

# Milestones in Autonomous Driving and Intelligent Vehicles Part II: Perception and Planning

Long Chen, *Senior Member, IEEE*, Siyu Teng, Bai Li, Xiaoxiang Na, Yuchen Li, Zixuan Li, Jinjun Wang, Dongpu Cao, *Senior Member, IEEE*, Nanning Zheng, *Fellow, IEEE* and Fei-Yue Wang, *Fellow, IEEE*

**Abstract**—Growing interest in autonomous driving (AD) and intelligent vehicles (IVs) is fueled by their promise for enhanced safety, efficiency, and economic benefits. While previous surveys have captured progress in this field, a comprehensive and forward-looking summary is needed. Our work fills this gap through three distinct articles. The first part, a "Survey of Surveys" (SoS), outlines the history, surveys, ethics, and future directions of AD and IV technologies. The second part, "Milestones in Autonomous Driving and Intelligent Vehicles Part I: Control, Computing System Design, Communication, HD Map, Testing, and Human Behaviors" delves into the development of control, computing system, communication, HD map, testing, and human behaviors in IVs. This part, the third part, reviews perception and planning in the context of IVs. Aiming to provide a comprehensive overview of the latest advancements in AD and IVs, this work caters to both newcomers and seasoned researchers. By integrating the SoS and Part I, we offer unique insights and strive to serve as a bridge between past achievements and future possibilities in this dynamic field.

**Index Terms**—Autonomous Driving, Intelligent Vehicles, Perception, Planning, Control, System Design, Communication, HD Map, Testing, Human Behaviors, Survey of Surveys.

## I. INTRODUCTION

AUTONOMOUS driving (AD) and intelligent vehicles (IVs) have recently attracted significant attention from academia as well as industry because of a range of potential benefits. Surveys on AD and IVs occupy an essential position in gathering research achievements, generalizing entire technology development, and forecasting future trends. However, a large majority of surveys only focus on specific tasks and lack

Manuscript received Sep 30, 2022; revised May 30, 2023. (Corresponding author: Fei-Yue Wang.)

This work is supported by the National Natural Science Foundation of China (62006256) and the Key Research and Development Program of Guangzhou (202007050002 202007050004).

Long Chen and Fei-Yue Wang are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, and Long Chen is also with Waytous Ltd. (e-mail: long.chen@ia.ac.cn; feiyue.wang@ia.ac.cn).

Siyu Teng and Yuchen Li are with BNU-HKBU United International College, Zhuhai, 519087, China and Hong Kong Baptist University, Kowloon, Hong Kong, 999077, China (e-mail: siyteng@ieee.org; liyuchen2016@hotmail.com).

Bai Li is with the College of Mechanical and Vehicle Engineering, Hunan University (e-mail: libai@zju.edu.cn).

Xiaoxiang Na is with the Department of Engineering, University of Cambridge (e-mail: xnna2@cam.ac.uk).

Zixuan Li is with Waytous Ltd. (e-mail: lizixuan981258655@gmail.com).

Jinjun Wang and Nanning Zheng are with the College of Artificial Intelligence, Xi'an Jiaotong University (e-mail: jinjun@mail.xjtu.edu.cn; nnzheng@mail.xjtu.edu.cn).

Dongpu Cao is with the School of Mechanical Engineering, Tsinghua University (e-mail: dp\_cao2016@163.com).

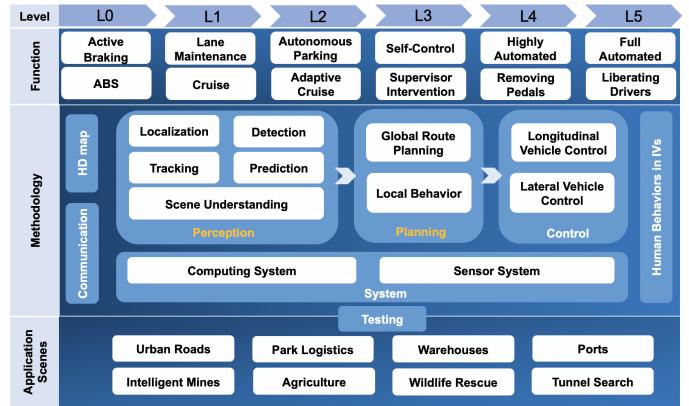


Fig. 1. The structure of autonomous driving with the function, methodology and application scenes

systematic summaries and research directions in the future. As a result, they may have a negative impact on conducting research for abecedarians. Our work consists of 3 independent articles including a Survey of Surveys (SoS) [1] and two surveys on crucial technologies of AD and IVs. Here is the third part (Part II of the survey) to systematically review the development of perception and planning. Combining with the SoS and the second part (Part I of the survey on control, system design, communication, High Definition map (HD map), testing, and human behaviors in IVs) [2], we expect that our work can be considered as a bridge between past and future for AD and IVs.

According to the different tasks in AD, we divide them into 8 sub-sections, perception, planning, control, system design, communication, HD map, testing, and human behaviors in IVs as Fig. 1. In Part I, we briefly introduce the function of each task and the intelligent levels for AD. Here, we describe classical applications in different AD scenes including urban roads, park logistics, warehouses, ports, intelligent mines, agriculture, wildlife rescue and tunnel search. It is more common for citizens to realize the AD in urban roads such as private IVs, AD taxis and buses. IVs in parts and ports require controllers to follow specific rules and achieve high efficiency. Warehouses and mines are classic closed scenes in indoor and outdoor environments. Modified IVs or called professional intelligent robots can be employed in wild to replace the human harbour in agricultural operations, wildlife rescue, tunnel search, etc. Indeed, AD and IVs could conduct a number of tasks in different scenes and play a crucial role

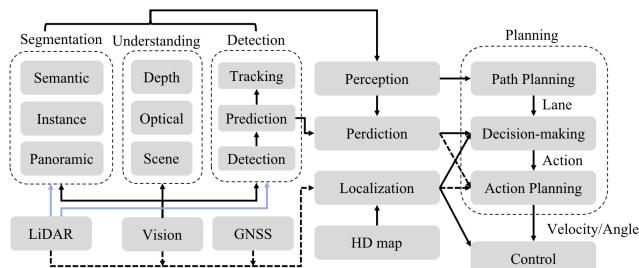


Fig. 2. The relationship of sub-tasks in perception, planning and the relationship of perception, planning and control

in our daily life.

In this paper, we consider 2 sub-sections as independent chapters, and each of them includes task definition, functional divisions, novel ideas, and a detailed introduction to milestones of AD and IVs, and the relationship of perception, planning and control can be seen in Fig. 2. The most important thing is that the research of them have rapidly developed for a decade and now entered a bottleneck period. We wish this article could be considered as a comprehensive summary for abecedarians and bring novel and diverse insights for researchers to make breakthroughs. We summarize three contributions of this article:

1. We provide a more systematic, comprehensive, and novel survey of crucial technology development with milestones on AD and IVs.
2. We introduce a number of deployment details, testing methods and unique insights throughout each technology section.
3. We conduct a systematic study that attempts to be a bridge between past and future on AD and IVs, and this article is the third part of our whole research (Part II for the survey).

## II. PERCEPTION

Perception is a fundamental module for AD. This module provides surrounding environmental information to the ego-vehicle. As can be seen in Fig. 3, perception is divided into localization, object detection, scene understanding, target prediction, and tracking.

### A. Localization

Localization is the technology for the driving platform to obtain its own position and attitude. It is an important prerequisite for the planning and control [3]. Currently, localization strategies are divided into four categories: Global Navigation Satellite System (GNSS) and Inertial Measurement Unit (IMU), visual Simultaneously Localization and Mapping (SLAM), LiDAR SLAM, and fusion-based SLAM [4].

1) *GNSS and IMU*: The GNSS [5] is a space-based radio navigation and localization system that can provide users with three-dimensional (3D) coordinates, velocity, and time information on the earth's surface. The IMU [6] is commonly composed of three-axis accelerometers and gyroscopes (additional three-axis magnetometers for 9 Degree of Freedom

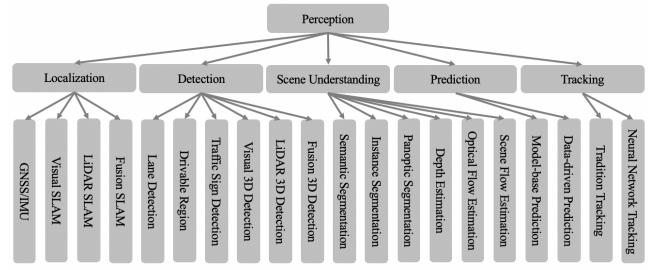


Fig. 3. The structure of perception methodology

(DOF) IMUs). By updating the localization information in low frequency from GNSS with dynamic states from IMUs, the platform could obtain the localization info in a high update frequency. Although the fusing the GNSS and IMU method is all-weather, the satellite signals will be interfered with by urban buildings [7].

2) *Visual SLAM*: Visual SLAM adopts the change of frames from cameras to estimate the ego-vehicle motion and this type of algorithms is divided into three categories by sensors: the monocular, multiply views, and depth. Specifically, visual SLAM algorithms only require images as inputs that means the cost of the localization system is relatively cheap [8]. However, they are dependent on abundant features and slight variation of illumination. In addition, optimization is a crucial module for visual localization system which updates each frame estimation after considering the global information and optimization methods include filter-based and graph-based [9].

There are two typical categories of visual SLAM from the perspective of feature extraction, key points [10–13] and optical flow methods [14–17]. Key points methods utilize points extraction approaches like SIFT, SURF, ORB, and descriptors to detect the same characteristics at different images and then compute relative motion among the frames. As points extraction approaches can extract crucial points stably and accurately, key points visual SLAM systems can offer significant benefits in structured roads and urban areas. However, the system may suffer difficulties when operating on an unstructured road or facing a flat white wall. Besides, earlier algorithms could not run in real-time and ignored most of the pixel information in the image. Optical flow methods assume that the photometric is invariant among the frames and attempt to estimate the camera motion by minimizing the photometric error on the images. This kind of method has several advantages as follows: 1) low computing overhead and high real-time performance; 2) weak dependence on key points; 3) considering whole pixels in the frames. However, due to the photometric assumption, it is sensitive for optical flow methods to the luminosity change between two images. visual SLAM systems also could be categorized into filter-based and optimization-based strategies from the optimization perspective, however, graph-based optimization methods have made a number of breakthroughs in accuracy and efficiency. Thus, researchers will continue to focus on the latter point in the future.

