A Research on Advanced Technology of Target Detection in Unmanned Driving

Bosi Wang¹, Renjie Wang², Bin Tang², Linfeng Cai², Nianbing Zhong¹, Mingfu Zhao¹, Zourong Long²

¹School of Automation, Chongqing University, Chongqing, China
²Intelligent optical fiber sensing technology Chongqing University Engineering Research Center, Chongqing University of Technology, Chongqing, China

*Corresponding author: Zourong Long(longzourong@cqut.edu.cn)

Abstract: Unmanned driving leads the development of smart cities and safe transportation. It relies on a large amount of complex data generated during driving. This paper reviews the state-of-the-art research status of data collection and analysis techniques. We divide the data required for unmanned driving into two types: traffic scene data and driving behavior data. Firstly, by fully considering the different needs of enterprises, universities and related institutions, we have fully studied the algorithms proposed by scholars all over the world. Then we introduce mainstream 2D and 3D target detection algorithms based on RGBD and laser point clouds. Finally, the article introduces some of the most widely used urban road data sets.

1. Introduction

Unmanned driving technology can significantly improve road traffic efficiency, reduce the incidence of traffic accidents and improve energy efficiency. It has become a research hotspot in the automotive industry in recent years [1,2]. However, the entire process of unmanned driving technology development is inseparable from the support of data, so a clearly labeled data set can provide a large amount of high-quality data for algorithm development and training. In addition, a rich variety of scene libraries can make simulation verification more realistic, thereby reducing the development cost of unmanned driving technology and shortening the development cycle.

The unmanned driving database mainly consists of three parts: data collection, data processing and data storage. First of all, data collection is based on different mission objectives, equipped with appropriate sensor combinations on different vehicles to obtain the required data stream. According to the different equipment used, data collection can generally be divided into two types: the existing collection equipment of the original vehicle (driving recorder, car GPS, etc.) and special equipment for installation. Then data processing refers to the enhancement and classification of the original data through the computer to generate new usable data. New data can be combined, archived and classified according to semantics, regulations and experience, which is helpful for developers to use. Finally, the storage of data refers to the rapid storage, retrieval and recall of massive data through distributed management and other methods.

The driverless database is divided into two types: traffic scene database and driving behavior database. The former focuses on the surrounding environment when the vehicle is driving. The latter focuses on the operation of the driver and the behavior of the vehicle itself. The two types merge with each other to form all the driving actions of the vehicle.
2. Data collection and analysis of traffic scenarios
Traffic scene refers to the scene where the vehicle is in road traffic. It is the basis for the development of unmanned driving algorithms and the specific context for the application of the algorithm. Typical traffic scenes mainly come from policies and regulations, driving experience, driving environment, etc.

Zhao et al. [3] of the University of Michigan divided the natural driving data into six key driving scenarios: free driving, car following, lane change, front vehicle insertion, front crossing and side crossing. They established the TrafficNet scene library, however its classification method did not take into account other scene information such as urban and rural environment, traffic signals and traffic conditions. Liu et al. [4] of the China Automotive Technology Research Center proposed a classification method based on the four elements of the vehicle, other traffic participants, road traffic conditions and urban environmental conditions.

2.1 Data collection of traffic scenarios
Data collection of traffic scenes refers to the use of a certain sensor or combination of sensors, which collects the traffic information of the scene where the vehicle is in the form of data for subsequent use. The development of traditional vehicles has little research on the data collection and reconstruction of the scene itself. In recent years, with the continuous development of unmanned driving technology, the executor of driving tasks has changed from people to vehicles. Engineers must consider the vehicle's ability to process different scenes, so data collection technology for traffic scenes is getting more and more attention.

2.1.1 Early traffic scene data collection system
Early traffic scene data collection systems mostly used cameras and Global Navigation Satellite System (GNSS) to record and reproduce scene data. They are mounted on road vehicle platforms. Normally, the recorded scene data is relatively single, which is generally the reconstruction of road scenes. These data hardly contain traffic information such as other traffic participants and urban and rural environments.

