[36] T. Wang, N. Zheng, J. Xin, and Z. Ma, "Integrating millimeter wave radar with a monocular vision sensor for on-road obstacle detection applications," *Sensors*, vol. 11, no. 9, pp. 8992–9008, 2011. doi: 10.3390/s110908992.

[37] X. Liu, Z. Sun, and H. He, "On-road vehicle detection fusing radar and vision," in *Proc. IEEE Int. Conf. Veh. Electron. Safety*, 2011, pp. 150–154.

[38] B. Aytekin and E. Altuğ, "Increasing driving safety with a multiple vehicle detection and tracking system using ongoing vehicle shadow information," in *Proc. IEEE Int. Conf. Syst., Man Cybernetics*, 2010, pp. 3650–3656.

[39] J. Woo, J. Lee, and N. Kim, "Obstacle avoidance and target search of an autonomous surface vehicle for 2016 maritime robotX challenge," in *Proc. IEEE Underwater Technol. (UT)*, 2017, pp. 1–5. doi: 10.1109/UT.2017.7890308.

[40] N. Srinivasa, Y. Chen, and C. Daniell, "A fusion system for real-time forward collision warning in automobiles," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2003, vol. 1, pp. 457–462.

[41] A. Sole, O. Mano, G. P. Stein, H. Kumon, Y. Tamatsu, and A. Shashua, "Solid or not solid: Vision for radar target validation," in *Proc. IEEE Intell. Veh. Symp.*, 2004, pp. 819–824.

[42] N. Srinivasa, "Vision-based vehicle detection and tracking method for forward collision warning in automobiles," in *Proc. Intell. Veh. Symp.*, 2002, vol. 2, pp. 626–631.

[43] E. Richter, R. Schubert, and G. Wanielik, "Radar and vision based data fusion-advanced filtering techniques for a multi object vehicle tracking system," in *Proc. IEEE Intell. Veh. Symp.*, 2008, pp. 120–125.

[44] W. Huang, Z. Zhang, W. Li, and J. Tian, "Moving object tracking based on millimeter-wave radar and vision sensor," *J. Appl. Sci. Eng.*, vol. 21, no. 4, pp. 609–614, 2018.

[45] B. Steux, C. Laurgeau, L. Salesse, and D. Wautier, "Fade: A vehicle detection and tracking system featuring monocular color vision and radar data fusion," in *Proc. Intell. Veh. Symp.*, 2002, vol. 2, pp. 632–639.

[46] F. Garcia, P. Cerri, A. Broggi, A. de la Escalera, and J. M. Armingol, "Data fusion for overtaking vehicle detection based on radar and optical flow," in *Proc. IEEE Intell. Veh. Symp.*, 2012, pp. 494–499.

[47] H.-t. Kim and B. Song, "Vehicle recognition based on radar and vision sensor fusion for automatic emergency braking," in *Proc. 13th Int. Conf. Control, Automation Syst. (ICCAS 2013)*, 2013, pp. 1342–1346.

[48] F. J. Botha, C. E. van Daalen, and J. Treurnicht, "Data fusion of radar and stereo vision for detection and tracking of moving objects," in *Proc. Pattern Recogn. Assoc. South Africa Robot. Mechatron. Int. Conf. (PRASA-RobMech)*, 2016, pp. 1–7.

[49] M. Bertozzi, A. Broggi, and S. Castelluccio, "A real-time oriented system for vehicle detection," *J. Syst. Arch.*, vol. 43, nos. 1–5, pp. 317–325, 1997. doi: 10.1016/S1383-7621(96)00106-3.

[50] Y. Cho, J. Park, M. Kang, and J. Kim, "Autonomous detection and tracking of a surface ship using onboard monocular vision," in *Proc. 12th Int. Conf. Ubiquitous Robots Ambient Intell. (URAI)*, 2015, pp. 26–31. doi: 10.1109/URAI.2015.7358921.

[51] D. Y. Kim and M. Jeon, "Data fusion of radar and image measurements for multi-object tracking via Kalman filtering," *Inf. Sci.*, vol. 278, pp. 641–652, 2014. doi: 10.1016/j.ins.2014.03.080.