3) *LiDAR SLAM*: Compared with the visual SLAM methods, LiDAR SLAM systems detect surrounding environments actively with accurate 3D information because of the LiDARs properties. Similar to visual systems, the LiDAR SLAM also could be categorized into 2D such as Gmapping, Cartographer, Karto, and 3D [18–21] methods by sensors or filter-based like Gmapping and optimization-based by the process of optimization. Gmapping adopts the particle filter approach and separates the localization and mapping processes. During the optimization, each particle is responsible for maintaining a map. LOAM [18] operates two parallel algorithms, one is to calculate the motion transformation between frames in a low frequency through point cloud matching methods, and the other attempts to construct a map and correct the odometry in a high frequency. Segmap [21] utilizes deep neural networks to extract semantic feature information, which could reduce the computational resource consumption, and solve the data compression problem in real-time for indoor intelligent robotics and IVs. SUMA [20] transfers the point clouds into 2D space and adopts an extended RGB-D SLAM structure to generate a local map. Besides, it maintains and updates the surfel map by Iterative Closest Point (ICP) matching method for point clouds. LiDAR SLAM systems have the advantages of high accuracy, achieving a dense map, and weak dependency on lightness. However, no semantic information and environmental disturbance are two main challenges for LiDAR SLAM systems. In addition, researchers have to spend lots of time and effort to maintain and repair LiDARs installed on the IVs.

4) *Fusion-based SLAM*: In order to avoid the problems with failures in single sensor or low robustness, fusing multiple modalities data methods have been introduced by researchers including visual-inertial [22–24], LiDAR-inertial [25–28], visual-LiDAR inertial [29–31] and other fusion, such as adding sonars [32] or radars [33], SLAM approaches. We found that fusion methods usually introduce IMU data with higher updating frequency to SLAM systems. Loose fusion methods [18, 29] treat the external observation data from cameras or LIDARs and the internal motion data from IMUs as two independent modules, while the tight fusion approaches [24, 26, 30, 31] design a unit optimization module to solve and fuse multiple modalities data. Former methods could be considered as extended visual or LiDAR SLAM systems and are friendly for researchers to deploy on testing platforms and IVs. However, to increase the Robustness and adaptability, the tight fusion strategies provide appropriate solutions including introducing bundle adjustment into the visual odometry system [30] and adopting association optimization [31]. In summary, fusion-based SLAM methods solve several difficulties for a single sensor but still introduce a few challenges for jointing systems such as calibration, synchronization, and complex processing. The advantages and disadvantages of different methods for localization are shown in TABLE I.

### B. Object Detection

The purpose of object detection is to detect the static and dynamic targets in the field of view of the sensors. The results of some detection tasks can be seen in Fig. 4.

TABLE I  
THE ADVANTAGES AND DISADVANTAGES OF DIFFERENT LOCALIZATION METHODS

Method	Advantages	Disadvantages
GNSS	1.All-weather 2.Easy configuration 3.Non external sensors	1.Site requirements 2.GPS update frequency 3.High-cost (high accuracy)
Visual	1.Economy 2.Rich semantic info	1.Lightness effect 2.Similar feature 3.Low accuracy
LiDAR	1.High accuracy. 2.Dense map	1.Non semantic info 2.Environmental disturbance 3.High-cost
Fusion	1.High accuracy 2.Robustness	1.Complexity 2.Calibration 3.Synchronization

1) *Lane Detection*: Lane Detection is to recognize the lane in the views of sensors, to assist driving. For universal process, it involves three sections, including image pre-processing, lane detection, and tracking. The purpose of image pre-processing, such as Region of Interest (RoI) extraction, inverse perspective mapping, and segmentation, is to reduce the computing cost and eliminate noise. The methods of lane detection and tracking can be divided into the Computer Vision based (CV-based) method and the learning-based method [34].

CV-based methods in lane detection are broadly utilized nowadays, primarily because of their light computing cost and easy reproduction. A morphological top-hat transform is utilized to eliminate the irrelevant objects in the field [35]. After that, the Hough transform is applied to extract the edge pixel of the image and construct the straight lines. However, the disadvantage is that it is hard to detect the curve lines, so a number of researchers have introduced some effective methods on the Hough transform [35]. Some other lines estimation approaches involve the Gaussian Mixture Models (GMM) [36], Random Sample Consensus (RANSAC) [37], Kalman filter [38] in complex scenes.

Learning-based methods can be deployed on abundant scenes but they need a great deal of data to train the network with plenty of parameters. [39] attempts to design novel multiple sub-headers structures to improve the lane detection performance. To our knowledge, lane detection is integrated into the Advanced Driver Assistance Systems (ADAS) to keep the lane or follow the former vehicle, and researchers pay more attention to 3D lanes [40], lanes in closed areas, and unstructured roads.

2) *Driving Region Detection*: Driving region detection increases the obstacle information compared to lane detection, which offers the base information for obstacle avoidance function and path planning tasks. We also categorize this task into the CV-based and learning-based approaches.

Driving region detection can be converted to lane detection when the road surface is not obscured by obstacles. Otherwise, it can be seen as a combination of lane detection and 2D target detection. When considering driving region detection as an independent task, it needs to distinguish the road pixel from targets and non-driving regions. The color histogram can meet the requirement and some researchers develop methods on color [41] and efficiency [42] to tackle the poor performance

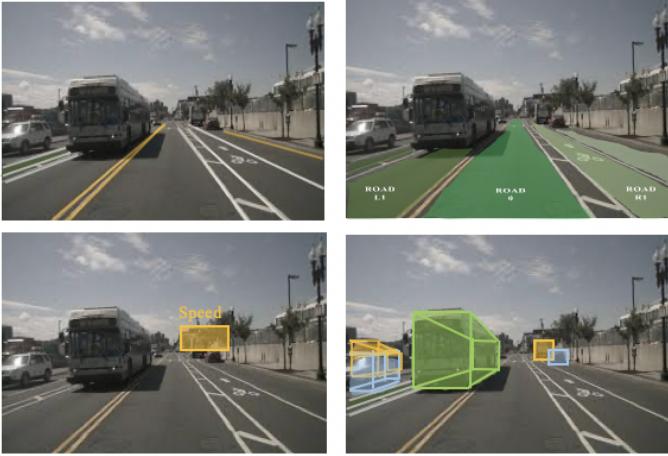


Fig. 4. The crucial tasks in the object detection for AD. The top-left, top-right, bottom-left, bottom-right figure represents the lane detection, drivable region, traffic sign detection and 3D object detection.

on color varying and low effect of it. Region growth methods [43] are more robust than the color histogram methods.

The learning-based methods in driving region detection are similar to image segmentation. For machine learning algorithms, features such as the RGB color, Walsh Hadamard, Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP), Haar, and LUV channel, can be extracted by the feature extractors and the classification header, such as Support Vector Machine (SVM), Conditional Random Field (CRF), to obtain the final results. The deep neural network can replace the feature extractors and some improvements, such as employing the large visual regions convolutional kernels [44], connection by multiple layers [45], to achieve competitive performance. We found that learning-based driving region detection results are usually one of the branches of the scene understanding task and researchers attempt to tackle a few challenges including 2D-3D transformation, complex driving regions, etc.

**3) Traffic Sign Detection:** Traffic signs contain plenty of crucial traffic information, such as road conditions, speed limitations, driving behavior restrictions, and other information. We also divide it into CV-based approaches and learning-based approaches.

For CV-based detection approaches, the conditions for the approximate color composition of traffic signs in a specific region (in a certain country or city) are similar. In the relatively simple original image, better results can be obtained by threshold separation of specific colors, which can be obtained by adopting the color space distribution, segmentation head, and the SVM classifier [46]. Some research develops the methods by introducing extra color channels, such as the normal RGB model, the dynamic color threshold, the probability model, and edge information. However, these approaches are hard to solve the problem caused by illumination, fading, occlusion, and bad weather. Some researchers tried to utilize the though transform on the triangle [47], circular traffic signs [48] or a coding gradient detection method [49], to handle occlusion and conjunction. The shape-based detection method can solve the problem of in-stable results caused by color change, but it has

little capability to overcome the problem caused by occlusion and deformation.

The traffic sign recognition algorithm based on machine learning usually uses the sliding window method to traverse the given traffic sign image. [50] proposed a variant histogram feature of gradient direction, and trained a single classifier to detect traffic signs through an extreme learning machine. With the continuous research of deep learning algorithms, more and more scholars use deep learning algorithms to detect traffic signs. Readers can regard this classification method as handling feature extraction, including pre-processing and classification [51]. To extract deeper information, the deeper encoder, integrated Space Pyramid Pooling (SPP) layer, cascaded RCNN network, depth separable convolution, and clipping strategy are introduced to achieve the detection accuracy and high inference speed [52]. The deep learning method has a satisfactory tolerance for the variation of the color and shape of signs, however, this type of method requires vast amounts of data and manual annotation. Besides, detection networks should have the capability of recognizing different regions with diffident signs and detecting signs over a long distance.

**4) Visual-based 3D Object Detection:** Visual-based 3D object detection refers to the process of acquiring 3D information (location, dimension, and rotation) about all targets in the field from the image. We divided it into monocular-based and stereo-based detection.

**4.1) Monocular 3D Object Detection:** Monocular 3D object detection is widely developed and the accuracy has been improving in recent years. Directly associating classification and regression methods, inheriting from 2D object detection networks like [53], have straightforward structures but perform unsatisfied due to the ill-posed problem of recovering 3D attributions from a single image. There are two main kinds of strategies to solve the shortcoming. 1) Some [54, 55] introduce the geometric connections between 3D and projected 2D candidates. For example, GS3D [56] decouples the objects into several parts to analyze the surface attributes and instance relationships. Monopair [54] and Monet3D [55] consider the relationships between the target and its two nearest neighbors. 2) Besides regressing the 3D bounding candidates, networks also take into account the local or full depth map [57, 58] from stereo vision or LiDAR data during the training stage. CaDDN [57] provides a fully differentiable end-to-end approach for combining depth estimation and object detection tasks. DDMP3D [58] utilizes the feature representation of context and depth estimation heads to achieve competitive performance. In addition, [59] introduces successive frames as inputs, which attempts to update the 3D results by associating detection and tracking. Although these methods have no obvious advantages in accuracy, extensive academic research and low cost make them attractive.