Cornelis et al. [5] of ETH Zurich used the data of the video stream to reconstruct the city's 3D scene and provide the city scene data for vehicle navigation. He et al. [6] of Queensland University of Technology combined the stereo images taken by the vehicle-mounted camera with the positioning and map data provided by GNSS to provide raw data for subsequent processing. Zhang of Wuhan University and Lu of National University of Defense Technology [7] used the driving recorder as a data acquisition device and used the video data provided by them to reconstruct the road scene.

2.1.2 Current traffic scene data collection system
With the continuous development of unmanned driving technology, its requirements for scene information are getting higher and higher. The current traffic scene data acquisition system generally uses a combination of multiple sensors such as lidar, millimeter-wave radar, binocular cameras, wide-angle cameras, ultrasonic sensors, and GNSS. They are mounted on a variety of different platforms such as drive test equipment and aircraft. The recorded scene data is relatively rich, including roads, urban and rural environments, and other traffic participants. The data types cover both point clouds and images.

Koppanyi et al. [8] of Ohio State University used vehicles equipped with lidar, GNSS and cameras with different resolutions to collect data of typical traffic scenes. They also developed a simple interface to export these highly redundant data. Correspondingly, China Automotive Technology Research Center [4] has also built a multi-sensor fusion data acquisition platform on a variety of models, which is equipped with monocular vision, binocular vision, lidar, millimeter wave radar, etc.

Some institutions and scholars have produced related data sets to cope with the increased data demand, such as CityScapes data set [9], Oxford RobotCar data set [10] and CARLA data set [11]. These data sets use a large number of sensors to collect different types of scenes in different cycles. Take the Oxford RobotCar dataset [10] as an example. It focuses on the temporal and spatial characteristics of traffic scenes, using 4 cameras, 3 lidars and GPS to perform data on the same continuous road under different weather, pedestrian and road traffic conditions. collection. Its collection
time is as long as one year, and the collection mileage reaches 1010km. It provides the basis for long-distance autonomous driving tests. However, the above-mentioned data sets collect all foreign traffic scenes, which are still different from the domestic road traffic conditions. There are currently few open data sets similar to them in China.

In addition to drive test equipment, aircraft-based data acquisition systems have also begun to be applied. Apeltauer et al. [12] used drones to collect aerial images of intersections and extracted the trajectories of vehicles from them. Chen et al. [13] used aerial video collected by aircraft to model traffic flow and achieved good results.

In order to improve safety and efficiency, Khan et al. [14] proposed a general workflow for collecting traffic scene data using drones and other aircraft. Zhang et al. [15] of Nanjing University of Science and Technology used high-resolution aerial data to reconstruct road scenes. Zhao et al. [16] of Xidian University also use real-time video images collected by drones as scene data for analysis and processing. Compared with drive test equipment, the data acquisition system with aircraft as the carrying platform has the advantages of wide field of view, strong flexibility, and low perspective distortion. With the widespread application of UAVs, the cost of aircraft platforms has been further reduced. However, the types of data collected at present are still mainly images, which are relatively single compared with drive test equipment.

In addition, scene data can not only be collected from the real environment, but also generated and acquired through data enhancement technology. Data enhancement technology obtains new data by performing appropriate translation, rotation, jitter and interpolation operations on the original data. For example: Zhang et al. [17] of Harbin Institute of Technology expanded the image data of autonomous driving by generating a confrontation network. Wu et al. [18] of Hunan University used methods such as masking and changing the illumination to expand the traffic sign database.

Data collection technology for traffic scenes has developed rapidly in the past 5 years. Early data acquisition systems used a single type of sensor. The platform used is mainly drive test equipment and there are fewer types of collection scenarios. With the continuous development of technology, existing data acquisition systems use different sensor combinations and are mounted on different platforms such as aircraft and vehicles. At the same time, the original data can be effectively expanded with the help of data enhancement technology to finally achieve the purpose of enriching the data set.

2.2 Data analysis of traffic scenarios

Data analysis of a traffic scene refers to the identification, classification and labeling of elements in the scene from the collected scene data through a certain method. Engineers proposed a series of technical methods to realize the understanding of traffic scenes. The technology of identifying, classifying and labeling scene elements developed earlier, while the technology of scene understanding has only developed rapidly in recent years.