[52] C. Otto, W. Gerber, F. P. León, and J. Wirnitzer, "A joint integrated probabilistic data association filter for pedestrian tracking across blind regions using monocular camera and radar," in *Proc. IEEE Intell. Veh. Symp.*, 2012, pp. 636–641.

[53] R. O. Chavez-Garcia and O. Aycard, "Multiple sensor fusion and classification for moving object detection and tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 525–534, 2015. doi: 10.1109/TITS.2015.2479925.

[54] S. Han, X. Wang, L. Xu, H. Sun, and N. Zheng, "Frontal object perception for Intelligent Vehicles based on radar and camera fusion," in *Proc. 35th Chinese Control Conf. (CCC)*, 2016, pp. 4003–4008. doi: 10.1109/ChiCC.2016.7553978.

[55] Y. Oishi and I. Matsunami, "Radar and camera data association algorithm for sensor fusion," *IEICE Trans. Fund. Electron., Commun. Comput. Sci.*, vol. E100.A, no. 2, pp. 510–514, 2017. doi: 10.1587/transfun.E100.A.510.

[56] X.-p. Guo, J.-s. Du, J. Gao, and W. Wang, "Pedestrian detection based on fusion of millimeter wave radar and vision," in *Proc. Int. Conf. Artif. Intell. Pattern Recogn.*, 2018, pp. 38–42. doi: 10.1145/3268866.3268868.

[57] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proc. 6th Int. Conf. Comput. Vision (IEEE Cat. No. 98CH36271)*, 1998, pp. 555–562.

[58] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proc. Int. Conf. Image Process.*, 2002, vol. 1, pp. I–I.

[59] W.-C. Chang and C.-W. Cho, "Online boosting for vehicle detection," *IEEE Trans. Syst., Man, Cybernetics B, Cybernetics*, vol. 40, no. 3, pp. 892–902, 2009.

[60] U. Kadow, G. Schneider, and A. Vukotich, "Radar-vision based vehicle recognition with evolutionary optimized and boosted features," in *Proc. IEEE Intell. Veh. Symp.*, 2007, pp. 749–754. doi: 10.1109/IVS.2007.4290206.

[61] A. Haselhoff, A. Kummert, and G. Schneider, "Radar-vision fusion for vehicle detection by means of improved haar-like feature and adaboost approach," in *Proc. 15th European Signal Process. Conf.*, 2007, pp. 2070–2074.

[62] Y.-W. Hsu, Y.-H. Lai, K.-Q. Zhong, T.-K. Yin, and J.-W. Perng, "Developing an on-road object detection system using monovision and radar fusion," *Energies*, vol. 13, no. 1, p. 116, 2020. doi: 10.3390/en13010116.

[63] Z. Ji and D. Prokhorov, "Radar-vision fusion for object classification," in *Proc. 11th Int. Conf. Inf. Fusion*, 2008, pp. 1–7.

[64] J. Lorenzo, I. Parra, F. Wirth, C. Stiller, D. F. Llorca, and M. A. Sotelo, "RNN-based pedestrian crossing prediction using activity and pose-related features," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2020, pp. 1801–1806. doi: 10.1109/IV47402.2020.9304652.

[65] H. Königshof and C. Stiller, "Learning-based shape estimation with grid map patches for realtime 3D object detection for automated driving," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, 2020, pp. 1–6. doi: 10.1109/ITSC45102.2020.9294743.

[66] Z. Ji, M. Luciw, J. Weng, and S. Zeng, "Incremental online object learning in a vehicular radar-vision fusion framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 402–411, 2010. doi: 10.1109/TITS.2010.2094188.

[67] Q. Feng, S. Qi, J. Li, and B. Dai, "Radar-vision fusion for correcting the position of target vehicles," in *Proc. 10th Int. Conf. Intell. Human–Machine Syst. Cybernetics (IHMSC)*, 2018, vol. 2, pp. 352–355. doi: 10.1109/IHMSC.2018.10186.

[68] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[69] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, 2015. doi: 10.1109/TPAMI.2015.2437384.

[70] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recogn.*, 2020, pp. 10,781–10,790.

