A Review on Autonomous Driving: Multi-sensor Object Detection and Semantic Segmentation

Runzhe Jiang
SUSTECH
11912511@mail.sustech.edu.cn

Zhuohang Zhuang
SUSTECH
11912528@mail.sustech.edu.cn

Abstract

Autonomous driving is a smart car technology that uses computer systems to drive without a driver. In a new era of booming new technologies led by artificial intelligence and big data, the trend of intelligent development in the automotive sector is no exception. In order to obtain more robust and accurate perceptual results, autonomous driving usually needs to be equipped with different sensors that complement each other in different operating conditions. Typical sensor modes include: camera, radar, Lidar, high precision map, etc. Multi-modal sensor fusion, which means information complementarity, stability and safety, has long been an important part of autonomous driving perception. However, the insufficient use of information, the noise of the original data and the misalignment between the various sensors (such as time stamps out of sync), and other factors have resulted in the limited fusion performance. This survey comprehensively investigates the advantages and disadvantages of existing sensors, multi-modal autonomous driving perception algorithms, and focuses on object detection and semantic segmentation. In this survey we summarize more than 20 literatures and analyze the existing problems in the current field and provide reference for future research directions.

1. Introduction

Google began testing its Google autonomous driving car in 2010. The goal is to drive vehicles without any human intervention. Figure 1 shows its core architecture. So far 600,000 km has been tested and a legal test plate has been obtained[2]. Other companies such as Tesla, Volvo, BMW are also making deep research into autonomous driving technology, with the recent positioning of advanced assisted driving on highways. Meanwhile, China's Baidu, Didi, Huawei and other companies have also conducted a lot of research on autonomous driving.

Automobile industry is a special industry for it involves the safety of passengers. Any accident in this area is unacceptable, so there are almost harsh requirements for safety and reliability. Therefore, the accuracy and robustness of sensors and algorithms are highly required in the study of autonomous driving. On the other hand,

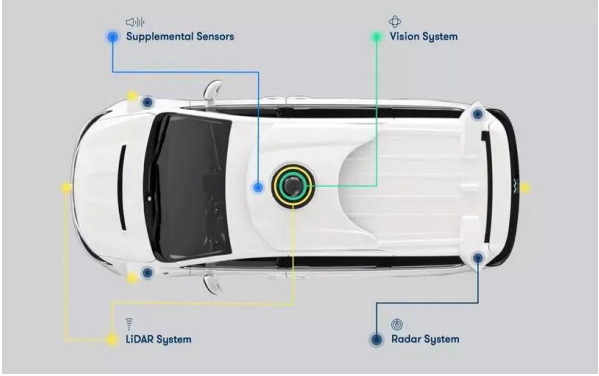

Figure 1

autonomous driving vehicles are products for ordinary consumers, so costs need to be controlled. In the past, it has been difficult to resolve the contradiction between high-precision sensors that make algorithmic results more accurate and expensive (such as Lidar)[3]. Today, the high accuracy brought by deep learning technology promotes the development of autonomous vehicle systems in multiple core areas such as object detection, semantic segmentation , decision making and sensor application. Deep learning technologies, typical of which are convolutional neural network (CNN)[7], are widely used in various kinds of image processing, and are highly applicable to the field of autonomous driving.

The realization of deep learning puts forward higher requirements for sensor technology, which requires multi-sensor fusion technology. In the design of software and hardware architecture of autonomous vehicles, sensors, as the source of data information, are of great importance. However, combined with the application and implementation of deep learning, no matter what kind of sensor has its own advantages and disadvantages. For example, the range information obtained by Lidar is very accurate, but there are some problems such as lack of texture, little feature information and much noise, which is not conducive to the application of deep learning. And the features of the camera are exactly the opposite of the radar. The fusion of sensors like Lidar and camera is of great significance for the accurate perception and cognition of autonomous driving[2].