2.2.1 Recognition, classification and labeling of scene elements

Early scene element recognition methods, namely traditional computer vision target detection methods, were mainly based on the relevant knowledge of digital image processing [19]. In recent years, with the rise of machine learning, recognition, classification and labeling methods have combined traditional theoretical methods with machine learning techniques to improve development efficiency and algorithm performance. The typical algorithms of the two methods are shown in Fig 1.
Fig 1 Typical scene recognition and classification algorithm

The recognition and classification of scene elements began with the introduction of computer vision theory. In the 1980s, Professor Marr of the Massachusetts Institute of Technology and others [20] first proposed the theory of human visual computing, which laid a theoretical foundation for traditional computer vision detection methods. Subsequently, Mardia et al. [21] of the University of Leeds proposed an image segmentation algorithm based on the threshold method, which classifies pixels by setting feature thresholds. The clustering method proposed by Hall et al. [22] of the University of South Florida realizes the extraction and recognition of image features by clustering pixels. The research of traditional computer vision detection methods has been carried out earlier, and there are many related studies and mature theoretical foundations. But traditional methods can only detect and recognize digital images. Not only can it not process point clouds and other types of data, but it also has shortcomings such as low accuracy and heavy workload. Today, machine learning-based recognition methods are mainly used, which follow part of the theoretical framework of traditional methods, such as Scale Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG). Some algorithms use neural networks and other techniques for feature extraction. This method can not only process various types of data including point clouds, but its accuracy rate also exceeds that of traditional detection methods. Therefore, machine learning technology is widely used in unmanned driving recognition algorithms.

In the identification and classification of other traffic participants, Cornelis et al. [5] of ETH Zurich adopted a vehicle detection method based on Implicit Shape Mode (ISM). It has strong robustness, and can detect the specific position of the vehicle under conditions of blur, partial occlusion, and sharp changes in illumination. Zhang et al. of Wuhan University and Lu et al. of National University of Defense Technology [7] proposed Haarlike features based on the AdaBoost classifier, which can complete the detection and recognition of the front and rear surfaces of the vehicle and part of the traffic guardrail in a few seconds. Fu et al. [23] of Beijing University of Posts and Telecommunications designed a classification method based on hierarchical support vector machines (SVM) to realize vehicle detection and recognition in congested traffic scenarios. Yang et al. [24] of Huazhong University of Science and Technology considered the complexity and variability of traffic scenes and proposed a scene-adaptive vehicle detection and recognition algorithm based on convolutional neural networks.

In the recognition and classification of road scenes, He et al. [6] of Queensland University of Technology designed a method to generate accurate and dense 3D road scene semantic maps on a large scale to meet the huge demand of unmanned driving for semantic maps. It uses a car camera to take stereoscopic pictures. The algorithm uses data-driven and non-parametric methods to parse the scene and achieve semantic segmentation. Sun et al. [25] of the University of Waterloo in Canada considered
the high cost of high-precision semantic map and related data collection, and proposed an image-to-map data collection and annotation framework based on proximity. The algorithm uses neural networks and the available tags in the online database for learning, and realizes the automatic labeling of road scenes in different weather. In terms of road alignment feature extraction, Gao et al. [26] of Jilin University proposed a method for estimating road curvature parameters based on the radar information of the preceding vehicle. Based on the YOLOv3 neural network framework, Wu et al. [18] of Hunan University proposed a two-stage detection method for target location and classification recognition, which realized the real-time detection of small objects (such as traffic signs, etc.) in the road scene.

Since the on-board processor may not have enough computing power, some tasks need to be uploaded to a remote server for processing, but the uploaded compressed data is time-consuming during reconstruction. In response to the above problems, Kwan et al. [27] of the Massachusetts Institute of Technology proposed a Pixel Coded Exposure (PCE) method, which combined with deep learning to achieve real-time vehicle detection and classification.

There are multiple types of targets in the actual traffic scene, which will interfere with the recognition algorithm. In response to this problem, Yao et al. [28] of Northwestern Polytechnical University proposed a method to automatically detect and classify multiple targets such as roads, cars, and pedestrians with binocular cameras under the same framework. Guo et al. [29] of Nanjing University established an autonomous driving scene database in accordance with three modules: data verification and submission, data labeling and data statistics. They realize target detection and semantic segmentation of various data such as images and point clouds based on the YOLO neural network algorithm. Zhong et al. [30] of the University of Science and Technology of China converted the lidar point cloud into a range image. They segmented it semantically through a two-dimensional convolutional neural network, and back-projected the obtained semantic tags onto the point cloud. The results show that the method has high accuracy.