[71] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1440–1448.

[72] 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.*, 2015, pp. 91–99.

[73] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[74] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2961–2969.

[75] M. Dimitrievski, L. Jacobs, P. Veelaert, and W. Philips, "People tracking by cooperative fusion of RADAR and camera sensors," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, 2019, pp. 509–514.

[76] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recogn.*, 2020, pp. 13,906–13,915.

[77] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "Cornernet-lite: Efficient keypoint-based object detection," 2019, arXiv:1904.08900.

[78] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, arXiv:1904.07850.

[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2016, pp. 770–778.

[80] S. Li, L. Yang, J. Huang, X.-S. Hua, and L. Zhang, "Dynamic anchor feature selection for single-shot object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 6609–6618.

[81] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv:1409.1556.

[82] C.-Y. Wang, H.-Y. Mark Liao, P.-Y. Chen, and J.-W. Hsieh, "Enriching variety of layer-wise learning information by gradient combination," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, 2019, pp. 2477–2484.

[83] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of cnn," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recogn. Workshops*, 2020, pp. 390–391.

[84] J. Nie, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Enriched feature guided refinement network for object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 9537–9546.

[85] J. Cao, Y. Pang, J. Han, and X. Li, "Hierarchical shot detector," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 9705–9714.

[86] T. Wang, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Learning rich features at high-speed for single-shot object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 1971–1980.

[87] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2018, pp. 8759–8768.

[88] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," in *Advances in Neural Information Processing Systems*, 2018, pp. 1963–1972.

[89] S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, and S.-J. Ko, "Parallel feature pyramid network for object detection," in *Proc. European Conf. Comput. Vision (ECCV)*, 2018, pp. 234–250.

[90] X. Long et al., "PP-YOLO: An effective and efficient implementation of object detector," 2020, arXiv:2007.12099.

[91] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2018, pp. 4203–4212.

[92] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. European Conf. Comput. Vision (ECCV)*, 2018, pp. 385–400.

[93] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. European Conf. Comput. Vision*, 2016, pp. 21–37.

[94] P. Chao, C.-Y. Kao, Y.-S. Ruan, C.-H. Huang, and Y.-L. Lin, "Hardnet: A low memory traffic network," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 3552–3561.

[95] Z. Liu, T. Zheng, G. Xu, Z. Yang, H. Liu, and D. Cai, "Training-time-friendly network for real-time object detection," in *Proc. AAAI Conf. Artificial Intell.*, 2020. vol. 34, no. 7, pp. 11,685–11,692. doi: 10.1609/aaai.v34i07.6838.

[96] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, arXiv:1804.02767.

[97] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, arXiv:2004.10934.

[98] P. Mahto, P. Garg, P. Seth, and J. Panda, "Refining Yolov4 for vehicle detection," *Int. J. Adv. Res. Eng. Technol. (IJARET)*, vol. 11, no. 5, pp. 409–419, June 16, 2020.

[99] Z. Qin et al., "ThunderNet: Towards real-time generic object detection on mobile devices," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 6718–6727.

[100] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2016, pp. 779–788.

[101] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2017, pp. 7263–7271.

[102] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proc. European Conf. Comput. Vision (ECCV)*, 2019, pp. 6569–6578.

[103] J. Huang et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2017, pp. 7310–7311.

[104] M. E. Russell, A. Crain, A. Curran, R. A. Campbell, C. A. Drubin, and W. F. Miccioli, "Millimeter-wave radar sensor for automotive intelligent