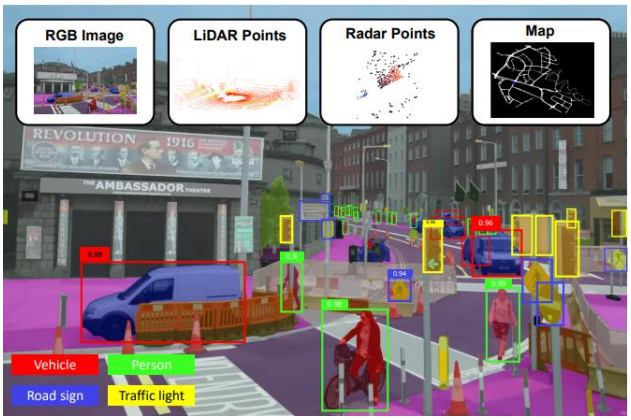

Figure 2 (A complex urban scenario for autonomous driving. The driverless car uses multi-modal signals for perception, such as RGB camera images, LiDAR points, Radar points, and map information. It needs to perceive all relevant traffic participants and objects accurately, robustly,and in real-time. For clarity, only the bounding boxes and classification scores for some objects are drawn in the image. The RGB image is adapted from [4].)

2. Background

2.1. Visual and Thermal Cameras

Images taken by visual and thermal cameras provide detailed textural information about the vehicle's surroundings. The followers of visual solutions believe that humans can become qualified drivers through visual information and brain processing. Camera + deep learning neural network + computer hardware can also achieve similar effects. Visual cameras are very sensitive to light and weather conditions. Thermal imagers, on the other hand, are sensitive to day-night changes because they detect infrared radiation, which is related to an object's heat[10]. By means of infrared optical system that can pass infrared radiation, the infrared radiation of the scene in the field of view is focused on the device that can convert infrared radiation energy into physical quantities that can be easily measured -- infrared detector[10]. The infrared detector then converts the radiation signals of varying strength into corresponding electrical signals. After amplification and video processing, video images can be formed for human eyes to observe. However, neither type of camera provides depth information directly..

2.2. LiDARs

LiDARs: which is also named Laser Radar. It is a radar system which transmits laser beam to detect the characteristic quantity of target such as speed and position. It works by firing a laser beam , which is a detection signal, at a target and then comparing the transmitted signal to the signal that is reflected back from the target. When the beam hits an object, it bounces back and is sensed and recorded by a sensor. Then, the system will calculate the distance of the beam to the object and back. After a series of processing, we can obtain the relevant information of the target, such as its orientation, speed, attitude, distance and shape parameters. And we can then create a 3D map of the area around the vehicle to enable safe navigation of the vehicle. Laser radar accuracy can reach mm level, such as a pedestrian reflection point can reach tens of thousands of points or even more, thus giving clear 3D three-dimensional graphics. Moreover, the laser radar has a long detection range and its angular resolution is several grades higher than millimeter wave radar, easily reaching 0.1°[11]. That is to say, if the power is enough, it can distinguish two small targets with a distance of 3KM for a 5M gap[6]. Flash LiDAR was developed to produce detailed object information similar to camera images. Of course, the FMCW Lidar can provide speed information.

Figure 3

2.3. Radars

Radars(radio detection and ranging) transmits radio waves that bounce off obstacles, measures the signal's travel time, and estimates the object's radial velocity through the Doppler effect. They are highly adaptable to all kinds of light and weather conditions, but it is very challenging to classify objects by radar due to its low resolution. Radar is often used in adaptive cruise control (ACC) and traffic jam assistance systems[11]. In 1999, Distronic (DTR) radar control system was installed on the Mercedes-Benz 220 Series S-Class sedan, providing basic ACC function by controlling the distance between the car and the car in front of it at 40km to 160km per hour[8]. At present, the radar of major international manufacturers has developed to the 5th generation, basically maintaining the

update speed of 2 to 3 years. At the same time, domestic millimeter wave radar manufacturers have also had about 5 years of development, the head company's mass production products have been close to the international advanced level. Although the accuracy and detection range have been greatly improved, but overall its development stage has not yet achieved essential breakthrough.

2.4. Ultrasonics

Ultrasonic radar uses ultrasonic waves to detect obstacles ahead. The ultrasonic generator emits ultrasonic sound waves in a certain direction. When the sound waves are propagated in the air, they meet obstacles and return to the original path. The ultrasonic receiver receives the echo and stops the timing. The distance between the sensor and the obstacle can be calculated according to the speed and time difference of the ultrasonic wave in the air. Ultrasonic radar is an important sensor for parking function due to its low price (a few yuan to 10 yuan) and short detection distance. In addition, the ultrasonic radar (APA) on the side of the car can also act on the lateral driving auxiliary alarm function[6]. Ultrasonic radar theoretically works on all objects that can reflect ultrasonic waves, including solids and liquids, but its ability to detect them is compromised if the target is an angled plane that can reflect sound waves to other angles, or an object such as a sponge or foam that can absorb ultrasonic waves. Therefore, they are usually used in close object detection and low speed scenarios.