**4.2) Stereo 3D Object Detection:** Stereo 3D object detection approaches[60–63] are inspired by the parallax analysis from binocular vision. The precise depth value can be reckoned with through the distance between the binocular centers and the associated pair of pixels. Disp-RCNN [63], OC-Stereo [61] add segmentation modules paired images from stereo cameras to induce accurate association. YOLOStereo3D [62]

provides a lightweight model, outperforms a great number of stereo methods based on the complicated disparity convolution operations, and significantly reduces the length of training and testing time. In sum, stereo-based methods could avoid the ill-posed problem of monocular images and are convenient for manufacturers to deploy and maintain in IVs, but accurate measurement on the baseline, the time cost of binocular matching, and the requirement of image preprocessing pose challenges to researchers.

*5) LiDAR-based 3D Object Detection:* LiDAR-based 3D object detection methods recognize targets 3D properties from point clouds data captured by LiDARs. We categorized it into voxel-wise and point-wise detection.

*5.1) Voxel-wise Object Detection:* Voxel-wise object detection methods represent the point cloud features in the Birds Eyes View (BEV) and the BEV map is divided into a series of independent voxels manually. The structural design of this type of detection network evolves from point cloud segmentation frameworks, such as PointNet [64] and PointNet++ [65], which fit the detection task at the input or output side, and its overall architecture needs to balance performance and efficiency. Taking the classic VoxelNet [66] and PointPillar [67] as examples, VoxelNet normalizes the voxels after mapping point clouds, and subsequently applies local feature extraction using several Voxel Feature Encoding (VFE) layers to each non-empty voxel. The voxel-wise features are further abstracted by 3D convolutional middle layers (increasing the receptive field and learning the geometric spatial representation), and finally, the object is detected and classified using a Region Proposal Network (RPN) with position regression.

*5.2) Point-wise Object Detection:* Point-wise object detection such as [68, 69], are inspired by PointNet, a classical network for indoor 6D pose estimation with point clouds. Point-RCNN [68] is a two-stage point cloud detection framework including candidates generation with semantic segmentation analysis at the first stage and the position revision during the second stage. VoteNet [70] extends 2D detection structures to the 3D framework to establish a generic detection framework for point clouds. It basically follows the PointNet++ to reduce the information loss in point cloud transformation. VoteNet also introduces a novel voting mechanism inspired by the Hough transform to locate the targets' centers instead of point on the surface, compared with other 3D networks. It should be noticed that the number of discarded points and modality distinction due to the distance in point clouds detection should be significantly considered for researchers.

*6) Fusion-based 3D Object Detection:* LiDARs, radars, and cameras are widely deployed in IVs for perception tasks and combination of these types of sensors could make the vehicles robust and able to detect targets full-time. However, this does not mean that fusion-based methods will outperform the approaches with a single sensor. There are two main reasons for the disadvantage of fusion-based methods. 1) It is challenging for the network to fill the modalities gap from various sensors; 2) The system error and measurement errors such as from calibration and synchronization are hard to eliminate and they would be propagated and amplified in the networks. Most researchers propose solutions to handle these

TABLE II  
THE PERFORMANCE OF 3D OBJECT DETECTION METHODS IN KITTI

Method	Sensors	Moderate(%)	Easy(%)	Hard(%)
VPFNet[71]	Cam+LiD	83.21	91.02	78.20
DVF[72]	Cam+LiD	82.45	89.40	77.56
CLOCs[73]	Cam+LiD	82.28	89.16	77.23
F-ConvNet[74]	Cam+LiD	76.39	87.36	66.69
Point-RCNN[68]	LiDAR	75.64	86.96	70.70
PointPillars[67]	LiDAR	74.31	82.58	68.99
F-PointNet[75]	Cam+LiD	69.79	82.19	60.59
AVOD[76]	Cam+LiD	66.47	76.39	60.23
MV3D[77]	Cam+LiD	63.63	74.97	54.00
Disp-RCNN[63]	Stereo	43.27	67.02	36.43
YOLOStereo3D[62]	Stereo	41.25	65.68	30.42
OC-Stereo[61]	Stereo	37.60	55.15	30.25
Stereo-RCNN[60]	Stereo	30.23	47.58	23.72
MonoDETR[78]	Mono	16.26	24.52	13.93
MonoDTR[79]	Mono	15.39	21.99	12.73
CaDDN[57]	Mono	13.41	19.17	11.46
DDMP-3D[58]	Mono	12.78	19.71	9.80
GS3D[56]	Mono	2.90	4.47	2.47

difficulties and achieve some competitive outcomes. In this section, we categorize the fusion-based objection detection task based on the types of sensors.

*6.1) Camera and LiDAR:* Cameras and LiDARs are two crucial sensors for AD and researcher firstly focus on fusion parallel methods, which extract point clouds and images information at the same time. MV3D [77] and AVOD [76] utilize the shared 3D anchors on the point cloud and the corresponding images. ContFuse [80] and MMF [81] adopt tightly-coupled fusion approaches with a consecutive fusion layer. 3D-CVF [82] introduces a cross-view spatial feature fusion method to fuse the images and point clouds. In addition, EPNet [83] focuses on the point cloud system and projects the images on it with point-based strategy on the geometric space.

Compared with parallel approaches, sequential methods are readable and deployable because of no need to introduce association structures to reduce the gaps. F-PointNet [75] and F-ConvNet [74] attempt to reduce the searching areas by generating 3D bounding boxes within 2D candidates. PointPainting [84] outputs semantic information and projects each point on the corresponding point to improve 3D object detection accuracy. CLOCs [73] fuses the data after the independent extractors and achieve a competitive result on KITTI [85]. DVF [72] adopts the 2D truth as guidance and then extract 3D properties by the point clouds.

*6.2) Camera and Radar:* Combining the images and data from Radars can effectively reduce the cost and maintain accuracy. [86] projects radar detection results to the image space and utilize them to boost the object detection accuracy for distant objects. CRF-Net [87] develops the method with a vertical presentation.

*6.3) LiDAR and Radar:* This type of fusion focuses on extremely harsh weather conditions and distinct targets. Radar-Net [88] fuses radar and LiDAR data via a novel-based early fusion approach. It leverages the radar's long sensing range via an attention-based fusion. MVDNet [89] generates proposals from two sensors and then fuses region-wise features between multi-modal sensor streams to improve final detection results. ST-MVDNet [90] develops the structure by enforcing output

consistency between a Teacher network and a Student network, and introducing missing modalities to tackle the degeneration problem when one type of data is missing.

*6.4) Camera, LiDAR and Radar:* In this fusion type, researchers attempt to design a robust perception system in different weather conditions. [91] obtains object detection outputs with a PointNet [64] architecture by projecting the images onto the point cloud directly. Parallel to the previous frame, the point cloud from the radar is processed to predict velocity which is then associated with the final detection output. RVF-Net [92] fuses all of the data on the input procedure and achieves satisfying results on the nuScenes [93] data set.

*6.5) Others:* Ultrasonic radar judges the distance of obstacles through the time of sound transmission in the air, and the accuracy can achieve a centimeter scale within 5 meters. This sensor is mostly used in autonomous parking scenes. An infrared camera with an infrared lamp can capture the infrared spectral characteristics to achieve the effect of night vision imaging. Besides, research on event cameras is one of the hot topics nowadays. Event cameras process data based on pipeline timestamps, rather than processing individual pixels in a frame plane. Because the data has the nature of time sequence, the traditional network structure can not process the data, so how to fuse with other sensors will be one of the research points in the future.

The performance of 3D object detection methods with various combination of different sensor types in KITTI [85] is shown in TABLE II. Here, KITTI divides whole data into three evaluation scenes (easy, moderate and hard) through the frame's complexity and computes 3D-AP, an extended method from 2D-AP [94] on these three scenes. We summarize: 1) Adopting fusion strategies could achieve competing results for 3D object detection tasks mainly because of introducing more initial information. But this type of method requires researchers to eliminate or reduce modal differences. 2) Due to the characteristics of sensors, limited resolutions of cameras, and the definition of the reference system in KITTI, the performance of visual-based methods is weaker than LiDAR-based. However, visual-based methods attract a number of researchers because of their maintainability, economy, and easy deployment. 3) The self-attention mechanism (Transformer structure) and BEV method [71, 78, 79] could improve the accuracy of cross-modality fusion, feature extraction, etc. In addition, to address data hungry and model robustness, current research studies train and test models on additional data such as unScenes[93], Waymo[95].

### C. Scene Understanding

We define scene understanding in our paper as the multiple outputs for each pixel or point instead of each target. This section, we divide it into three sub-sections, segmentation, depth, and flow estimation. We only focus on academic research and applications in AD areas.

*1) Segmentation in Autonomous Driving:* The target of semantic segmentation is to partition a scene into several meaningful parts, usually by labeling each pixel in the image

TABLE III  
RESULTS ON CITYSCAPE FOR PANOPTIC SEGMENTATION AND KITTI FOR DEPTH ESTIMATION

Method	PQ(%)	SQ(%)	RQ(%)
Axial-D[109]	62.7	82.2	75.3
TASC[110]	60.7	81.0	73.8
Method	SILog(log(m))	sqErrorRel(%)	iRMSE(1/km)
BANet[111]	0.1155	2.31	12.17
VNL[112]	0.1265	2.46	13.02
SDNet[113]	0.1468	3.90	15.96
MultiDepth[114]	0.1605	3.89	18.21

with semantics (semantic segmentation), by simultaneously detecting objects and distinguishing each pixel from each object (instance segmentation), or by combining semantic and instance segmentation (panoptic segmentation) [96]. The segmentation is one of the crucial tasks in computer vision and researchers evaluate their models on ADE20K [97], Pascal-VOC [94], CityScape [98], etc. However, in AD scenarios, the classic 3D CV area, it is hard to complete the perception task independently. It is usually involved in lane detection, driving region detection, visual interface module, or combined with point clouds to provide semantic information. We will briefly introduce the general background based on segmentation, and then highlight segmentation research on AD.

*1.1) Semantic Segmentation:* Fully Convolutional Network (FCN) [99] is a popular structure for semantic segmentation which adopts convolutional layers to recover the size of output maps. Some work extends FCN by introducing an improved encoder-decoder [100], the dilated convolutions [101], CRFs [102], atrous spatial pyramid pooling(ASPP) [103]. In addition, the above approaches attend to fixed, square context regions because of the pooling and dilation convolution operations. The relational context method [104] extracts the relationship between pixels. [105] pursues high resolution by channel concatenation and skip connection, especially in the medical field. In the field of AD, the semantic segmentation networks may be familiar with the common structures, and researchers should pay more attention to the special categories, and occlusion, and evaluate their models on data sets of road scenarios [98]. To achieve SOTA results on data sets, the researcher introduces the multiple scale attention mechanism [106], boundary-aware segmentation module [107]. Besides, some research focuses on the targets' attributes on roads like considering the intrinsic relevance among the cross-class objects [108] or semi-supervised segmentation mechanism because of the lack of labeled data on AD scenarios.