cruise control (ICC)," *IEEE Trans. Microw. Theory Techn.*, vol. 45, no. 12, pp. 2444–2453, 1997. doi: 10.1109/22.643858.
[105] "The KITTI Vision Benchmark Suite," cvlibs.net, July 2019. [Online]. Available: http://www.cvlibs.net/datasets/kitti/
[106] D. Langer and T. Jochem, "Fusing radar and vision for detecting, classifying and avoiding roadway obstacles," in *Proc. IEEE Conf. Intell. Veh.*, 1996, pp. 333–338.
[107] U. Chipengo and M. Commens, "A 77 GHz simulation study of roadway infrastructure radar signatures for smart roads," in *Proc. 16th IEEE European Radar Conf. (EuRAD)*, 2019, pp. 137–140.
[108] C. Blanc, R. Aufrere, L. Malaterre, J. Gallice, and J. Alizon, "Obstacle detection and tracking by millimeter wave radar," *IFAC Proc. Vol.*, vol. 37, no. 8, pp. 322–327, 2004. doi: 10.1016/S1474-6670(17)31996-1.
[109] MOTChallenge: The Multiple Object Tracking Benchmark. Accessed: July 12, 2020. [Online]. Available: https://motchallenge.net/
[110] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," 2015, arXiv:1504.01942.
[111] Y. Zhan, C. Wang, X. Wang, W. Zeng, and W. Liu, "A simple baseline for multi-object tracking," 2020, arXiv:2004.01888.
[112] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2016, pp. 3464–3468. doi: 10.1109/ICIP.2016.7533003.
[113] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," 2016, arXiv:1603.00831.
[114] R. Ravindran, M. J. Santora, and M. M. Jamali, "Multi-object detection and tracking, based on DNN, for autonomous vehicles: A review," *IEEE Sensors J.*, vol. 21, no. 5, pp. 5668–5677, 2021. doi: 10.1109/JSEN.2020.3041615.
[115] A. Palffy, J. Dong, J. F. Kooij, and D. Gavrila, "CNN based road user detection using the 3D radar cube," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1263–1270, 2020. doi: 10.1109/LRA.2020.2967272.
[116] S. Kim, S. Lee, S. Doo, and B. Shim, "Moving target classification in automotive radar systems using convolutional recurrent neural networks," in *Proc. 26th European Signal Process. Conf. (EUSIPCO)*, 2018, pp. 1482–1486. doi: 10.23919/EUSIPCO.2018.8553185.
[117] O. Schumann, C. Wöhler, M. Hahn, and J. Dickmann, "Comparison of random forest and long short-term memory network performances in classification tasks using radar," in *Proc. Sensor Data Fusion: Trends, Solutions, Appl. (SDF)*, 2017, pp. 1–6. doi: 10.1109/SDF.2017.8126350.
[118] A. Danzer, T. Griebel, M. Bach, and K. Dietmayer, "2d car detection in radar data with pointnets," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, 2019, pp. 61–66.
[119] J. F. Tilly et al., "Detection and tracking on automotive radar data with deep learning," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, 2020, pp. 1–7. doi: 10.23919/FUSION45008.2020.9190261.
[120] K. Fatseas and M. J. Bekooij, "Neural network based multiple object tracking for automotive FMCW radar," in *Proc. Int. Radar Conf. (RADAR)*, 2019, pp. 1–5. doi: 10.1109/RADAR41533.2019.171248.
[121] Á. F. García-Fernández, Y. Xia, K. Granström, L. Svensson, and J. L. Williams, "Gaussian implementation of the multi-Bernoulli mixture filter," in *Proc. 22th Int. Conf. Inf. Fusion (FUSION)*, 2019, pp. 1–8.