2.5. GNSS and HD Maps

GNSS is the abbreviation of "Global Navigation Satellite System". It refers to all satellite navigation systems in general, including global, regional and enhanced, such as the US GPS, Russia's Glonass, Europe's Galileo, China's Beidou navigation Satellite System[4]. Also it related enhanced systems, such as WAAS (Wide Area Augmentation System) in the United States, EGNOS (European Static Navigation Overlapping System) in Europe and MSAS (Multifunctional Transportation Satellite Augmentation System)[4] in Japan, and other navigation satellite systems under construction or to be built in the future. The international GNSS system is a complex composite system with multiple systems, layers and modes. In other words, it is a large system composed of multiple satellite navigation and positioning systems and their enhanced systems. As a result, GNSS positioning accuracy is much higher than civilian GPS, but the error is still several meters. To achieve safe autonomous driving, ensure a high precision GNSS is the key. Originally introduced as a navigation tool for driving assistance in cars, GNSS is now also used in conjunction with high-definition maps for autonomous vehicle path planning and personal vehicle location. At the present stage, GNSS positioning alone is still some distance from fully meeting the position accuracy requirements of automatic driving. According to analysis, autonomous driving on expressways requires a minimum error of 1.5 meters, while autonomous driving in urban areas requires a centimeter-level accuracy. In addition, it is easy to lose signal in navigation when entering tunnels or built-up metropolitan areas, which is why it is difficult to use GNSS alone to locate.

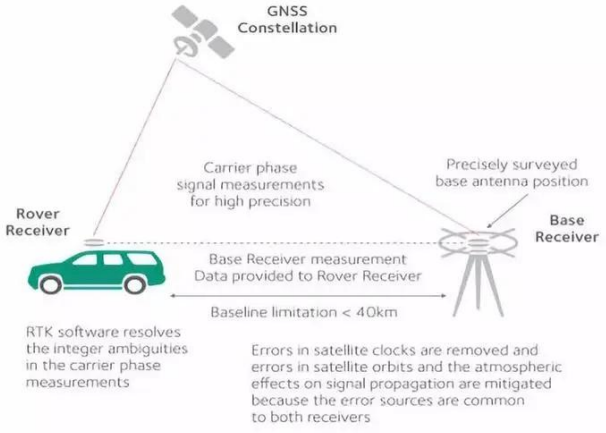

Figure 4

2.6. Inertial Measurement Unit and Odometer

The Inertial Measurement Unit (IMU) and odometer provide internal information about the vehicle. It can be used as an effective supplement when sensor data is missing. The IMU measures three-dimensional linear acceleration and three-dimensional angular velocity, and from this information, the vehicle's attitude (pitch and roll Angle), heading, speed and position change can be calculated. IMUs can be used to fill in the gaps between GNSS signal updates and even dead reckoning if the GNSS and other sensors in the system fail. The key advantage of the IMU is that it works well in all weather and geographical conditions. As an independent data source, it can be used for short-term navigation and to verify information from other sensors, and is not invalidated by weather, lens dirt, radar and lidar signal reflections, or the urban canyon effect. As an independent sensor, IMU is regarded as the sensor that supplements and confirms the data of other sensors, namely the "last sensor", which is used to ensure the safety of the vehicle and stop the vehicle in a controlled manner when other sensors are damaged or disabled. Therefore, we call IMU the anchor of the autonomous driving system[2].

2.7. Object Detection

Object detection is an important task in computer vision,

which develops computational models for a specific class of visual objects (such as humans, animals or cars). If we think of object detection today as a technological aesthetic under deep learning, then go back 20 years and we will witness the "wisdom of the age of cold weapons". Most of the early target detection algorithms were built based on manual features. Due to the lack of effective image representation at the time, people had no choice but to design complex features and utilize various acceleration techniques to pk limited computing resources. There are three main algorithms: Viola Jones detector, HOG detection and DPM algorithm[1].