In addition, in view of the changes of various traffic elements over time and the time-varying viewpoints caused by shooting videos on moving vehicles, Dwivedi et al. of Columbia University and Narayanan et al. of Honda Research Institute [31] proposed a new classification method. This method has temporal and spatial consistency, which can dynamically identify road conditions. It uses the spatio-temporal characteristics of the data set for semantic segmentation and scene classification. The test results show that the motion of the scene elements captured by the model can better describe the actual scene compared to the static images of the scene elements directly.

### 2.2.2 Understanding of traffic scenarios

The understanding of traffic scenes is based on semantic segmentation and element recognition of traffic scenes. Some scholars believe that the task of understanding the scene is to realize the semantic understanding of the entire scene area by analyzing and predicting the interactive behavior and movement behavior of each element in the scene [32]. With the rise of statistical learning methods represented by neural networks, scene understanding has achieved rapid development in recent years.

Roth et al. [33] of Darmstadt University of Technology conducted earlier research on traffic scene understanding. They use a three-dimensional probability model to understand the interaction between objects in the scene. Jin et al. [34] of Northeastern University established a model that can identify and label the scene area based on the conditional random field theory.

Geiger et al. [35] of the Karlsruhe Institute of Technology proposed a probabilistic generation model for multi-target traffic scene understanding based on mobile platforms. The model uses only visual data for analysis and understanding. It has achieved good results in inferring the layout of the scene.

The environment model output by basic perception modules such as element recognition will focus on expressing specific perception tasks, resulting in different patterns but redundant information. Oeljeklaus et al. [36] constructed a multi-task environment model based on convolutional neural networks, and provided an effective method for generating complementary representations in an integrated manner.

In order to enhance the generalization ability of the scene understanding model, Di et al. [37] of
Beijing University of Posts and Telecommunications established a transfer learning method based on dense correspondence. They conducted experiments in different weather and lighting environments.

In summary, with the continuous breakthroughs of neural network and other technologies in related fields, data analysis technology for traffic scenes has made great progress. The research on the recognition and classification of scene elements has matured day by day, but the development of scene understanding technology is still at a relatively preliminary stage. At present, the task division between traffic element recognition and traffic scene understanding is not completely clear. The semantic classification standards of scenes are not completely uniform. Solving the above problems is of great significance to the establishment of the scene library and the implementation of unmanned driving.

3. Target recognition algorithm for traffic scenes
Object detection is a traditional task in the field of computer vision. Different from image recognition, target detection not only recognizes the type of object on the image, but also gives the position of the object in the form of a minimum bounding box. Normally, RGB images are used for target detection, and then divided into 2D and 3D according to the different output results. The method of outputting the object category and the smallest bounding frame on the image is called 2D target detection. The detection of output object category, length, width, height, rotation angle and other information in three-dimensional space is called 3D target detection.

With the emergence of neural network-based recognition algorithms such as Faster-RCNN, SSD, and Yolo, the accuracy of 2D target detection has exceeded the human eye. However, for the automatic driving of vehicles, it is also necessary to obtain data such as the three-dimensional size and rotation angle of the target object, which cannot be achieved by 2D target detection. Sensors such as lidar, millimeter-wave radar, ultrasound, and RGB cameras provide necessary data for unmanned vehicles, but different types of sensors have advantages and disadvantages. A single sensor cannot meet the requirements of unmanned driving functionality and safety. For example, compared with the RGB camera, the lidar has a poorer resolution, but its ranging ability and environmental adaptability are stronger. The RGB camera is greatly affected by bad weather, but it is better than lidar and millimeter wave radar to detect target speed. Ultrasonic has a better effect on distance detection of close targets and cannot perform long-distance ranging. Therefore, sensor fusion can significantly improve the redundancy and fault tolerance of the system, and improve the accuracy of the 3D target detection algorithm in unmanned driving, thereby ensuring the speed and correctness of decision-making.

3.1 3D target detection based on RGB-D image
In the target detection of 2D images, frameworks such as R-CNN and YoLo proposed by the academic community have been put into commercial use. However, in driverless application scenarios, 2D target detection is still not enough to describe the 3D real-world scene.