[122] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013. doi: 10.1177/0278364913491297.
[123] V. Milanés et al., "Intelligent automatic overtaking system using vision for vehicle detection," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3362–3373, 2012. doi: 10.1016/j.eswa.2011.09.024.
[124] J.-G. Wang, S. J. Chen, L.-B. Zhou, K.-W. Wan, and W.-Y. Yau, "Vehicle detection and width estimation in rain by fusing radar and vision," in *Proc. 15th Int. Conf. Control, Automation, Robot. Vision (ICARCV)*, 2018, pp. 1063–1068. doi: 10.1109/ICARCV.2018.8581246.
[125] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T. K. Kim, "Multiple object tracking: A literature review," *Artif. Intell.*, vol. 293, p. 103448, 2020. doi: 10.1016/j.artint.2020.103448.
[126] K. Suzuki, C. Yamano, and N. Ikoma, "Multiple target tracking in automotive FCM radar by multi-Bernoulli filter with elimination of other targets," in *Proc. 21st Int. Conf. Inf. Fusion (FUSION)*, 2018, pp. 527–534.
[127] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960. [Online]. Available: https://doi.org/10.1115/1.3662552
[128] T. Wang, R. Aggarwal, and A. Somani, "Human tracking using Delphi ESR-vision fusion in complex environments," in *Proc. Int. Conf. Image Process., Comput. Vision, Pattern Recogn. (IPCV)*, 2015, p. 198.
[129] S. W. Kim, M. Byeon, K. Kim, and J. Y. Choi, "MAP-based online data association for multiple people tracking in crowded scenes," in *Proc. 22nd Int. Conf. Pattern Recogn.*, 2014, pp. 1212–1217.
[130] J. Ren, Y. Wang, Y. Han, and R. Zhang, "Information fusion of digital camera and radar," in *Proc. IEEE MTT-S Int. Microw. Biomed. Conf. (IMBioC)*, 2019, vol. 1, pp. 1–4. doi: 10.1109/IMBIOC.2019.8777799.
[131] B. D. Anderson and J. B. Moore, *Optimal Filtering*. North Chelmsford, MA: Courier Corp., May 23, 2012.
[132] S. Wu, S. Decker, P. Chang, T. Camus, and J. Eledath, "Collision sensing by stereo vision and radar sensor fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 606–614, 2009. doi: 10.1109/TITS.2009.2032769.
[133] M. Feng, Y. Chen, T. Zheng, M. Cen, and H. Xiao, "Research on information fusion method of millimetre wave radar and monocular camera for intelligent vehicle," in *J. Phys., Conf. Ser.*, vol. 1314, no. 1, p. 012059, 2019. doi: 10.1088/1742-6596/1314/1/012059.
[134] N. J. Gordon, D. J. Salmond, and A. F. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," in *IEE Proc. F (Radar Signal Processing*, vol. 140, no. 2, pp. 107–113, 1993. doi: 10.1049/ip-f-2.1993.0015.
[135] S. Scheidegger, J. Benjaminsson, E. Rosenberg, A. Krishnan, and K. Granström, "Mono-camera 3d multi-object tracking using deep learning detections and PMBM filtering," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2018, pp. 433–440. doi: 10.1109/IVS.2018.8500454.
[136] K. Granström, M. Fatemi, and L. Svensson, "Poisson multi-Bernoulli mixture conjugate prior for multiple extended target filtering," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 1, pp. 208–225, 2019. doi: 10.1109/TAES.2019.2920220.