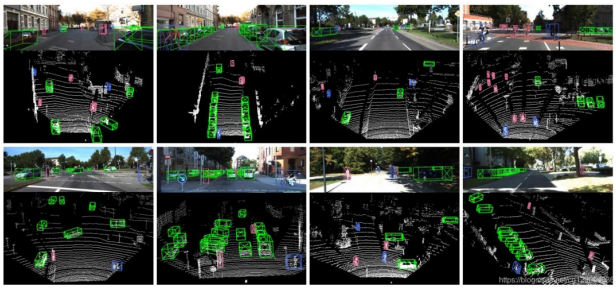

Figure 5

Currently, most of the frameworks used in object detection are based on convolutional neural networks (CNN)[14], which can effectively extract image features. Because of different task requirements, object detection is divided into two categories: object detection algorithm of two stages and object detection algorithm of one stage. The target detection in the Two stages is to first generate a series of candidate boxes as samples by the heuristic candidate region generation algorithm, and then classify samples by convolutional neural network. In one stage, the problem of target frame location is directly transformed into a regression problem without generating candidate boxes. Compared with the two, each has its advantages and disadvantages. The target detection algorithm of the two stages has higher detection accuracy and positioning accuracy, which is suitable for tasks requiring high-precision identification. However, the target detection algorithm of one stage is faster and suitable for real-time recognition tasks[7]. The target detection algorithm of the Two stages is mainly represented by the RCNN series. The main idea is to generate area proposal and separate the target classification detection, so as to obtain more accurate detection effect. There are also SPP-Net algorithm, Fast R-CNN algorithm and Feature Pyramid Networks[13].

Although Faster R-CNN, the peak of the two stages, has made great progress in computing speed, it still cannot meet the requirements of real-time detection. Therefore, researchers are thinking about how to achieve real-time target detection. So a method based on regression has been

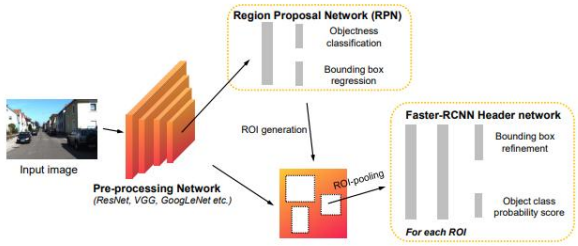

Figure 6 (The Faster R-CNN object detection network. It consists of three parts: a pre-processing network to extract highlevel image features, a Region Proposal Network (RPN) that produces region proposals, and a Faster-RCNN head which fine-tunes each region proposal.)[1]

proposed to directly return the location and type of the target object from the picture. The target detection algorithm of one stage can realize the sharing feature of a complete single training, and the speed is greatly improved under the premise of ensuring a certain accuracy. Two typical algorithms are YOLO and SSD[12].

Target detection can provide automatic driving with traffic sign recognition, pedestrian and vehicle recognition, traffic signal recognition and other basic functions. Currently, vehicles capable of autonomous driving generally rely on ultrasonic sensors in the front and rear sides of the vehicle, millimeter-wave radar in the center of the front, and forward-looking cameras under the rearview mirrors to achieve autonomous driving. Front side and rear side ultrasonic sensors are used to detect close distance obstacles, help automatic parking, etc. Millimeter wave radar is used to detect distant obstacles. The forward-looking camera can complete the detection of lane lines and obstacles on the road surface.

2.8. Semantic Segmentation

In recent years, with the rapid development of deep learning technology, great breakthroughs have been made in many tasks that are difficult to be solved by traditional methods in the field of computer vision. Especially in the field of image semantic segmentation, deep learning technology plays a particularly prominent role. Image semantic segmentation is a fundamental and challenging task in computer vision. Its goal is to assign corresponding semantic labels to each pixel in the image. The result is to divide a given image into several visually meaningful or interesting areas, which is conducive to the subsequent image analysis and visual understanding.