3D target detection can get a 3D regression box, which contains information including RGB images and corresponding depth information maps. The depth map is similar to a grayscale image. Each pixel value is the actual distance of the sensor from the object. The RGB image and the depth map are registered, so there is a one-to-one correspondence between pixels. In the environment of unmanned vehicles, detection targets include small objects, occlusions, shadows and other non-ideal situations. Therefore, a large number of scholars have made corresponding improvements to the 2D algorithm, as shown in Table 1.

In 2014, GUPTA et al. [38] added depth information to detect the contour of the image on the basis of the 2D target detection framework R-CNN, and generated a proposal of 2.5D (including the parallax, height, and tilt angle of each pixel). They used Depth CNN to learn features on DepthMap, combined with RGB CNN to learn 2D image features and used SVM for classification. The algorithm realizes target classification and detection with in-depth information. In 2015, CHEN et al. [39] faced unmanned 3D target detection scenes. They conducted a detailed discussion on the method of using the energy equation to obtain the 3D regression frame based on the Fast-RCNN framework. However, there is still no practical and effective method for how to effectively detect the occluded target. In 2016, SONG et
al. [40] proposed a 3D Region Proposal Network based on the Faster R-CNN architecture, which takes the 3D volume scene in the RGB-D image as input and outputs the 3D object bounding box to generate the 3D bounding box of the object. The experimental results show that the algorithm can provide real-world dimensions in the entire range of the object, regardless of truncation or occlusion. In 2017, XU et al. [41] changed the network input to 2.5D (extracting appropriate expressions on the basis of RGB-D) in order to improve the speed of the target detection algorithm, and obtained certain results. Based on the SSD framework, KEHL et al. [42] used the method of synthesizing data and decomposing the pose space of the model to improve the detection rate. However, if the target is occluded, the accuracy of the recognition frame will be greatly reduced. In 2019, Wang et al. [43] proposed the Dense Fusion network architecture. The algorithm merges the depth information of each pixel with the image information, and infers the local fine appearance and geometric spatial information of the target, which is used to deal with the situation of heavy occlusion. By integrating iterative fine-tuning in the neural network, the real-time processing speed of the model is improved.

### 3.2 3D target detection based on laser point cloud

Different from the structural information contained in the image, the laser point cloud has the characteristics of disorder, sparseness, and limited amount of information. In order to make more effective use of the laser point cloud information, 2D, 3D and cut processing schemes are generally used.

#### 3.2.1 Laser point cloud 2D processing

Projecting point cloud data onto a two-dimensional plane is one of the commonly used processing methods. This method does not directly process the three-dimensional point cloud data, but first projects the point cloud to some specific perspectives and then processes it, such as the front view perspective and the bird's-eye view perspective. As shown in Figure 2, VeloFCN [44] projects the coordinates (x, y, z) onto the cylindrical surface to obtain the coordinates (r, c) to generate a dense front view. The bird's-eye view uses multi-channel input to divide the point cloud data according to height. Generate a 2D bird's-eye view projection for each slice.

![Height Maps](image1)
![Density](image2)
![Intensity](image3)

**Fig 2. Aerial view and front view of laser point cloud**

In 3D object detection, the bird's-eye view has three advantages compared with the front view. First, the object retains its physical size when projected to the bird's-eye view, so the size change is small. Second, the objects in the bird's-eye view occupy different spaces, avoiding the problem of occlusion. Third, in a road scene, since objects are usually located on the ground plane and have small changes in
vertical positions, the bird's-eye view position is more important to obtain an accurate 3D bounding box.

3.2.2 Laser point cloud 3D processing
The point cloud has strong arrangement invariance, so the neural network that directly calculates the point cloud provides a unified architecture for the application of target classification, component segmentation to scene semantic analysis. QI et al. [45] proposed a PointNet network that can directly process laser point cloud data. The algorithm extracts laser point cloud data features through MaxPooling symmetric function, and solves the problem of disorder of laser point cloud data. The algorithm realizes the data alignment of the laser point cloud data by training a small neural network, and ensures the invariance of the laser point cloud data in the process of realizing rotation or translation conversion. However, because the PointNet network simply connects all points, it is not effective for multi-classification problems with multiple instances.