[137] Á. F. García-Fernández, L. Svensson, J. L. Williams, Y. Xia, and K. J. I. T. o S. P. Granström, "Trajectory Poisson multi-Bernoulli filters," *IEEE Trans. Signal Process.*, vol. 68, pp. 4933–4945, 2020. doi: 10.1109/TSP.2020.3017046.
[138] Z. Su, H. Ji, and Y. Zhang, "A Poisson multi-Bernoulli filter with target spawning," in *Proc. 22th Int. Conf. Inf. Fusion (FUSION)*, 2019, pp. 1–6.
[139] M. Fröhle, K. Granström, and H. J. I. A. Wymeersch, "Decentralized Poisson multi-Bernoulli filtering for vehicle tracking," *IEEE Access*, vol. 8, pp. 126,414–126,427, 2020. doi: 10.1109/ACCESS.2020.3008007.
[140] A. Scheel, S. Reuter, and K. Dietmayer, "Vehicle tracking using extended object methods: An approach for fusing radar and laser," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, 2017, pp. 231–238. doi: 10.1109/ICRA.2017.7989029.
[141] H. Shen-Tu, Y. Rong, Y. Guo, J.-A. Luo, and Y. Shi, "Gaussian mixtures match and fusion algorithms for multi-sensor multi-target tracking," in *Proc. 22th Int. Conf. Inf. Fusion (FUSION)*, 2019, pp. 1–8.
[142] Z. Fu, S. M. Naqvi, and J. A. Chambers, "Collaborative detector fusion of data-driven PHD filter for online multiple human tracking," in *Proc. 21st Int. Conf. Inf. Fusion (FUSION)*, 2018, pp. 1976–1981.
[143] A. Scheel and K. Dietmayer, " Tracking multiple vehicles using a variational radar model," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3721–3736, 2018. doi: 10.1109/TITS.2018.2879041.
[144] Y. Tan, F. Han, and F. Ibrahim, "A radar guided vision system for vehicle validation and vehicle motion characterization," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2007, pp. 1059–1066. doi: 10.1109/ITSC.2007.4357754.
[145] 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*, 2019, pp. 482–490.
[146] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 3029–3037.
[147] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang, "Online multi-object tracking with dual matching attention networks," in *Proc. European Conf. Comput. Vision (ECCV)*, 2018, pp. 366–382.
[148] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 3645–3649. doi: 10.1109/ICIP.2017.8296962.
[149] S. Kapania, D. Saini, S. Goyal, N. Thakur, R. Jain, and P. Nagrath, "Multi object tracking with UAVs using deep SORT and YOLOv3 RetinaNet detection framework," in *Proc. 1st ACM Workshop Autonom. Intell. Mobile Syst.*, 2020, pp. 1–6. doi: 10.1145/3377283.3377284.
[150] F. A. Alencar, L. A. Rosero, C. Massera Filho, F. S. Osório, and D. F. Wolf, "Fast metric tracking by detection system: radar blob and camera fusion," in *2015 12th Latin American Robot. Symp. 2015 3rd Brazilian Symp. Robot. (LARS-SBR)*, 2015, pp. 120–125.
[151] Q. Jiang, L. Zhang, and D. Meng, "Target detection algorithm based on MMW radar and camera fusion," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, 2019, pp. 1–6.
[152] M. Meyer and G. Kuschk, "Deep learning based 3d object detection for automotive radar and camera," in *Proc. 16th European Radar Conf. (EuRAD)*, 2019, pp. 133–136.
[153] R. Nabati and H. Qi, "RRPN: Radar region proposal network for object detection in autonomous vehicles," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2019, pp. 3093–3097. doi: 10.1109/ICIP.2019.8803392.