Image segmentation task requires each pixel of the original image to predict its category pixel by pixel. Since it is necessary to predict the category of the object pixel by pixel, this poses great challenges to the deep learning model. For example, some objects are small in size and difficult to identify, and some objects are mostly blocked,

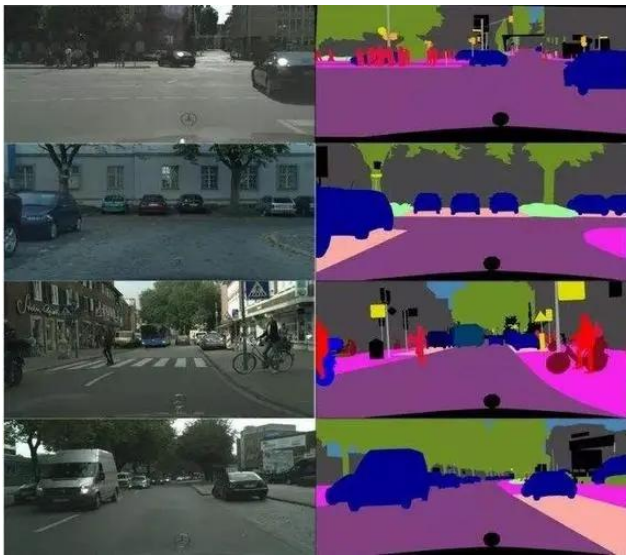

Figure 7

resulting in reduced recognition.

Although semantic segmentation was first introduced to process camera images, many methods have been proposed to segment lidar points. A number of data sets for semantic segmentation have been released such as Cityscape, KITTI, Toronto City, Mapillary Vistas and ApolloScape[11]. These data sets advance deep learning research on semantic segmentation in autonomous driving.

Semantic segmentation can also be classified as two stages and one stage. In two stage pipelines, domain proposals are first generated and then fine-tuned, mainly for instance level segmentation (e.g. R-CNN, SDS, Mask-RCNN[1]). For semantic segmentation, a more common approach is the single-stage pipeline based on the full convolutional network originally proposed by Long et al. In this work, the full join layer in the CNN classifier used to predict classification scores is replaced by a convolutional layer to produce a rough output map. These maps are then up-sampled to dense pixel labels by reverse convolution (i.e. deconvolution). Kendall et al. extended FCN by introducing encod-decoder CNN architecture[13]. The purpose of the encoder is to generate a layered image representation through the CNN backbone (removing the fully connected layer). Instead, the decoder restores these low-dimensional features to their original resolution through a set of up-sampling and convolution layers. The recovered feature map is finally used for pixel label prediction.

3. Datasets

As the methods of multi-sensor object detection and semantic segmentation are mostly based on supervised learning, the quality of datasets becomes quite significant to good perception ability in autonomous driving. Instead of the datasets that only include camera images for other traditional tasks in computer vision field, training neural networks in this case requires large quantities of multi-modal datasets with accurate labels. However, to collect such multi-modal data will not only face difficulties in both hardware and software technology, but also require a great deal of time and money. The following are several aspects of datasets that should be taken into consideration. Each of them can have a dramatic influence on the performance, thus indicating the potential method to improve the performance.

Please use footnotes[iii] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

3.1. Sensing modalities

In terms of sensing modalities, the mainstream methods mainly fuse RGB images taken by vision camera either with thermal images or with 3D point clouds. Therefore, three most common types of data in current datasets are RGB images, thermal images and LiDAR point clouds. For example, KITTI[14][15], a widely used dataset for vision recognition system provides a benchmark developed by their autonomous driving platform including RGB camera images recorded by four high solution stereo cameras and LiDAR point clouds recorded by a laser scanner. An GPS/IMU inertial navigation system is also involved to provide the path and position information of the car[15]. Ha et al.[16] not only made contributions to semantic segmentation by producing a new CNN architecture named Multi-spectral Fusion Networks (MFNet), but also creating a groundbreaking public RGB-thermal image dataset for training in relative methods. KAIST multi-spectral dataset provided both LiDAR cloud points and thermal images with their self-developed multi-sensor platform[17]. For all-day vision of autonomous driving, creating a dataset involving several sensing modalities is quite challenging because many sensors need to be aligned. To solve this problem, they proposed a calibration technique so that their autonomous driving platform can support the use of a co-aligned group of sensors including a stereo vision camera, an RGB-thermal camera, a 3D LiDAR and GPS/IMU sensor groups, thus providing several combinations of modalities.