In order to solve the above problems, QI et al. [46] proposed an improved network PointNet++, which obtained the deep semantic features of the target by cascading the processing modules of the sampling layer, combination layer, and feature extraction layer. The algorithm uses two combination strategies of multi-scale combination and multi-resolution combination to ensure more accurate target feature extraction.

3.2.3 Laser point cloud segmentation processing
The segmentation method of laser point cloud is to divide it into spatially related shapes, that is, the point cloud data is voxel meshed by means of 3D convolution. Apple proposed VoxelNet on the basis of PointNet++ [47]. Firstly, VoxelNet divides the 3D point cloud into a certain number of voxels. Then the data is randomly sampled and normalized. Finally, the algorithm extracts non-empty voxel features locally. In 2019, Wang et al. [48] proposed PointRCNN to perform 3D object detection through the original point cloud to improve the accuracy of the detection framework. The comparison of the pros and cons of each algorithm, as shown in Table 1, is based on the test results of the data set KITTI. Because the point cloud data processing scheme is closely related to the fusion scheme, the test results are uniformly compared.

Table 1 3D target detection network detection accuracy

| Method      | Type of data | Car Easy | Car Medium | Car Hard | Pedestrian Easy | Pedestrian Medium | Pedestrian Hard | Cycling Easy | Cycling Medium | Cycling Hard |
|-------------|--------------|----------|------------|----------|-----------------|-------------------|-----------------|--------------|----------------|--------------|
| VeloFCN     | Point cloud  | 60.30    | 47.50      | 47.50    | -               | -                 | -               | -            | -              | -            |
| VoxelNet    | Point cloud  | 81.97    | 65.46      | 62.85    | 57.86           | 53.42             | 48.87           | 67.17        | 47.65          | 45.11        |
| Pointnet    | Point cloud  | 72.60    | -          | -        | -               | -                 | -               | -            | -              | -            |
| Pointnet++  | Point cloud  | 75.80    | -          | -        | -               | -                 | -               | -            | -              | -            |
| Point R CNN | Point cloud  | 84.32    | 75.42      | 75.42    | 49.43           | 41.78             | 38.63           | 73.93        | 59.60          | 5.59         |

4. Road target detection data set
In the past ten years, several data sets have been created for pedestrian detection, including INRIA, ETH, TUD-Brussels, and Daimler. But because these data sets are too small (INRIA, ETH, TUD-Brussels) or only provide grayscale images (Daimler).

Recently, larger and richer data sets such as Caltech, KITTI and CityPersons have been widely used. Among them, Caltech includes approximately 250,000 frames and 185,000 pedestrian bounding boxes. Since the video is only taken from daytime recordings in one city, the data set still lacks diversity.

The KITTI data set focuses on the field of multi-sensor settings composed of cameras, laser scanners and GPS/IMU positioning systems, providing data for multiple tasks such as stereo matching, optical flow, SLAM, target detection, and 3D estimation. In the pedestrian detection part, the data set is relatively small.

The Cityscapes dataset consists of a large number of different stereoscopic video sequences recorded on the streets of 27 cities in Germany and neighboring countries. Provided high-quality border annotations for 35,000 pedestrians in 5,000 images. In addition, fine pixel-level annotations for 30 visual categories are provided. Fine annotations include instance tags for people and vehicles. However, the
data set does not contain night scenes or background images.

The KAIST dataset is currently the only public dataset that contains nighttime images of pedestrian detection (5 out of 10 records are at night). However, the data is taken from a single city and a single season, which limits the diversity of the data. Moreover, the recording camera is a consumer-grade camera, resulting in a lot of image noise in the data set.

In summary, the current open source datasets differ greatly in the number, types, and environmental factors of obstacles. The 3D target detection algorithm is divided into pure vision and sensor fusion, so the test data sets used also have great similarities and differences. For the convenience of readers' retrieval,

5. Conclusions and prospects
This article summarizes the current research status of data collection and analysis technology for autonomous driving. We classified the detection algorithms based on different data and summarized different algorithms.