*1.2) Instance Segmentation:* Instance segmentation is to predict a mask and its corresponding category for each object instance. Early method [115] designs an architecture to realize both object detection and segmentation missions. Mask-RCNN [115] extends Faster-RCNN to identify each pixel's category with binary segmentation and pools image features from Region of Interest (RoI) following a Region Proposal Network (RPN). Some researchers develop the base structure by introducing a coefficients network [116], the IoU score for each mask, and shape priors to refine predictions. Similar with the 2D object detection methods, [117] replaces the detectors

with the one-stage structures. [118] attempts to avoid the effect of detection and achieve remarkable performances. To achieve competitive segmentation results on AD datasets, researchers focus on the geometric information on 3D space [119], boundary recognition [120], combining the semantic segmentation (panoptic segmentation) [121] or intruding multiple frames (video-base) [122].

**1.3) Panoptic Segmentation:** Panoptic segmentation is proposed to unify pixel-level and instance-level semantic segmentation [123], and [124] designs a different branch to regress the semantic and instance segmentation results. Panoptic-FCN [125] aims to represent and predict foreground things and background stuff in a unified fully convolutional pipeline. Panoptic SegFormer [126] introduces a concise and effective framework for panoptic segmentation with transformers. For AD scenarios, TASC [110] proposes a new differentiable approach to reduce the gap between the two sub-tasks during training. Axial-DeepLab [109] builds a stand-alone attention model with a global receptive field and a position-sensitive attention layer to capture the positional information with low computational cost. Besides, researchers address the multiple scales on roads by introducing a novel crop-aware bounding box regression loss and a sample approach [127], and capture the targets' boundary by a combinatorial optimization strategy. These methods achieve competitive results on the task of CityScape [98] or Mapillary Vistas [128].

**2) Depth Estimation in Autonomous Driving:** This type of task is to present the depth information on the camera plane, which is an effective way to enhance the visual-based 3D object detection and a potential bridge to connect the LiDAR and camera.

The depth completion task is a sub-problem of depth estimation [129]. In the sparse-to-dense depth completion problem, researchers infer the dense depth map of a 3D scene from a sparse depth map by computational methods or multiple data from sensors. The main difficulties include: 1) the irregularly spaced pattern in the sparse depth, 2) the fusion methods for multiple sensor modalities (optional), and 3) the lack of dense pixel-level ground truth for some data and the real world (optional).

Depth estimation is the task of measuring the distance of each pixel relative to the camera. The depth value is extracted from either monocular or stereo images with supervised (the dense map obtained by depth completion) [130], unsupervised [131], LiDAR guidance [132] or stereo computing [133]. Some approaches [134, 135] introduce the CRF module, multi-tasks structure, global extractor, and the piece-wise planarity priors to achieve competitive performances in popular benchmarks such as KITTI [85] and NYUv2 [136]. Models are typically evaluated according to an RMS metric [85].

For outdoor monocular depth estimation, DORN [137] adopts a multi-scale network structure to capture the contextual information. MultiDepth [114] makes use of depth interval classification as an auxiliary task. HGR [138] proposes a hierarchical guidance and regularization learning framework to estimate the depth. SDNet [113] improves the results by utilizing a dual independent estimation head involving depth and semantics. VNL [112] designs a novel structure

that includes local planar guidance layers at multiple stages. [139] uses the geometric constraints of normal directions determined by randomly sampled three points to improve the depth prediction accuracy. BANet [111] introduces bidirectional attention modules which adopt the feed-forward feature maps and incorporate the global information to eliminate ambiguity. The Unsupervised method [140] attracts plenty of researchers because it could reduce the requirements on the labeled data and eliminate the over-fit problem. In addition, the pure monocular depth estimation only obtains the relative depth value because of the ill-posed problem, and the stereo guidance methods could obtain the absolute depth value. [141] introduces the Transformer structures to achieve competitive results. The stereo depth estimation methods can be found in the stereo disparity estimation task.

**3) Flow Estimation in Autonomous Driving:** Similar to the segmentation and depth estimation tasks, flow estimation focuses on the image plane and it presents the pixel movement during a data frame. It attracts interest nowadays and its research can be used in event camera methods.

**3.1) Optical Flow Estimation:** Optical flow refers to the pixels' movement in the imaging system including two directions, the horizontal and vertical. Similar to unsupervised video-based depth estimation, the pixel motion [142] can be deduced by minimizing differences between the target and source images. SPyNet [143] proposes a lightweight framework that adopts classical spatial-pyramid formulation for optical flow estimation. In addition, it attempts to estimate large-displacement movement and accurate sub-pixel flow. PWC-Net [144] includes three sub-nets, the feature pyramid extractor, warping layer, and cost volume layer, to improve the quality of optical flow.

**3.2) Scene Flow Estimation:** Scene flow estimation indicates a 3-dimensional movement field which can be treated as the extension of optical flow. Therefore, it is the combination of optical flow and depth estimation in 3D scenarios. Monocular images are seldom utilized in the holistic training step for scene flow, and the structure takes the binocular videos as input to regress disparity to restore the scale. DRISF [145] treats the inference step of Gaussian Newton (GN) as a Recurrent Neural Network (RNN) which means it can be trained in an end-to-end method. FD-Net [146] further extends the unsupervised depth estimation and disentangles the full flow into object flow (targets pixels) and rigid flow (background pixels) to assess the characteristics respectively, which is able to avoid the warping ambiguity due to the occlusion and truncation. Competitive Collaboration (CC) [147] sets the scene flow estimation as a game with three players. Two of them compete for the resource and the last one acts as a moderator. GeoNet [148] consists of two modules, a monocular depth with the 6 DoF ego-motion estimation, and a residual network to learn the object's optical flow.

The performance of panoptic segmentation and depth estimation on CityScape and KITTI is shown in TABLE III. PQ, SQ, RQ refer the panoptic segmentation, segmentation quality, and recognition quality respectively in [123], and for depth estimation, SILog (Scale invariant logarithmic error), sqErrorRel (Relative squared error), and iRMSE (Root mean

squared error of the inverse depth) are classical metrics in KITTI. Similar to detection, researchers introduce the self-attention mechanism, extra training data and novel network units to develop the accuracy in scene understanding tasks. And we mention that above tasks do not directly provide outputs to the downstream tasks such as planning and control in AD. In the actual tasks, semantic segmentation, depth estimation and optical flow estimation will be combined with each other to provide richer pixel semantic information, so as to improve the accuracy of cross-modality data fusion, spatial detection and tracking for moving targets.

#### D. Prediction

In order to safely and efficiently navigate in complex traffic scenarios, an AD framework should be able to predict the way in which the other traffic agents (such as vehicles and pedestrians) will behave in the near future. Prediction can be defined as probable results according to past perceptions. Let  $X_t^i$  be a vector with the spatial coordinates of agents  $i$  at observation time  $t$ , with  $t \in \{X_1^i, X_2^i, \dots, X_{T_{obs}}^i\}$ .

1) *Model-based Approaches*: These methods predict the behaviors of agents, such as changing lanes, turning left, and so on. One of the simplest methods to predict the probability distribution of vehicle behavior is the autonomous multiple models (AMM) algorithm. This algorithm computes the maximum probability trajectory of each agent.

2) *Data-driven Approaches*: These methods are mainly composed of the neural network. After training on the perception dataset, the model makes a prediction of the next behavior. DESIRE [149] proposes an encoder-decoder framework that innovatively incorporates the scenario context and the interactions between traffic agents. SIMP [150] discretizes the output space, calculates the distribution of the vehicle's destination, and predicts an estimated time of arrival and a spatial offset. FaF [151] pioneers the unification of detection and short-term motion forecasting based on LiDAR point clouds. The prediction module is sometimes separated from the perception, mainly because the downstream planning module receives both the perception and the prediction results. Future research on prediction will focus on the formulation of generalized rules, the universality of scenarios and the simplicity of modules.

#### E. Tracking

The tracking problem begins with a sequence of vehicle-mounted sensor data. Depending on if neural network is embedded in the tracking framework, we divide them into the traditional method and the neural network method.

1) *Tradition Method*: The Kalman filter[152] is a famous algorithm, particularly with regard to tracking agents. Because of the low computational cost, the Kalman-based method [153] has quick response time even on low-spec hardware in simple scenarios.

The tracking problem also can be shown as a graph search problem [154]. Compared with Kalman-based methods, The most important advantage of graph-based approach is that it is better for the multi-tracking problems. [155] exploits graph-based methods using the min-cost approach to solve tracking problems.

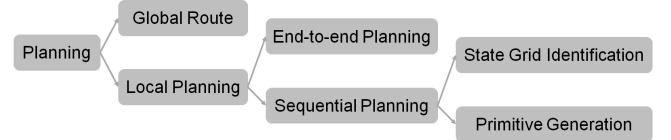


Fig. 5. The structure of planning methodology

2) *Neural Network Method*: Neural networks have the advantage of being able to learn important and robust features given training data that is relevant and with sufficient quantity.

CNN is widely used in agents tracking. [156] handles multi-agent tracking using combinations of values from convolutional layers. [157] proposes appropriate filters for information drawn from shallow convolutional layers, achieving the same level of robustness compared with deeper layers or a combination of multiple layers.

RNN also provides a smart method to solve temporal coherence problem in tracking task. [158] uses an LSTM-based classifier to track agents across multiple frames in time. Compared with CNN method, the LSTM-based approach is better suited to remove and reinsert candidate observations particularly when objects leave or reenter the visible area of the scene. Joint perception and tracking can achieve the SOTA results in these two tasks. In reality, stable tracking can reduce the requirements of the system for real-time detection and can also correct the detection results. At present, the strategy of joint task learning has been favored by more and more researchers.

### III. PLANNING

The planning module is responsible for finding a local trajectory for the low-level controller of the ego vehicle to track.

The planning module is responsible for finding a local trajectory for the low-level controller of the ego vehicle to track. Herein, “local” means that the resultant trajectory is short in its spatial or temporal range; otherwise the ego vehicle cannot react to risks beyond the sensor ranges. The planning module typically contains three functions, namely global route planning, local behavior planning, and local trajectory planning [159]. Global route planning provides a road-level path from the start point to the destination on a global map; local behavior planning decides a driving action type (e.g., car-following, nudge, side pass, yield, and overtake) for the next several seconds while local trajectory planning generates a short-term trajectory based on the decided behavior type. This section reviews the techniques related to the three functions in the planning module as Fig. 5.