[154] L.-q. Li and Y.-l. Xie, "A feature pyramid fusion detection algorithm based on radar and camera sensor," in *Proc. 15th IEEE Int. Conf. Signal Process. (ICSP)*, 2020, vol. 1, pp. 366–370.
[155] S. Chang et al., "Spatial attention fusion for obstacle detection using mmWave radar and vision sensor," *Sensors*, vol. 20, no. 4, p. 956, 2020. doi: 10.3390/s20040956.
[156] R. Yadav, A. Vierling, and K. Berns, "Radar+ RGB fusion for robust object detection in autonomous vehicle," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2020, pp. 1986–1990.
[157] N. S. Zewge, Y. Kim, J. Kim, and J.-H. Kim, "Millimeter-wave radar and RGB-D camera sensor fusion for real-time people detection and tracking," in *Proc. 7th Int. Conf. Robot Intell. Technol. Appl. (RiTA)*, 2019, pp. 93–98. doi: 10.1109/RITAPP.2019.8932892.
[158] A. Sengupta, F. Jin, and S. Cao, "A DNN-LSTM based target tracking approach using mmWave radar and camera sensor fusion," in *Proc. IEEE National Aerospace Electron. Conf. (NAECON)*, 2019, pp. 688–693.
[159] J. Ma, Z. Tian, Y. Li, and M. Cen, "Vehicle tracking method in polar coordinate system based on radar and monocular camera," in *Proc. Chinese Control Decision Conf. (CCDC)*, 2020, pp. 93–98. doi: 10.1109/CCDC49329.2020.9164534.
[160] K. Aziz, E. De Greef, M. Rykunov, A. Bourdoux, and H. Sahli, "Radar-camera fusion for road target classification," in *Proc. IEEE Radar Conf. (RadarConf20)*, 2020, pp. 1–6. doi: 10.1109/RadarConf2043947.2020.9266510.
[161] J. X. Lu, J. C. Lin, M. Vinay, P.-Y. Chen, and J.-I. Guo, "Fusion technology of radar and RGB camera sensors for object detection and tracking and its embedded system implementation," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2020, pp. 1234–1242.
[162] K. R. Kurapati, M. Suma, and A. Chavan, "Multiple object tracking using radar and vision sensor fusion for autonomous vehicle," in *Proc. IEEE Int. Conf. Innova. Technol. (INOCON)*, 2020, pp. 1–6. doi: 10.1109/INOCON50539.2020.9298297.
[163] J. Bai, S. Li, L. Huang, and H. Chen, "Robust detection and tracking method for moving object based on radar and camera data fusion," *IEEE Sensors J.*, vol. 21, no. 9, 2021. doi: 10.1109/JSEN.2021.3049449.
[164] Z. Liu et al., "Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions," *IEEE Trans. Intell. Transp. Syst.*, early access, Feb. 24, 2021.
[165] Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi-object tracking," 2019, arXiv:1909.12605.
[166] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "MOTS: Multi-object tracking and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recogn.*, 2019, pp. 7942–7951.
[167] Y. Xu, A. Osep, Y. Ban, R. Horaud, L. Leal-Taixé, and X. Alameda-Pineda, "How to train your deep multi-object tracker," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recogn.*, 2020, pp. 6787–6796.
[168] J. Ebert, T. Gumpp, S. Münzner, A. Matskevych, A. P. Condurache, and C. Gläser, "Deep radar sensor models for accurate and robust object tracking," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, 2020, pp. 1–6. doi: 10.1109/ITSC45102.2020.9294735.
[169] T. Akita and S. Mita, "Object tracking and classification using millimeter-wave radar based on LSTM," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, 2019, pp. 1110–1115.
[170] H. Liu, H. Zhang, and C. Mertz, "DeepDA: LSTM-based deep data association network for multi-targets tracking in clutter," in *Proc. 22th Int. Conf. Inf. Fusion (FUSION)*, 2019, pp. 1–8.
[171] S. Richter, S. Wirges, H. Königshof, and C. J. t-T. M. Stiller, "Fusion of range measurements and semantic estimates in an evidential framework (Fusion von Distanzmessungen und semantischen Größen im Rahmen der Evidenztheorie)," *tm-Technisches Messen*, vol. 86, no. s1, pp. 102–106, 2019. doi: 10.1515/teme-2019-0052.
[172] S. Richter, J. Beck, S. Wirges, and C. Stiller, "Semantic evidential grid mapping based on stereo vision," in *Proc. IEEE Int. Conf. Multisensor Fusion Integration Intell. Syst. (MFI)*, 2020, pp. 179–184. doi: 10.1109/MFI49285.2020.9235217.
[173] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000. doi: 10.1109/34.888718.
[174] L. Bombini, P. Cerri, P. Medici, and G. Alessandretti, "Radar-vision fusion for vehicle detection," in *Proc. Int. Workshop on Intell. Transp.*, 2006, pp. 65–70.