First, data collection of traffic scenes. Data collection systems for current traffic scenarios are mainly mounted on platforms such as ground vehicles and aircraft. They collect data through a combination of sensing devices such as lidar, millimeter-wave radar, GNSS, and cameras, and use data augmentation and other technologies to appropriately supplement the available data. However, there are currently fewer scenario data sets established for China's traffic environment, and the advancement of related work will have great significance for the application of driverless technology in China.

Second, there are still a lot of shortcomings in target detection algorithms: (1) In the actual application of unmanned driving, there is a problem of reduced accuracy due to occlusion and surrounding environment chaos. (2) Most detection algorithms are based on R-CNN, Yolo, etc. They still cannot meet the real-time requirements in unmanned driving. (3) The 3D target detection algorithm based on laser point cloud information has the problem of weak semantic segmentation information ability and cannot accurately classify target types.

A complete database system and mature target recognition algorithm are the basis for the realization of unmanned driving. The development and advancement of unmanned driving data collection and analysis technology will promote the establishment of a traffic scene database and driving behavior database, and will play a vital role in the development, optimization and personalized customization of unmanned driving. With the unmanned driving boom that our country has entered, more excellent algorithms will surely lead us to realize true smart transportation.

Acknowledgments
This work was partially supported by National Natural Science Foundation of China (61805029), Innovation research group of universities in Chongqing (No. CXQT21035)

Reference
[1] Paden B, Čáp M, Yong S Z, et al. A survey of motion planning and control techniques for self-driving urban vehicles[J]. IEEE Transactions on intelligent vehicles, 2016, 1(1): 33-55.
[2] Chen Huiyan, Xiong Guangming, Gong Jianwei. Introduction to Driverless Cars[M]. Beijing Institute of Technology Press, 2014.
[3] Zhao D, Guo Y, Jia Y J. Trafficnet: An open naturalistic driving scenario library[C]//2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2017: 1-8.
[4] Liu Sheng. Research and application of data collection for intelligent networked automobile driving scenes[J]. Automobile Aspects, 2018(8):74-75.
[5] Cornelis N, Leibe B, Cornelis K, et al. 3d urban scene modeling integrating recognition and reconstruction[J]. International Journal of Computer Vision, 2008, 78(2): 121-141.
[6] He H, Upcroft B. Nonparametric semantic segmentation for 3d street scenes[C]//2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2013: 3697-3703.
[7] Zhang Y, Li Q, Lu H, et al. Optimized 3d street scene reconstruction from driving recorder
[8] Koppanyi Z, Toth C K. EXPERIENCES WITH ACQUIRING HIGHLY REDUNDANT SPATIAL DATA TO SUPPORT DRIVERLESS VEHICLE TECHNOLOGIES[J]. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences, 2018, 4(2).

[9] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 3213-3223.

[10] Maddern W, Pascoe G, Linegar C, et al. 1 year, 1000 km: The oxford robotcar dataset[J]. The International Journal of Robotics Research, 2017, 36(1): 3-15.

[11] Dosovitskiy A, Ros G, Codevilla F, et al. CARLA: An open urban driving simulator[C]//Conference on robotics learning. PMLR, 2017: 1-16.

[12] Apeltauer J, Babinec A, Herman D, et al. Automatic vehicle trajectory extraction for traffic analysis from aerial video data[J]. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015, 40(3): 9.

[13] Chen X, Li Z, Yang Y, et al. High-resolution vehicle trajectory extraction and denoising from aerial videos[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(5): 3190-3202.

[14] Khan M A, Ectors W, Bellemans T, et al. UAV-based traffic analysis: A universal guiding framework based on literature survey[J]. Transportation research procedia, 2017, 22: 541-550.

[15] Zhang Chudong. Research on road reconstruction technology based on UAV aerial images[D]. Nanjing University of Science and Technology, 2019.

[16] Zhao Kaidi. Design and implementation of vehicle and lane detection system based on UAV [D]. Xidian University.

[17] Zhang Yaolong. Research on key technologies of data platform for autonomous driving [D]. Harbin Institute of Technology, 2019.

[18] Wu Y, Li Z, Chen Y, et al. Real-time traffic sign detection and classification towards real traffic scene[J]. Multimedia Tools and Applications, 2020, 79(25): 18201-18219.

[19] Wei Zhenxing. Overview of target recognition algorithms [J]. China Equipment Engineering, 2019, 412(01):101-104.