Figure 8 below demonstrates clearly different sensing modalities available in KAIST multi-spectral dataset: calibrated RGB-stereo image pairs (the 1st column), 3D LiDAR data (the 3rd column), the thermal images co-aligned with the left-view RGB image using the beam-splitter(the 4th column)[18]. What's more, the chaotic bottom image of the 2nd column has shown us the

weakness of RGB-stereo images in providing information at night when light condition is poor. To fulfill unsupervised depth estimation for all-day vision, Kim et al.[18] proposed the multi-task framework named the Multi-spectral Transfer Network (MTN) to generate depth images from a single thermal image that are not sensitive to the light condition. This indicate the significance of multi-modal fusion that it can leverage complementary advantages of various sensing modalities to improve the performance of the deep neural networks in object detection and semantic segmentation.

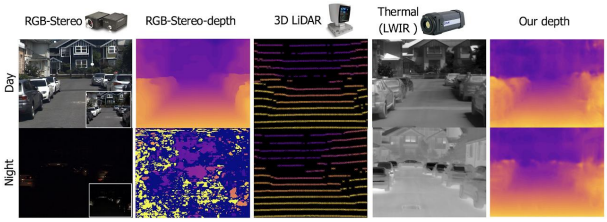

Figure 8[18] Different sensing modalities in KAIST datasets

3.2. Environmental conditions

As an autonomous driving system is likely to confront all types of complex environmental conditions and is expected to deal with them correctly every time to ensure the absolute road security, an ideal multi-modal dataset used for training should cover various environmental conditions as widely as possible, including different weather conditions such as rainstorm, snowstorm, thick fog, different time periods of the day such as early morning, noon and late night. To further improve the accuracy and stability of object detection, the datasets should also be recorded in different locations to enhance its diversity. Nevertheless, many datasets are less efficient for their lack in diversity of environmental conditions. Although the well known dataset in autonomous driving, KITTI[14][15], are widely recognized in this field, it has great limitations as the whole dataset is recorded in a single city, Karlsruhe, Germany. Some other datasets have take the diversity of environmental conditions into consideration. For instance, nearly 20 million images of the Oxford RobotCar Dataset are collected by the Oxford RobotCar platform with its 6 visual cameras, LiDAR, GPS and inertial sensors when it travel along the same route in Oxford twice a week from May 2014 to December 2015[20]. Figure 9[20] shows 9 of over 100 images taken by the robot car at the same location of the route, illustrating great differences between the appearances over a series of conditions. The group of the images vary from overall lighting condition to seasonal variations, all these changes produce a profusion of different combination of environmental conditions to this dataset for training.

Figure 9[20] Some of the images taken by the Oxford RobotCar platform at the same location under different environmental conditions

3.3. Dataset size

Due to the complexity of the multi-modal datasets and the difficulties in creating a new one, datasets for multi-sensor object detection and semantic segmentation are relatively smaller and fewer. The widely used KITTI[14][15] only has nearly 7000 frames for training, while KAIST provides nearly 100K frames and 100K objects in the datasets[17]. Some other datasets include continuous video image frames providing similar information that are less useful for training. We can see that advanced researches in this field are still in great need of large-scale and high-quality datasets.

3.4. Labels

Labels are given in most datasets used for training in object detection and semantic segmentation tasks. Depending on the purposes that the datasets are constructed for, in different datasets, even the same object can be labeled in different ways. For example, KAIST multi-spectral pedestrian dataset focuses on recognizing people, thus only including labels for distinguishable single person and non-distinguishable individuals and cyclists[17]. KITTI provides different labels for pedestrians and the sitting people, it can also distinguish different vehicles like cars, trucks and vans[14][15]. One potential problem is that in most of the datasets, vehicles and people are dominant in labeled objects, which might weaken the ability of the trained deep neural networks to recognize other objects like road signs or trees.

4. Methods

4.1. Processing point clouds

When fusing LiDAR point clouds with RGB images in multi-sensor object detection and semantic segmentation tasks, one of the key points is that besides the visual information provided by RGB images, 3D object detection also needs accurate depth information provided by LiDAR point clouds to estimate the distance between the vehicle and the object. So there are several researches concentrating on a variety of methods to properly represent and process the point clouds, so that the deep neural networks can learn the features as accurate as possible.