#### A. Global Route Planning

Global route planning is responsible for finding the best road-level path in a road network, which is presented as a directed graph containing millions of edges and nodes. A route planner searches in the directed graph to find the minimal-cost sequence that links the starting and destination nodes. Herein, the cost is defined based on the query time, preprocessing

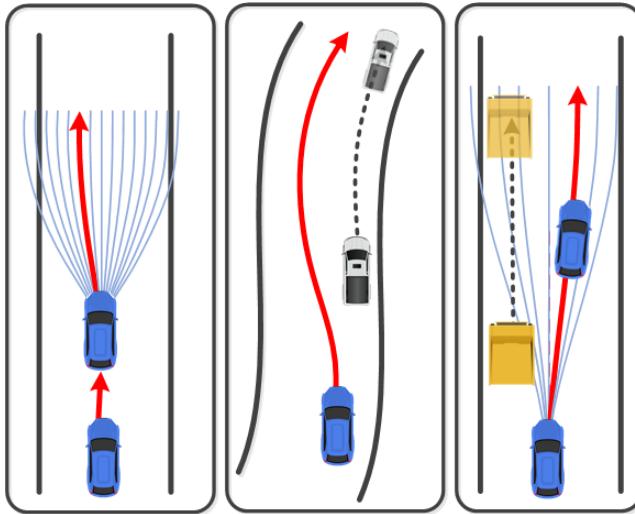


Fig. 6. The route section for local behavior planning

complexity, memory occupancy, and/or solution robustness considered.

The development history of global route planning techniques is much longer than that of autonomous vehicle techniques because global route planning also serves manually driven vehicles. As indicated by [160], the existing global route planning methods are classified as goal-directed methods, separator-based methods, hierarchical methods, bounded-hop methods, and their combinations.

### B. Local Behavior/Trajectory Planning

Local behavior planning and local trajectory planning functions work together to output a local trajectory along the identified global route as shown in Fig. 6. Since the resultant trajectory is local, the two functions have to be implemented in a receding-horizon way unless the global destination is not far away [161]. It deserves to emphasize that the output of the two functions should be a trajectory rather than a path [162], otherwise extra efforts are needed for the ego vehicle to evade the moving obstacles in the environment.

Broadly speaking, the two functions would work in two different ways. One is the end-to-end way, i.e., to develop an integrated system that receives the raw data from the on-board sensors and outputs a local trajectory directly. The other way is to implement the local behavior planning and local trajectory planning functions sequentially.

1) *End-to-end Methods*: Compared with the sequential-planning solution reviewed in the next subsection, an end-to-end solution nominally deals with vehicle-environment interactions more efficiently because there is not an external gap between the perception and planning modules[163]. The input of an end-to-end system is the large amount of raw data obtained by the on-board sensors whereas the output is a local trajectory. Since the relationship between the input and output is too intricate to be summarized as complete rules [164], machine learning methods are commonly used, most of which are classified as imitation-learning-based and reinforcement-learning-based methods [165].

An imitation-learning-based method builds a neuro network based on training samples [166, 167]. Challenges lie in how to collect massive training samples that are consistent and how to guarantee learning efficiency (e.g., free from overfitting). Reinforcement-learning-based methods obtain knowledge by trial-and-error operations, thus they rely less on the quality and quantity of external training samples [168].

End-to-end methods are still not mature, thus most of them are trained/tested in simulations rather than real-world scenarios . Recent research efforts focus on how to enhance learning interpretability, security, and efficiency.

2) *Sequential-planning-based Methods*: As opposed to the aforementioned end-to-end solution, applying local behavior planning and trajectory planning functions sequentially has been a common and conventional choice in the past decade. However, the boundary between local behavior planning and trajectory planning is rather blurred [159], e.g., some behavior planners do more than just identify the behavior type. For the convenience of understanding, this paper does not distinguish between the two functions strictly and the related methods are simply regarded as trajectory planning methods.

Nominally, trajectory planning is done by solving an optimal control problem (OCP), which minimizes a predefined cost function with multiple types of hard constraints satisfied [169]. The solution to the OCP is presented as time-continuous control and state profiles, wherein the desired trajectory is reflected by (part of) the state profiles. Since the analytical solution to such an OCP is generally not available, two types of operations are needed to construct a trajectory.

Concretely, the first type of operation is to identify a sequence of state grids while the second type is to generate primitives between adjacent state grids.

2.1) *State Grid Identification*: State grid identification can be done by search, selection, optimization, or potential minimization. Search-based methods abstract the continuous state space related to the aforementioned OCP into a graph and find a link of states there. Prevalent search-based methods include A\* search [170] and dynamic programming (DP) [171]. Selection-based methods decide the state grids in the next one or several steps by seeking the candidate with the optimal cost/reward function value. Greedy selection [172] and Markov decision process (MDP) series methods typically [173] fall into this category. An optimization-based method discretizes the original OCP into a mathematical program (MP), the solution of which are high-resolution state grids [174, 175]. MP solvers are further classified as gradient-based and non-gradient-based ones; gradient-based solvers typically solve nonlinear programs [169, 175], quadratic programs [176, 177], quadratically constrained quadratic programs [178] or mix-integer programs [179]; non-gradient-based solvers are typically represented by metaheuristics [180]. Potential-minimization-based methods adjust the state grid positions by simulating the process they are repelled or attracted by forces or in a heuristic potential field. Prevalent methods in this category include the elastic band (EB) series [181], artificial potential field methods [162], and force-balance model [182].

The capability of each state grid identification method is different. For example, gradient-optimization-based and

potential-minimization-based methods are generally more flexible and stable than typical search-/selection-based methods [183], but search-/selection-based methods are more efficient to explore the entire state space globally [181, 184, 185]. Different methods could be combined jointly as a coarse-to-fine strategy, as has been implemented by many studies such as [169, 175, 176, 178].

**2.2 Primitive Generation:** Primitive generation is commonly done by closed-form rules, simulation, interpolation, and optimization. Closed-form rules refer to methods that generate primitives by analytical methods with closed-form solutions. Typical methods include the Dubins/Reeds-Shepp curves [186], polynomials [172], and theoretical optimal control methods [187]. Simulation-based methods generate trajectory/path primitives by forward simulation, which runs fast because it has no degree of freedom [188]. Interpolation-based methods are represented by splines or parameterized polynomials. Optimization-based methods solve a small-scale OCP numerically to connect two state grids [189].

State grid identification and primitive generation are two necessary operations to construct a trajectory. Both operations may be organized in various ways. For example, [188] integrates both operations in an iterative loop; [189] builds a graph of primitives offline before online state grid identification; [176] identifies the state grids before generating connective primitives.

If a planner only finds a path rather than a trajectory, then a time course should be attached to the planned path as a post-processing step [190]. This strategy, denoted as path velocity decomposition (PVD), has been commonly used because it converts a 3D problem into two 2D ones, which largely facilitates the solution process. Conversely, non-PVD methods directly plan trajectories, which has the underlying merit to improve the solution optimality [171, 191, 192].

Recent studies in this research domain include how to develop specific planners that fit specific scenarios/tasks particularly [161], and how to plan safe trajectories with imperfect upstream/downstream modules [193].

#### IV. CONCLUSION

This article is the third part of our work (Part II for the technology survey). In this paper, we provide a review of wide introductions on research development with milestones of perception and planning in AD and IVs. In addition, we provide a few experiment results and unique opinions for these two tasks. In combination with the other two parts, we expect that our whole work will bring novel and diverse insights to researchers and abecedarians, and serve as a bridge between past and future.

#### REFERENCES

- [1] L. Chen, Y. Li, C. Huang, B. Li, Y. Xing, D. Tian, L. Li, Z. Hu, X. Na, Z. Li, S. Teng, C. Lv, J. Wang, D. Cao, N. Zheng, and F.-Y. Wang, "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046–1056, 2023.
- [2] L. Chen, Y. Li, C. Huang, Y. Xing, D. Tian, L. Li, Z. Hu, S. Teng, C. Lv, J. Wang, D. Cao, N. Zheng, and F.-Y. Wang, "Milestones in autonomous driving and intelligent vehicles—part 1: Control, computing system design, communication, hd map, testing, and human behaviors," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–17, 2023.
- [3] S. Kuutti, S. Fallah, K. Katsaros, M. Dianati, F. McCullough, and A. Mouzakitis, "A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 829–846, 2018.
- [4] T. T. O. Takleh, N. A. Bakar, S. A. Rahman, R. Hamzah, and Z. Aziz, "A brief survey on SLAM methods in autonomous vehicle," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 38–43, 2018.
- [5] S. Chen, B. Liu, C. Feng, C. Vallespi-Gonzalez, and C. Wellington, "3d point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 68–86, 2020.
- [6] C. Gentner, T. Jost, W. Wang, S. Zhang, A. Dammann, and U.-C. Fiebig, "Multipath assisted positioning with simultaneous localization and mapping," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6104–6117, 2016.
- [7] J. Zhang and S. Singh, "LOAM: Lidar odometry and mapping in real-time," in *Robotics: Science and Systems*, vol. 2, no. 9. Berkeley, CA, 2014, pp. 1–9.
- [8] S. Thrun, "Simultaneous localization and mapping," in *Robotics and Cognitive Approaches to Spatial Mapping*. Springer, 2007, pp. 13–41.
- [9] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 15–22.
- [10] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [11] R. Wang, M. Schworer, and D. Cremers, "Stereo dso: Large-scale direct sparse visual odometry with stereo cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3903–3911.
- [12] T. Pire, T. Fischer, G. Castro, P. De Cristóforis, J. Civera, and J. J. Berlles, "S-ptam: Stereo parallel tracking and mapping," *Robotics and Autonomous Systems*, vol. 93, pp. 27–42, 2017.
- [13] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.
- [14] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part II 13*. Springer, 2014, pp. 834–849.
- [15] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.
- [16] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 15–22.
- [17] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2320–2327.
- [18] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time," in *Robotics: Science and Systems*, vol. 2, no. 9. Berkeley, CA, 2014, pp. 1–9.
- [19] T. Shan and B. Englot, "Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4758–4765.
- [20] J. Behley and C. Stachniss, "Efficient surfel-based slam using