[20] MARR D V. A Computational Investigation into the Human Representation and Processing of Visual Information[M]. San Francisco: W H Freeman and Company, 1982.

[21] Mardia K V, Hainsworth T J. A spatial thresholding method for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 1988, 10(6): 919-927.

[22] Hall L O, Bensaid A M, Clarke L P, et al. A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain[J]. IEEE transactions on neural networks, 1992, 3(5): 672-682.

[23] Fu H, Ma H, Liu Y, et al. A vehicle classification system based on hierarchical multi-SVMs in crowded traffic scenes[J]. Neurocomputing, 2016, 211: 182-190.

[24] Yang D W, Wang H T, Chen Y J, et al. Adaptive Vehicle Detection and Classification Scheme for Urban Traffic Scenes Using Convolutional Neural Network[C]//2019 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW). IEEE, 2019: 1-2.

[25] Sun C, Vianney J M U, Li Y, et al. Proximity based automatic data annotation for autonomous driving[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(2): 395-404.

[26] Zhenhai G, Le Dinh D, Hongyu H, et al. Study lane curvature estimation algorithm based on radar[C]//2017 9th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). IEEE, 2017: 72-76.

[27] Kwan C, Gribben D, Chou B, et al. Real-time and deep learning based vehicle detection and classification using pixel-wise code exposure measurements[J]. Electronics, 2020, 9(6): 1014.

[28] Song Y, Yao J, Ju Y, et al. Automatic Detection and Classification of Road, Car, and Pedestrian Using Binocular Cameras in Traffic Scenes with a Common Framework[J]. Complexity, 2020, 2020.

[29] Guo Jianpeng. Design and Implementation of Autonomous Driving Scene Database Data System
[30] Zhong Zeyu. Semantic understanding method of autonomous driving scene based on lidar point cloud[D]. University of Science and Technology of China, 2020.

[31] Narayanan A, Dwivedi I, Dariush B. Dynamic traffic scene classification with space-time coherence[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 5629-5635.

[32] Xia Z X, Lai W C, Tsao L W, et al. A Human-Like Traffic Scene Understanding System: A Survey[J]. IEEE Industrial Electronics Magazine, 2020, 15(1): 6-15.

[33] Wojek C, Walk S, Roth S, et al. Monocular visual scene understanding: Understanding multi-object traffic scenes[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(4): 882-897.

[34] Jin Xuemei. Research on Traffic Scene Understanding Algorithm Based on Conditional Random Field [D]. Northeastern University, 2015.

[35] Geiger A, Lauer M, Wojek C, et al. 3d traffic scene understanding from movable platforms[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(5): 1012-1025.

[36] Oeljeklaus M, Hoffmann F, Bertram T. A fast multi-task cnn for spatial understanding of traffic scenes[C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 2825-2830.

[37] Di S, Zhang H, Li C G, et al. Cross-domain traffic scene understanding: A dense correspondence-based transfer learning approach[J]. IEEE transactions on intelligent transportation systems, 2017, 19(3): 745-757.

[38] Gupta S, Girshick R, Arbeláez P, et al. Learning rich features from RGB-D images for object detection and segmentation[C]//European conference on computer vision. Springer, Cham, 2014: 345-360.

[39] Chen X, Kundu K, Zhu Y, et al. 3d object proposals for accurate object class detection[C]//Advances in Neural Information Processing Systems. 2015: 424-432.

[40] Song S, Xiao J. Deep sliding shapes for amodal 3d object detection in rgb-d images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 808-816.

[41] Xu D, Ouyang W, Ricci E, et al. Learning cross-modal deep representations for robust pedestrian detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5363-5371.

[42] Kehl W, Manhardt F, Tombari F, et al. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1521-1529.

[43] Wang C, Xu D, Zhu Y, et al. Densfusion: 6d object pose estimation by iterative dense fusion[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3343-3352.

[44] Li B. 3d fully convolutional network for vehicle detection in point cloud[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017: 1513-1518.

[45] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

[46] Qi C R, Yi L, Su H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. arXiv preprint arXiv:1706.02413, 2017.

[47] Zhou Y, Tuzel O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4490-4499.

[48] Wang Z, Jia K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 1742-1749.