4.1.1 3D voxelization

One method is to divide the 3D space into little 3D voxels, where the concept of a voxel in the 3D space can be likened to a 2D pixel of a 2D image, so the 3D point cloud can be assigned to the voxels respectively. The advantage of this method is that voxelization can retain rich spatial information of the 3D object. But there are many remaining empty voxels because the point clouds are usually sparse, making the corresponding 3D-CNN and clustering methods quite time-consuming when processing such data, which is an unacceptable case in real-time autonomous driving. To handle sparse LiDAR point clouds, Zhou et al.[21] proposed VoxelNet, a universal 3D detection end-to-end trainable deep neural network. They introduced a Voxel Feature Encoding (VFE) layer, as shown in the Figure 10, to process the point cloud efficiently. By conducting experiments on widely used KITTI dataset, the VoxelNet has shown outstanding performance among the state-of-the-art methods based on 3D LiDAR point clouds with an inference time of only 225ms.

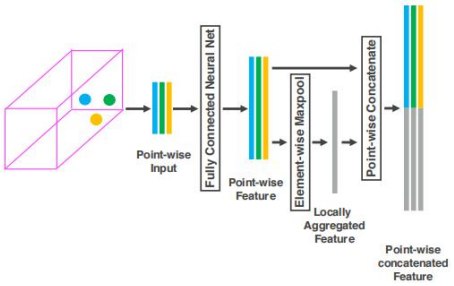

Figure 10[21] Voxel feature encoding(VFE) layer

4.1.2 Direct learning over points

Instead of voxelization, some related work based on 3D LiDAR directly conducts deep learning on point sets. PointNet++ is able to learn individual features from each point in the LiDAR point clouds and use max pooling to aggregate features of a set of points[22]. It leverages the points in neighborhood with grouping at different scales to make the method robust and accurate in detail.

4.1.3 Projection to 2D feature maps

Another way to process the 3D point cloud is to utilize the idea of processing 2D images. That is, we can project the LiDAR point cloud to various 2D feature maps, and then leverage 2D convolutional layers to process them. Figure 4 displays the RGB image and several types of 2D feature maps, such as spherical map, camera-plane map and Bird's Eye View (BEV) map, of a street scene[1]. The spherical map can represent the point cloud in a ideally dense way by projecting the points on a sphere. However, without preprocessing, such spherical map is difficult to align with RGB camera images for multi-modal fusing. Though the camera-plane map can directly be used for fusing, empty pixels exist in the sparse map, thus requiring up-sampling methods to fill these pixels and turn it into a dense map for further processing. The BEV map is another choice that keep the distribution of the objects on the ground plane, including their lengths, widths and distances between each other.

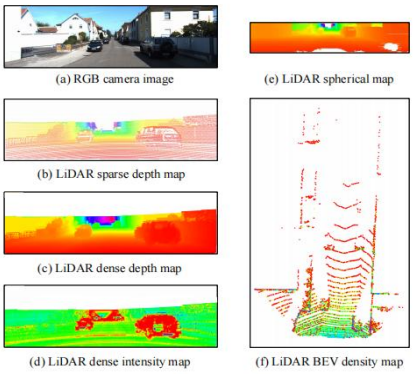

Figure 11[1] Different 2D LiDAR representations

4.2. Processing camera images

Most current methods of multi-sensor object detection and semantic segmentation take RGB images from vision camera or infrared images from thermal camera as one of the sensing modalities. They include rich visual information demonstrating the surroundings. Thermal camera is particularly good at capturing images when lighting condition is limited. However, one problem is that for an object in an camera image, the shape can be distorted and the size can vary on the camera plane, thus affecting object detection. To solve this problem, Roddick et al.[23] develop an Orthographic Feature Transform (OFT) algorithm to project the camera images to a BEV representation that is commonly used for processing a point cloud, which largely preserves the features of the objects.

4.3. Fusing point clouds and camera images

Most of related works fuse LiDAR point clouds and camera images on the network level. That is, they project the point clouds to 2D feature maps and use the convolutional neural networks to process them as we described in Part III - 1 - 3). These methods fuse two modalities by fusing the features obtained from the networks. Some methods conduct clustering directly on 3D point clouds to generate labels for region proposals. Several works also leverage the idea of projection to align features across different sensing modalities. For example, projecting RGB images onto BEV plane when the point cloud is represented as a BEV map.