- 3d laser range data in urban environments.” in *Robotics: Science and Systems*, vol. 2018, 2018, p. 59.
- [21] R. Dubé, A. Cramariuc, D. Dugas, J. Nieto, R. Siegwart, and C. Cadena, “Segmap: 3d segment mapping using data-driven descriptors,” *arXiv preprint arXiv:1804.09557*, 2018.
- [22] G. Chevrin, S. Changey, M. Rebert, D. Monnin, and J.-P. Lauffenburger, “Visual-inertial fusion on kitti using msf-ekf,” in *2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2022, pp. 35–40.
- [23] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, “Keyframe-based visual–inertial odometry using non-linear optimization,” *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.
- [24] T. Qin, P. Li, and S. Shen, “Vins-mono: A robust and versatile monocular visual-inertial state estimator,” *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.
- [25] H. Ye, Y. Chen, and M. Liu, “Tightly coupled 3d lidar inertial odometry and mapping,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3144–3150.
- [26] S. Zhao, Z. Fang, H. Li, and S. Scherer, “A robust laser-inertial odometry and mapping method for large-scale highway environments,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1285–1292.
- [27] C. Qin, H. Ye, C. E. Pranata, J. Han, S. Zhang, and M. Liu, “Lins: A lidar-inertial state estimator for robust and efficient navigation,” in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 8899–8906.
- [28] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, “Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping,” in *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2020, pp. 5135–5142.
- [29] J. Zhang and S. Singh, “Visual-lidar odometry and mapping: Low-drift, robust, and fast,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 2174–2181.
- [30] J. Graeter, A. Wilczynski, and M. Lauer, “Limo: Lidar-monocular visual odometry,” in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2018, pp. 7872–7879.
- [31] W. Shao, S. Vijayarangan, C. Li, and G. Kantor, “Stereo visual inertial lidar simultaneous localization and mapping,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 370–377.
- [32] M. S. Khan, S. S. Chowdhury, N. Niloy, F. T. Z. Aurin, and T. Ahmed, “Sonar-based slam using occupancy grid mapping and dead reckoning,” in *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE, 2018, pp. 1707–1712.
- [33] Z. Hong, Y. Petillot, A. Wallace, and S. Wang, “Radar slam: A robust slam system for all weather conditions,” *arXiv preprint arXiv:2104.05347*, 2021.
- [34] J. Ma, K. Zhang, and J. Jiang, “Loop closure detection via locality preserving matching with global consensus,” *IEEE/CAA Journal of Automatica Sinica*, pp. 1–16, 2022.
- [35] R. K. Satzoda, S. Sathyanarayana, T. Srikanthan, and S. Sathyanarayana, “Hierarchical additive hough transform for lane detection,” *IEEE Embedded Systems Letters*, vol. 2, no. 2, pp. 23–26, 2010.
- [36] R. Laxhammar, G. Falkman, and E. Sviestins, “Anomaly detection in sea traffic-a comparison of the gaussian mixture model and the kernel density estimator,” in *2009 12th International Conference on Information Fusion*. IEEE, 2009, pp. 756–763.
- [37] A. S. Huang, D. Moore, M. Antone, E. Olson, and S. Teller, “Finding multiple lanes in urban road networks with vision and lidar,” *Autonomous Robots*, vol. 26, no. 2, pp. 103–122, 2009.
- [38] A. Borkar, M. Hayes, and M. T. Smith, “Robust lane detection and tracking with ransac and kalman filter,” in *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009, pp. 3261–3264.
- [39] L. Liu, X. Chen, S. Zhu, and P. Tan, “Condlanenet: a top-to-down lane detection framework based on conditional convolution,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3773–3782.
- [40] F. Yan, M. Nie, X. Cai, J. Han, H. Xu, Z. Yang, C. Ye, Y. Fu, M. B. Mi, and L. Zhang, “Once-3dlanes: Building monocular 3d lane detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17143–17152.
- [41] Q. Li, L. Chen, M. Li, S.-L. Shaw, and A. Nüchter, “A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios,” *IEEE Transactions on Vehicular Technology*, vol. 63, no. 2, pp. 540–555, 2013.
- [42] C. Guo, S. Mita, and D. McAllester, “Drivable road region detection using homography estimation and efficient belief propagation with coordinate descent optimization,” in *2009 IEEE Intelligent Vehicles Symposium*. IEEE, 2009, pp. 317–323.
- [43] Q. Li, L. Chen, M. Li, S.-L. Shaw, and A. Nüchter, “A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios,” *IEEE Transactions on Vehicular Technology*, vol. 63, no. 2, pp. 540–555, 2013.
- [44] H. Xue, H. Fu, R. Ren, J. Zhang, B. Liu, Y. Fan, and B. Dai, “LiDAR-based drivable region detection for autonomous driving,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1110–1116.
- [45] D. Qiao and F. Zulkernine, “Drivable area detection using deep learning models for autonomous driving,” in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 5233–5238.
- [46] S. Maldonado-Bascón, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gómez-Moreno, and F. López-Ferreras, “Road-sign detection and recognition based on support vector machines,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 264–278, 2007.
- [47] S. Houben, “A single target voting scheme for traffic sign detection,” in *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2011, pp. 124–129.
- [48] M. Á. García-Garrido, M. Á. Sotelo, and E. Martín-Gorostiza, “Fast road sign detection using hough transform for assisted driving of road vehicles,” in *International Conference on Computer Aided Systems Theory*. Springer, 2005, pp. 543–548.
- [49] M. Boumediene, C. Cudel, M. Basset, and A. Ouamri, “Triangular traffic signs detection based on RSLD algorithm,” *Machine Vision and Applications*, vol. 24, no. 8, pp. 1721–1732, 2013.
- [50] Z. Huang, Y. Yu, J. Gu, and H. Liu, “An efficient method for traffic sign recognition based on extreme learning machine,” *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 920–933, 2016.
- [51] J. Zhang, Z. Xie, J. Sun, X. Zou, and J. Wang, “A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection,” *IEEE Access*, vol. 8, pp. 29742–29754, 2020.
- [52] J. Zhang, M. Huang, X. Jin, and X. Li, “A real-time chinese traffic sign detection algorithm based on modified yolov2,” *Algorithms*, vol. 10, no. 4, p. 127, 2017.
- [53] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [54] Y. Chen, L. Tai, K. Sun, and M. Li, “Monopair: Monocular 3d object detection using pairwise spatial relationships,” in

- Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 093–12 102.
- [55] X. Zhou, Y. Peng, C. Long, F. Ren, and C. Shi, “Monet3d: Towards accurate monocular 3d object localization in real time,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 503–11 512.
- [56] B. Li, W. Ouyang, L. Sheng, X. Zeng, and X. Wang, “Gs3d: An efficient 3d object detection framework for autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1019–1028.
- [57] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, “Categorical depth distribution network for monocular 3d object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8555–8564.
- [58] L. Wang, L. Du, X. Ye, Y. Fu, G. Guo, X. Xue, J. Feng, and L. Zhang, “Depth-conditioned dynamic message propagation for monocular 3d object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 454–463.
- [59] G. Brazil, G. Pons-Moll, X. Liu, and B. Schiele, “Kinematic 3d object detection in monocular video,” in *European Conference on Computer Vision*. Springer, 2020, pp. 135–152.
- [60] P. Li, X. Chen, and S. Shen, “Stereo r-cnn based 3d object detection for autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7644–7652.
- [61] A. D. Pon, J. Ku, C. Li, and S. L. Waslander, “Object-centric stereo matching for 3d object detection,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8383–8389.
- [62] Y. Liu, L. Wang, and M. Liu, “Yolostereo3d: A step back to 2d for efficient stereo 3d detection,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 018–13 024.
- [63] J. Sun, L. Chen, Y. Xie, S. Zhang, Q. Jiang, X. Zhou, and H. Bao, “Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 548–10 557.
- [64] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.
- [65] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [66] Y. Zhou and O. Tuzel, “Voxelnet: End-to-end learning for point cloud based 3d object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.
- [67] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, “Pointpillars: Fast encoders for object detection from point clouds,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.
- [68] S. Shi, X. Wang, and H. Li, “Pointrcnn: 3d object proposal generation and detection from point cloud,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 770–779.
- [69] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, “Pv-rcnn: Point-voxel feature set abstraction for 3d object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 529–10 538.
- [70] C. R. Qi, O. Litany, K. He, and L. J. Guibas, “Deep hough voting for 3d object detection in point clouds,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9277–9286.
- [71] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, “Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion,” *IEEE Transactions on Multimedia*, 2022.
- [72] A. Mahmoud, J. S. Hu, and S. L. Waslander, “Dense voxel fusion for 3d object detection,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 663–672.
- [73] S. Pang, D. Morris, and H. Radha, “CLOCs: Camera-LiDAR object candidates fusion for 3D object detection,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 386–10 393.
- [74] Z. Wang and K. Jia, “Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1742–1749.
- [75] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, “Frustum pointnets for 3d object detection from rgbd data,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927.
- [76] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, “Joint 3d proposal generation and object detection from view aggregation,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.
- [77] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3d object detection network for autonomous driving,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.
- [78] R. Zhang, H. Qiu, T. Wang, X. Xu, Z. Guo, Y. Qiao, P. Gao, and H. Li, “Monodetr: Depth-aware transformer for monocular 3d object detection,” *arXiv preprint arXiv:2203.13310*, 2022.
- [79] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, “Monodetr: Monocular 3d object detection with depth-aware transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4012–4021.
- [80] M. Liang, B. Yang, S. Wang, and R. Urtasun, “Deep continuous fusion for multi-sensor 3d object detection,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 641–656.
- [81] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, “Multi-task multi-sensor fusion for 3d object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7345–7353.
- [82] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, “3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection,” in *European Conference on Computer Vision*. Springer, 2020, pp. 720–736.
- [83] T. Huang, Z. Liu, X. Chen, and X. Bai, “Epnet: Enhancing point features with image semantics for 3d object detection,” in *European Conference on Computer Vision*. Springer, 2020, pp. 35–52.
- [84] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, “Pointpainting: Sequential fusion for 3d object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4604–4612.
- [85] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.
- [86] S. Chadwick, W. Maddern, and P. Newman, “Distant vehicle detection using radar and vision,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8311–8317.
- [87] 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 *2019 Sensor Data Fusion:*