[175] A. Wedel and U. Franke, "Monocular video serves radar-based emergency braking," in *Proc. IEEE Intell. Veh. Symp.*, 2007, pp. 93–98. doi: 10.1109/IVS.2007.4290097.
[176] H. Jha, V. Lodhi, and D. Chakravarty, "Object detection and identification using vision and radar data fusion system for ground-based navigation," in *Proc. 6th Int. Conf. Signal Process. Integrated Netw. (SPIN)*, 2019, pp. 590–593. doi: 10.1109/SPIN.2019.8711717.
[177] S. Chadwick, W. Maddetn, and P. Newman, "Distant vehicle detection using radar and vision," in *Proc. Int. Conf. Robot. Automat. (ICRA)*, 2019, pp. 8311–8317. doi: 10.1109/ICRA.2019.8794312.
[178] V. John and S. Mita, "RVNet: Deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments," in *Proc. Pacific-Rim Symp. Image Video Technol.*, 2019, pp. 351–364.
[179] V. Lekic and Z. Babic, "Automotive radar and camera fusion using generative adversarial networks," *Comput. Vision Image Understanding*, vol. 184, pp. 1–8, 2019. doi: 10.1016/j.cviu.2019.04.002.
[180] L. Li, W. Zhang, Y. Liang, and H. Zhou, "Preceding vehicle detection method based on information fusion of millimetre wave radar and deep learning vision," *J. Phys., Conf. Ser.*, vol. 1314, no. 1, p. 012063, 2019. doi: 10.1088/1742-6596/1314/1/012063.
[181] K.-C. Chang, R. K. Saha, and Y. Bar-Shalom, "On optimal track-to-track fusion," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 33, no. 4, pp. 1271–1276, 1997. doi: 10.1109/7.625124.
[182] X. Tian and Y. Bar-Shalom, "On algorithms for asynchronous track-to-track fusion," in *Proc. 13th Int. Conf. Inf. Fusion*, 2010, pp. 1–8.
[183] K. Yang, Y. Bar-Shalom, and P. Willett, "Track-to-track fusion with cross-covariances from radar and IR/EO sensor," in *Proc. 22th Int. Conf. Inf. Fusion (FUSION)*, 2019, pp. 1–5.
[184] K. Lu, K. Chang, and R. Zhou, "The exact algorithm for multi-sensor asynchronous track-to-track fusion," in *Proc. 18th Int. Conf. Inf. Fusion (Fusion)*, 2015, pp. 886–892.
[185] M. A. Khan, "Comparison of track to track fusion methods for nonlinear process and measurement models," in *Proc. Sensor Data Fusion: Trends, Solutions, Appl. (SDF)*, 2019, pp. 1–8. doi: 10.1109/SDF.2019.8916652.
[186] K.-E. Kim, C.-J. Lee, D.-S. Pae, and M.-T. Lim, "Sensor fusion for vehicle tracking with camera and radar sensor," in *Proc. 17th Int. Conf. Control, Automat. Syst. (ICCAS)*, 2017, pp. 1075–1077. doi: 10.23919/ICCAS.2017.8204375.
[187] D. Müller, J. Pauli, M. Meuter, L. Ghosh, and S. Müller-Schneiders, "A generic video and radar data fusion system for improved target selection," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2011, pp. 679–684. doi: 10.1109/IVS.2011.5940469.
[188] F. Liu, J. Sparbert, and C. Stiller, "IMMPDA vehicle tracking system using asynchronous sensor fusion of radar and vision," in *Proc. IEEE Intell. Veh. Symp.*, 2008, pp. 168–173.
[189] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *Proc. Sensor Data Fusion: Trends, Solutions, Appl. (SDF)*, 2019, pp. 1–7. doi: 10.1109/SDF.2019.8916629.
[190] S.-H. Bae, "Online multi-object tracking with visual and radar features," *IEEE Access*, vol. 8, pp. 90,324–90,339, 2020. doi: 10.1109/ACCESS.2020.2994000.
[191] V. Chandrakanth, A. Murthy, and S. S. Channappayya, "Target tracking in blind range of radars with deep learning," in *Proc. 21st Int. Radar Symp. (IRS)*, 2020, pp. 148–153. doi: 10.23919/IRS48640.2020.9253924.
[192] M. Bijelic, F. Mannan, T. Gruber, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep sensor fusion in the absence of labeled training data," 2019, arXiv:1902.08913.
[193] D. Feng et al., "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, 2020. doi: 10.1109/TITS.2020.2972974.
[194] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3943–3956, 2019. doi: 10.1109/TCSVT.2019.2920407.
[195] R. Kumar, R. Balasubramanian, and B. K. Kaushik, "Efficient method and architecture for real-time video defogging," *IEEE Trans. Intell. Transp. Syst.*, 2020.
[196] M. Hnewa and H. Radha, "Object detection under rainy conditions for autonomous vehicles," 2020, arXiv:2006.16471.
[197] A. M. Singh, S. Bera, and R. Bera, "Review on vehicular radar for road safety," in *Proc. Adv. Commun., Cloud, Big Data*, 2019, pp. 41–47.
[198] B. Li, B. Wu, J. Su, G. Wang, and L. Lin, "EagleEye: Fast sub-net evaluation for efficient neural network pruning," in *Proc. European Conf. Comput. Vision*, 2020. pp. 639–654.