5. Challenges

5.1. Data diversity

Ideal datasets with great diversity in environmental conditions and object labels with several sensing modalities are needed to improve the accuracy and robustness of multi-sensor object detection and semantic segmentation. So far, available datasets are still small in scale and few in number.

5.2. Data quality

Even the data labeled manually might include some errors in labels. These errors might lower the accuracy of object detection when the labeled data are used for training. What's more, different sensors that are not calibrated accurately can have temporal and spatial deviations, which might produce critical errors in datasets. Further researches are needed for sensor calibration to improve the quality of the datasets and the efficiency of calibration.

5.3. Sensor redundancy

Current works mainly focus on fusing different sensing modalities, but few conduct research on fusing data recorded by several sensors of the same type. Such sensor redundancy can not only collect more information from the surroundings, but also improve safety of autonomous driving.

5.4. Network architecture

Most of the works use convolutional neural networks for real-time perception without considering previous states. In the future, new network architectures that can process time series might be introduced to multi-sensor object detection and semantic segmentation task.

References

[1] Feng D , Haase-Schutz C , Rosenbaum L , et al. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, PP(99):1-20.

[2] Wang Z , Wu Y , Niu Q . Multi-Sensor Fusion in Automated Driving: A Survey[J]. IEEE Access, 2020, 8:2847-2868.

[3] J. Kim, J. Koh, Y. Kim, J. Choi, Y. Hwang, and J. W. Choi, "Robust deep multi-modal learning based on gated information fusion network," arXiv:1807.06233 [cs.CV], 2018

[4] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The ` Mapillary Vistas dataset for semantic understanding of street scenes," in Proc. IEEE Conf. Computer Vision, Oct. 2017, pp. 5000–5009.

[5] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," Sensors, vol. 18, no. 10, p. 3337, 2018.

[6] C. Urmson et al., "Autonomous driving in urban environments: Boss and the urban challenge," J. Field Robotics, vol. 25, no. 8, pp. 425–466, 2008.

[7] B. Xu and Z. Chen, "Multi-level fusion based 3D object detection from monocular images," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2018, pp. 2345–2353.

[8] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik, "Learning rich features ´ from RGB-D images for object detection and segmentation," in Proc. Eur. Conf. Computer Vision. Springer, 2014, pp. 345–360.

[9] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner, "Three decades of driver assistance systems: Review and future perspectives," IEEE Intell. Transp. Syst. Mag., vol. 6, no. 4, pp. 6–22, 2014.

[10] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," IEEE Transactions on Intelligent Transportation Systems, pp. 1–14, 2019.

[11] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," arXiv:1704.06857 [cs.CV], 2017.

[12] L. Liu et al., "Deep learning for generic object detection: A survey," arXiv:1809.02165 [cs.CV], 2018.

[13] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3d lidar using fully convolutional network," in Proc. Robotics: Science and Systems, Jun. 2016.

[14] A. Geiger, P. Lenz and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354-3361, doi: 10.1109/CVPR.2012.6248074.

[15] A. Geiger, P. Lenz, C. Stiller, R. Urtasun. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research. 2013;32(11):1231-1237. doi:10.1177/0278364913491297

[16] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 5108-5115, doi: 10.1109/IROS.2017.8206396.

[17] Y. Choi et al., "KAIST Multi-Spectral Day/Night Data Set for Autonomous and Assisted Driving," in IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 3, pp. 934-948, March 2018, doi: 10.1109/TITS.2018.2791533.

[18] N. Kim, Y. Choi, S. Hwang, I. S. Kweon (2018). Multispectral Transfer Network: Unsupervised Depth Estimation for All-Day Vision. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.12297

[19] The nuScene dataset. nuScene. [Online]. Available: https://www.nuscenes.org/

[20] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000km: The Oxford RobotCar dataset," Int. J. Robotics Research, vol. 36, no. 1, pp. 3–15, 2017.

[21] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3d object detection, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2018.

[22] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in Advances in Neural Information Processing Systems, 2017, pp. 5099–5108.

[23] T. Roddick, A. Kendall, and R. Cipolla, " Orthographic feature transform for monocular 3d object detection, " arXiv:1811.08188 [cs.CV], 2018.