- Trends, Solutions, Applications (SDF)*. IEEE, 2019, pp. 1–7.
- [88] E. Hayashi, J. Lien, N. Gillian, L. Giusti, D. Weber, J. Yamamoto, L. Bedal, and I. Poupyrev, “Radarnet: Efficient gesture recognition technique utilizing a miniature radar sensor,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.
  - [89] K. Qian, S. Zhu, X. Zhang, and L. E. Li, “Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 444–453.
  - [90] Y.-J. Li, J. Park, M. O’Toole, and K. Kitani, “Modality-agnostic learning for radar-lidar fusion in vehicle detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 918–927.
  - [91] L. Wang, T. Chen, C. Anklam, and B. Goldluecke, “High dimensional frustum pointnet for 3D object detection from camera, LiDAR, and radar,” in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1621–1628.
  - [92] F. Nobis, E. Shafiei, P. Karle, J. Betz, and M. Lienkamp, “Radar voxel fusion for 3D object detection,” *Applied Sciences*, vol. 11, no. 12, p. 5598, 2021.
  - [93] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, and O. Beijbom, “nuScenes: A multimodal dataset for autonomous driving,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
  - [94] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
  - [95] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou *et al.*, “Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9710–9719.
  - [96] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, “Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.
  - [97] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, “Semantic understanding of scenes through the ade20k dataset,” *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.
  - [98] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.
  - [99] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
  - [100] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
  - [101] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” *arXiv preprint arXiv:1511.07122*, 2015.
  - [102] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” *arXiv preprint arXiv:1412.7062*, 2014.
  - [103] ———, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
  - [104] Y. Yuan, X. Chen, and J. Wang, “Object-contextual representations for semantic segmentation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 173–190.
  - [105] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.
  - [106] A. Tao, K. Sapra, and B. Catanzaro, “Hierarchical multi-scale attention for semantic segmentation,” *arXiv preprint arXiv:2005.10821*, 2020.
  - [107] S. Borse, Y. Wang, Y. Zhang, and F. Porikli, “Inverseform: A loss function for structured boundary-aware segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5901–5911.
  - [108] Y. Cai, L. Dai, H. Wang, and Z. Li, “Multi-target pan-class intrinsic relevance driven model for improving semantic segmentation in autonomous driving,” *IEEE Transactions on Image Processing*, vol. 30, pp. 9069–9084, 2021.
  - [109] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, “Axial-deeplab: Stand-alone axial-attention for panoptic segmentation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 108–126.
  - [110] J. Li, A. Raventos, A. Bhargava, T. Tagawa, and A. Gaidon, “Learning to fuse things and stuff,” *arXiv preprint arXiv:1812.01192*, 2018.
  - [111] S. Aich, J. M. U. Vianney, M. A. Islam, and M. K. B. Liu, “Bidirectional attention network for monocular depth estimation,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11746–11752.
  - [112] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, “From big to small: Multi-scale local planar guidance for monocular depth estimation,” *arXiv preprint arXiv:1907.10326*, 2019.
  - [113] M. Ochs, A. Kretz, and R. Mester, “SDNet: Semantically guided depth estimation network,” in *German Conference on Pattern Recognition*. Springer, 2019, pp. 288–302.
  - [114] L. Liebel and M. Körner, “Multidepth: Single-image depth estimation via multi-task regression and classification,” in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1440–1447.
  - [115] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.
  - [116] W. Xu, H. Wang, F. Qi, and C. Lu, “Explicit shape encoding for real-time instance segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5168–5177.
  - [117] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, “Blendmask: Top-down meets bottom-up for instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8573–8581.
  - [118] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, “Solov2: Dynamic and fast instance segmentation,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17721–17732, 2020.
  - [119] J. Liang, N. Homayounfar, W.-C. Ma, Y. Xiong, R. Hu, and R. Urtasun, “Polytransform: Deep polygon transformer for instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9131–9140.
  - [120] C. Tang, H. Chen, X. Li, J. Li, Z. Zhang, and X. Hu, “Look closer to segment better: Boundary patch refinement for instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13926–13935.
  - [121] L.-C. Chen, H. Wang, and S. Qiao, “Scaling wide residual networks for panoptic segmentation,” *arXiv preprint arXiv:2011.11675*, 2020.
  - [122] L.-C. Chen, R. G. Lopes, B. Cheng, M. D. Collins, E. D.

- Cubuk, B. Zoph, H. Adam, and J. Shlens, "Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 695–714.
- [123] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9404–9413.
- [124] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6399–6408.
- [125] Y. Li, H. Zhao, X. Qi, L. Wang, Z. Li, J. Sun, and J. Jia, "Fully convolutional networks for panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 214–223.
- [126] Z. Li, W. Wang, E. Xie, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, and T. Lu, "Panoptic segformer: Delving deeper into panoptic segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1280–1289.
- [127] L. Porzi, S. R. Bulo, and P. Kotschieder, "Improving panoptic segmentation at all scales," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7302–7311.
- [128] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kotschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4990–4999.
- [129] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 11–20.
- [130] W. Zhuo, M. Salzmann, X. He, and M. Liu, "Indoor scene structure analysis for single image depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 614–622.
- [131] Y. Kuznetsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6647–6655.
- [132] J. M. U. Vianney, S. Aich, and B. Liu, "Refinedmpl: Refined monocular pseudolidar for 3d object detection in autonomous driving," *arXiv preprint arXiv:1911.09712*, 2019.
- [133] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5515–5524.
- [134] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5354–5362.
- [135] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 675–684.
- [136] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.
- [137] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.
- [138] Z. Zhang, C. Xu, J. Yang, Y. Tai, and L. Chen, "Deep hierarchical guidance and regularization learning for end-to-end depth estimation," *Pattern Recognition*, vol. 83, pp. 430–442, 2018.
- [139] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5684–5693.
- [140] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.
- [141] Z. Li, Z. Chen, X. Liu, and J. Jiang, "Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation," *arXiv preprint arXiv:2203.14211*, 2022.
- [142] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2758–2766.
- [143] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4161–4170.
- [144] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnn for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.
- [145] W.-C. Ma, S. Wang, R. Hu, Y. Xiong, and R. Urtasun, "Deep rigid instance scene flow," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3614–3622.
- [146] Y. Zou, Z. Luo, and J.-B. Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 36–53.
- [147] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12240–12249.
- [148] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1983–1992.
- [149] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 336–345.
- [150] Y. Hu, W. Zhan, and M. Tomizuka, "Probabilistic prediction of vehicle semantic intention and motion," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 307–313.
- [151] W. Luo, B. Yang, and R. Urtasun, "Fast and furious: Real time end-to-end 3D detection, tracking and motion forecasting with a single convolutional net," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3569–3577.
- [152] G. Guo and S. Zhao, "3D multi-object tracking with adaptive cubature kalman filter for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, pp. 1–1, 2022.
- [153] M. Bahramgiri, S. Nooshabadi, K. Olutomilayo, and D. Fuhrmann, "Automotive radar-based hitch angle tracking technique for trailer backup assistant systems," *IEEE Transactions on Intelligent Vehicles*, pp. 1–1, 2022.
- [154] H. Lv, C. Liu, X. Zhao, C. Xu, Z. Cui, and J. Yang, "Lane marking regression from confidence area detection to field inference," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 47–56, 2021.
- [155] T. Yan, Z. Xu, and S. X. Yang, "Consensus formation control

