

# Deep Learning for LiDAR Point Clouds in Autonomous Driving: A Review

Ying Li, Lingfei Ma, *Student Member, IEEE*, Zilong Zhong, *Student Member, IEEE*, Fei Liu, Dongpu Cao, *Senior Member, IEEE*, Jonathan Li, *Senior Member, IEEE*, and Michael A. Chapman *Senior Member, IEEE*,

**Abstract**—Recently, the advancement of deep learning in discriminative feature learning from 3D LiDAR data has led to rapid development in the field of autonomous driving. However, automated processing uneven, unstructured, noisy, and massive 3D point clouds is a challenging and tedious task. In this paper, we provide a systematic review of existing compelling deep learning architectures applied in LiDAR point clouds, detailing for specific tasks in autonomous driving such as segmentation, detection, and classification. Although several published research papers focus on specific topics in computer vision for autonomous vehicles, to date, no general survey on deep learning applied in LiDAR point clouds for autonomous vehicles exists. Thus, the goal of this paper is to narrow the gap in this topic. More than 140 key contributions in the recent five years are summarized in this survey, including the milestone 3D deep architectures, the remarkable deep learning applications in 3D semantic segmentation, object detection, and classification; specific datasets, evaluation metrics, and the state of the art performance. Finally, we conclude the remaining challenges and future researches.

**Index Terms**—Autonomous driving, LiDAR, point clouds, object detection, segmentation, classification, deep learning.

## I. INTRODUCTION

**A**CCURATE environment perception and precise localization are crucial requirements for reliable navigation, information decision and safely driving of autonomous vehicles (AVs) in complex dynamic environments[1, 2]. These two tasks need to acquire and process highly-accurate and information-rich data of real-world environments [3]. To obtain such data, multiple sensors such as LiDAR and digital cameras [4] are equipped on AVs or mapping vehicles to collect and extract target context. Traditionally, image data captured by the digital camera, featured with 2D appearance-based representation, low cost, and high efficiency, is the most commonly used data in perception tasks [5]. However, image data lack of 3D geo-referenced information [6]. Thus, the dense, geo-referenced, and accurate 3D point cloud data collected by LiDAR are exploited. Besides, LiDAR is not

Y.Li, L.Ma and J.Li are with the Department of Geography and Environmental Management, University of Waterloo, 200 University Avenue West, Waterloo, N2L 3G1, Canada (e-mail: y2424li@uwaterloo.ca, l53ma@uwaterloo.ca).

Z.Zhong is with School of Data and Computer Science Sun Yat-Sen University, Guangzhou, China, 510006 (email: zlzhong@ieee.org).

F.Liu is with Xilinx Technology Beijing Limited, Beijing, China, 100083 (email: fledliu@xilinx.com).

D.Cao is with Waterloo Cognitive Autonomous Driving Lab, University of Waterloo, N2L 3G1, Canada (e-mail: dongpu.cao@uwaterloo.ca).

J. Li is with the Departments of System Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: junli@uwaterloo.ca).

M. A. Chapman is with the Department of Civil Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: mchapman@ryerson.ca).

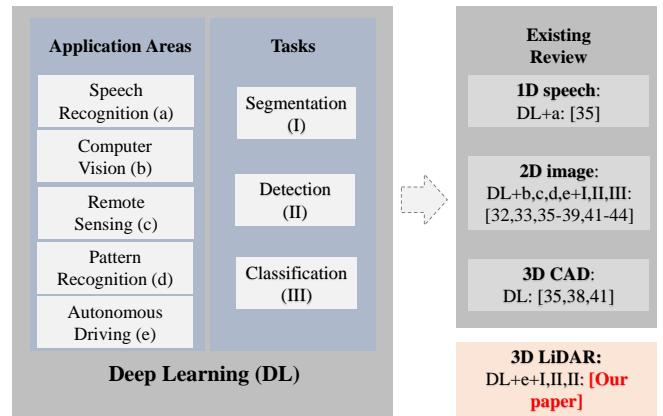


Fig. 1. Existing review paper related to DL and their application with different tasks. We summarize that our paper is the first one to survey the application of LiDAR point clouds in segmentation, detection and classification tasks for autonomous driving using DL techniques

sensitive to the variations of lighting conditions and can work under day and night, even with glare and shadows [7].

The application of LiDAR point clouds for AVs can be described in two aspects: (1) real-time environment perception and processing for scene understanding and object detection [8]; (2) high-definition (HD) maps and urban models generation and construction for reliable localization and referencing [2]. These applications have some similar tasks, which can be roughly divided into three types: 3D point cloud segmentation, 3D object detection and localization, and 3D object classification and recognition. Such a technique has led to an increasing and urgent requirement for automatic analysis of 3D point clouds [9] for AVs.

Driven by the breakthroughs brought by deep learning (DL) techniques and the accessibility of 3D point cloud, the 3D DL frameworks have been investigated based on the extension of 2D DL architectures to 3D data with a notable string of empirical successes. These frameworks can be applied to several tasks specifically for AVs such as: segmentation and scene understanding [10–12], object detection [13, 14], and classification [10, 15, 16]. Thus, we provide a systematic survey in this paper, which focuses explicitly on framing the LiDAR point clouds in segmentation, detection, and classification tasks for autonomous driving using DL techniques.