- for multiple auvsystems using distributed bioinspired sliding mode control,” *IEEE Transactions on Intelligent Vehicles*, pp. 1–1, 2022.
- [156] Y. Zhu, C. Li, J. Tang, and B. Luo, “Quality-aware feature aggregation network for robust RGBT tracking,” *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 121–130, 2021.
- [157] J. E. Hoffmann, H. G. Tosso, M. M. D. Santos, J. F. Justo, A. W. Malik, and A. U. Rahman, “Real-time adaptive object detection and tracking for autonomous vehicles,” *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 3, pp. 450–459, 2021.
- [158] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler, “Online multi-target tracking using recurrent neural networks,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017.
- [159] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, “A survey of motion planning and control techniques for self-driving urban vehicles,” *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 33–55, 2016.
- [160] H. Bast, D. Delling, A. Goldberg, M. Müller-Hannemann, T. Pajor, P. Sanders, D. Wagner, and R. F. Werneck, “Route planning in transportation networks,” in *Algorithm Engineering*. Springer, 2016, pp. 19–80.
- [161] B. Li, Z. Yin, Y. Ouyang, Y. Zhang, X. Zhong, and S. Tang, “Online trajectory replanning for sudden environmental changes during automated parking: A parallel stitching method,” *IEEE Transactions on Intelligent Vehicles*, 2022.
- [162] Y. Huang, H. Ding, Y. Zhang, H. Wang, D. Cao, N. Xu, and C. Hu, “A motion planning and tracking framework for autonomous vehicles based on artificial potential field elaborated resistance network approach,” *IEEE Transactions on Industrial Electronics*, vol. 67, no. 2, pp. 1376–1386, 2019.
- [163] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu, and L. Chen, “Motion planning for autonomous driving: The state of the art and future perspectives,” *IEEE Transactions on Intelligent Vehicles*, pp. 1–21, 2023.
- [164] M. Montemerlo, J. Becker, S. Bhat, H. Dahlkamp, D. Dolgov, S. Etinger, D. Haehnel, T. Hilden, G. Hoffmann, B. Huhnke *et al.*, “Junior: The stanford entry in the urban challenge,” *Journal of Field Robotics*, vol. 25, no. 9, pp. 569–597, 2008.
- [165] G. C. Y. Liu, X. Hu, “Review of end-to-end motion planning for autonomous driving with visual perception,” *Journal of Field Robotics*, vol. 25, no. 9, pp. 569–597, 2008.
- [166] E. Rehder, J. Quehl, and C. Stiller, “Driving like a human: Imitation learning for path planning using convolutional neural networks,” in *International Conference on Robotics and Automation Workshops*, 2017, pp. 1–5.
- [167] S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan, and X. Hu, “Hierarchical interpretable imitation learning for end-to-end autonomous driving,” *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 673–683, 2023.
- [168] X. Zhang, Y. Jiang, Y. Lu, and X. Xu, “A receding-horizon reinforcement learning approach for kinodynamic motion planning of autonomous vehicles,” *IEEE Transactions on Intelligent Vehicles*, 2022.
- [169] B. Li, T. Acarman, Y. Zhang, Y. Ouyang, C. Yaman, Q. Kong, X. Zhong, and X. Peng, “Optimization-based trajectory planning for autonomous parking with irregularly placed obstacles: A lightweight iterative framework,” *IEEE Transactions on Intelligent Transportation Systems*, 2021.
- [170] D. Fassbender, B. C. Heinrich, and H.-J. Wuensche, “Motion planning for autonomous vehicles in highly constrained urban environments,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4708–4713.
- [171] W. Xu, J. Pan, J. Wei, and J. M. Dolan, “Motion planning under uncertainty for on-road autonomous driving,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 2507–2512.
- [172] X. Li, Z. Sun, D. Cao, Z. He, and Q. Zhu, “Real-time trajectory planning for autonomous urban driving: Framework, algorithms, and verifications,” *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 2, pp. 740–753, 2015.
- [173] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee, “Intention-aware online POMDP planning for autonomous driving in a crowd,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 454–460.
- [174] K. Kondak and G. Hommel, “Computation of time optimal movements for autonomous parking of non-holonomic mobile platforms,” in *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, vol. 3. IEEE, 2001, pp. 2698–2703.
- [175] B. Li, Y. Ouyang, L. Li, and Y. Zhang, “Autonomous driving on curvy roads without reliance on frenet frame: A cartesian-based trajectory planning method,” *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [176] H. Fan, F. Zhu, C. Liu, L. Zhang, L. Zhuang, D. Li, W. Zhu, J. Hu, H. Li, and Q. Kong, “Baidu apollo em motion planner,” *arXiv preprint arXiv:1807.08048*, 2018.
- [177] P. Scheffe, T. M. Henneken, M. Kloock, and B. Alrifaei, “Sequential convex programming methods for real-time optimal trajectory planning in autonomous vehicle racing,” *IEEE Transactions on Intelligent Vehicles*, 2022.
- [178] W. Lim, S. Lee, M. Sunwoo, and K. Jo, “Hierarchical trajectory planning of an autonomous car based on the integration of a sampling and an optimization method,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 2, pp. 613–626, 2018.
- [179] C. Miller, C. Pek, and M. Althoff, “Efficient mixed-integer programming for longitudinal and lateral motion planning of autonomous vehicles,” in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1954–1961.
- [180] L. Claussmann, M. Revilloud, and S. Glaser, “Simulated annealing-optimized trajectory planning within non-collision nominal intervals for highway autonomous driving,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5922–5928.
- [181] C. Rösmann, F. Hoffmann, and T. Bertram, “Integrated online trajectory planning and optimization in distinctive topologies,” *Robotics and Autonomous Systems*, vol. 88, pp. 142–153, 2017.
- [182] T. Gu, J. Atwood, C. Dong, J. M. Dolan, and J.-W. Lee, “Tunable and stable real-time trajectory planning for urban autonomous driving,” in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 250–256.
- [183] D. Le, Z. Liu, J. Jin, K. Zhang, and B. Zhang, “Historical improvement optimal motion planning with model predictive trajectory optimization for on-road autonomous vehicle,” in *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society*, vol. 1. IEEE, 2019, pp. 5223–5230.
- [184] J. Park, S. Karumanchi, and K. Iagnemma, “Homotopy-based divide-and-conquer strategy for optimal trajectory planning via mixed-integer programming,” *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1101–1115, 2015.
- [185] J. Ziegler, P. Bender, T. Dang, and C. Stiller, “Trajectory planning for bertha—a local, continuous method,” in *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, 2014, pp. 450–457.
- [186] J. Reeds and L. Shepp, “Optimal paths for a car that goes both forwards and backwards,” in *American Journal of Mathematics*. IEEE, 1990, pp. 367–393.
- [187] W. N. Patten, H. C. Wu, and W. Cai, “Perfect parallel parking via pontryagin’s principle,” *Journal of Dynamic Systems Measurement and Control*, vol. 116, no. 4, pp. 723–728, 1994.
- [188] Y. Kuwata, S. Karaman, J. Teo, E. Frazzoli, J. P. How, and G. Fiore, “Real-time motion planning with applications to autonomous urban driving,” *IEEE Transactions on Control Systems Technology*, vol. 17, no. 5, pp. 1105–1118, 2009.

- [189] M. Rufli and R. Siegwart, "On the design of deformable input-state-lattice graphs," in *IEEE International Conference on Robotics & Automation*, 2010.
- [190] K. Bergman, O. Ljungqvist, and D. Axehill, "Improved path planning by tightly combining lattice-based path planning and optimal control," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 57–66, 2020.
- [191] M. McNaughton, C. Urmon, J. M. Dolan, and J.-W. Lee, "Motion planning for autonomous driving with a conformal spatiotemporal lattice," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 4889–4895.
- [192] L. Ma, J. Xue, K. Kawabata, J. Zhu, C. Ma, and N. Zheng, "Efficient sampling-based motion planning for on-road autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1961–1976, 2015.
- [193] C. Pek and M. Althoff, "Fail-safe motion planning for online verification of autonomous vehicles using convex optimization," *IEEE Transactions on Robotics*, vol. 37, no. 3, pp. 798–814, 2020.

## V. BIOGRAPHY SECTION



**Long Chen** (Senior Member, IEEE) received the Ph.D. degree in electrical and electronic engineering from Wuhan University in 2013.

He is currently a Professor with State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is also with the presidents office, Waytous Ltd., Beijing. His research interests include autonomous driving, robotics, and artificial intelligence, where he has contributed more than 100 publications.

Prof. Chen serves as an Associate Editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the IEEE/CAA JOURNAL OF AUTOMATIC SINICA, the IEEE TRANSACTIONS ON INTELLIGENT VEHICLES and the IEEE Technical Committee on Cyber-Physical Systems.



**Siyu Teng** received the M.S. degree in computer science and engineering from Jilin University, Changchun, China, in 2021. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China, and also with the Department of Computer Science, Hong Kong Baptist University, Hong Kong.

His main interests are end-to-end autonomous driving and interpretable deep learning.



and trajectory planning.

Dr. Li was a recipient of the International Federation of Automatic Control (IFAC) 2014–2016 Best Journal Paper Prize. He is an Associate Editor of IEEE TRANSACTIONS ON INTELLIGENT VEHICLES.



**Xiaoxiang Na** received the B.Sc. and M.Sc. degrees in automotive engineering from the College of Automotive Engineering, Jilin University, China, in 2007 and 2009, respectively. He received the Ph.D. degree in driver-vehicle dynamics from the Department of Engineering, University of Cambridge, U.K. in 2014.

He is currently a University Assistant Professor in Applied Mechanics at the Department of Engineering, University of Cambridge. From 2014 to 2023, he worked as a Research Associate and later Senior Research Associate at the Centre for Sustainable Road Freight (SRF) at the University of Cambridge. His main research interests include driver-vehicle dynamics, in-service monitoring of heavy goods vehicle (HGV) operations, and evaluation and modelling of energy performance of HGVs.



**Yuchen Li** received the B.E. degree in software engineering from the University of Science and Technology Beijing, Beijing, China, in 2016, and the M.E. degree in software engineering from Beihang University, Beijing, in 2020. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China, and also with the Department of Computer Science, Hong Kong Baptist University, Hong Kong.

His research interests cover computer vision, 3-D object detection, and autonomous driving.



**Zixuan Li** received the B.E. degree from Anhui University in 2017, and the M.E. degree from the University of Chinese Academy of Sciences in 2021. He is a intern in the presidents office, Waytous Ltd., Beijing.

His research interest cover computer vision ,communication engineering and autonomous driving.



**Jinjun Wang** received the B.E. and M.E. degrees from Huazhong University of Science and Technology, China, in 2000 and 2003, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2006.

From 2006 to 2009, he was with NEC Laboratories America, Inc., as a Research Scientist, and from 2010 to 2013, he was with Epson Research and Development, Inc., as a Senior Research Scientist. He is currently a Professor with Xi'an Jiaotong University. His research interests include pattern classification, image/video enhancement and editing, content-based image/video annotation and retrieval, and semantic event detection.



**Fei-Yue Wang** (Fellow, IEEE) received the Ph.D. degree in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

He is currently a Professor and the Director of the State Key Laboratory of Intelligent Control and Management of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Prof. Wang was the Founding Editor-in-Chief of the INTERNATIONAL JOURNAL OF INTELLIGENT CONTROL AND SYSTEMS from 1995 to 2000, the Series on Intelligent Control and Intelligent Automation from 1996 to 2004, and the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS. He was the Editor-in-Chief of the IEEE INTELLIGENT SYSTEMS from 2009 to 2011 and the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS. He is the Editor-in-Chief of the IEEE/CAA JOURNAL OF AUTOMATIC SINICA. He is a member of Sigma Xi and an Elected Fellow of INCOSE, IFAC, ASME, and AAAS. He was the President of the IEEE Intelligent Transportation Systems Society from 2005 to 2007, the Chinese Association for Science and Technology, USA, in 2005, and the American Zhu Kezhen Education Foundation from 2007 to 2008. He is currently the Vice President and the Secretary General of the Chinese Association of Automation.



**Dongpu Cao** (Senior Member, IEEE) received the Ph.D. degree in mechanical engineering from Concordia University, Montreal, QC, Canada, in 2008.

He is a Professor with the school of mechanical engineering, Tsinghua University, Beijing, China. He has contributed more than 200 papers and three books. His current research focuses on driver cognition, automated driving, and cognitive autonomous driving.

Dr. Cao received the SAE Arch T. Colwell Merit Award in 2012, and three Best Paper Awards from the ASME and IEEE conferences. Dr. Cao serves as an Associate Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE/ASME TRANSACTIONS ON MECHATRONICS, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE/CAA JOURNAL OF AUTOMATICA SINICA and ASME JOURNAL OF DYNAMIC SYSTEMS, MEASUREMENT AND CONTROL. He was a Guest Editor for VEHICLE SYSTEM DYNAMICS and IEEE TRANSACTIONS ON SMC: SYSTEMS. He serves on the SAE Vehicle Dynamics Standards Committee and acts as the Co-Chair of IEEE ITSS Technical Committee on Cooperative Driving.



**Nanning Zheng** (Fellow, IEEE) graduated from the Department of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China, in 1975. He received the M.S. degree in information and control engineering from Xi'an Jiaotong University in 1981 and the Ph.D. degree in electrical engineering from Keio University, Yokohama, Japan, in 1985.

He joined Xi'an Jiaotong University in 1975, where he is currently a Professor and the Director of the Institute of Artificial Intelligence and Robotics. His research interests include computer vision, pattern recognition and image processing, and hardware implementation of intelligent systems.

Dr. Zheng became a member of the Chinese Academy of Engineering in 1999, and he is the Chinese Representative on the Governing Board of the International Association for Pattern Recognition.