Several related surveys based on DL have been published in recent years. The basic and comprehensive knowledge of DL is described in detail in [17, 18]. These surveys normally

focused on reviewing DL applications in visual data [19, 20] and remote sensing imagery [21, 22]. Some are targeted at more specific tasks such as object detection [23, 24], semantic segmentation [25], recognition [26]. Although DL in 3D data has been surveyed in [27–29], these 3D data are mainly 3D CAD models [30]. In [1], challenges, datasets, and methods in computer vision for AVs are reviewed. However, DL applications in LiDAR point cloud data have not been comprehensively reviewed and analyzed. We summarize these surveys related to DL in Fig.1.

There also have several surveys published for LiDAR point clouds. In [31–34], 3D road object segmentation, detection, and classification from mobile LiDAR point clouds are introduced, but they are focusing on general methods not specific for DL models. In [35], comprehensive 3D descriptors are analyzed. In [36, 37], approaches of 3D object detection applied for autonomous driving are concluded. However, DL models applied in these tasks have not been comprehensively analyzed. Thus, the goal of this paper is to provide a systematic review of DL using LiDAR point clouds in the field of autonomous driving for specific tasks such as segmentation, detection/localization, and classification.

The main contributions of our work can be summarized as:

- An in-depth and organized survey of the milestone 3D deep models and a comprehensive survey of DL methods aimed at tasks such as segmentation, object detection/localization, and classification/recognition in AVs, their origins, and their contributions.
- A comprehensive survey of existing LiDAR datasets that can be exploited in training DL models for AVs.
- A detailed introduction for quantitative evaluation metrics and performance comparison for segmentation, detection, and classification.
- A list of the remaining challenges and future researches that help to advance the development of DL in the field of autonomous driving.

The remainder of this paper is organized as follows: Tasks in autonomous driving and the challenges of DL using LiDAR point cloud data are introduced in Section II. A summary of existing LiDAR point clouds datasets and evaluation metrics are described in Section III. Then the milestone 3D deep models with four data representations of LiDAR point clouds are described in Section IV. The DL applications in segmentation, object detection/localization, and classification/recognition for AVs based on LiDAR point clouds are reviewed and discussed in Section V. Section VI proposes a list of the remaining challenges for future researches. We finally conclude the paper in Section VII.

## II. TASKS AND CHALLENGES

### A. Tasks

In the perception module of autonomous vehicles, semantic segmentation, object detection, object localization, and classification/recognition constitute the foundation for reliable navigation and accurate decision [38]. These tasks are described as follows respectively:

- **3D point cloud semantic segmentation:** Point cloud segmentation is the process to cluster the input data into several homogeneous regions, where points in the same region have the identical attributes [39]. Each input point is predicted with a semantic label, such as ground, tree, building. The task can be concluded as: given a set of ordered 3D points  $X = \{x_1, x_2, x_i, \dots, x_n\}$  with  $x_i \in R^3$  and a candidate label set  $Y = \{y_1, y_2, \dots, y_k\}$ , assign each input point  $x_i$  with one of the  $k$  semantic labels [40]. Segmentation results can further support object detection and classification, as shown in Fig.2(a).
- **3D object detection/localization:** Given an arbitrary point cloud data, the goal of 3D object detection is to detect and locate the instances of predefined categories (e.g., cars, pedestrians, and cyclists, as shown in Fig.2(b)), and return their geometric 3D location, orientation and semantic instance label [41]. Such information can be represented coarsely using a 3D bounding box which is tightly bounding the detected object [13, 42, 42]. This box is commonly represented as  $(x, y, z, h, w, l, \theta, c)$ , where  $(x, y, z)$  denotes the object (bounding box) center position,  $(h, w, l)$  represents the bounding box size with width, length and height, and  $\theta$  is the object orientation. The orientation refers to the rigid transformation that aligns the detected object to its instance in the scene, which are the translations in each of the  $x$ ,  $y$ , and  $z$  directions as well as a rotation about each of these three axes [43, 44].  $c$  represents the semantic label of this bounding box (object).
- **3D object classification/recognition:** Given several groups of point clouds, the objectiveness of classification/recognition is to determine the category (e.g., mug, table, or car, as shown in Fig.2(c)) the group points belong to. The problem of 3D object classification can be defined as: given a set of 3D ordered points  $X = \{x_1, x_2, x_i, \dots, x_n\}$  with  $x_i \in R^3$  and a candidate label set  $Y = \{y_1, y_2, \dots, y_k\}$ , assign the whole point set  $X$  with one of the  $k$  labels [45].

### B. Challenges and Problems

In order to segment, detect, and classify the general objects using DL for AVs with robust and discriminative performance, several challenges and problems that must be addressed, as shown in Fig.2. The variation of sensing conditions and unconstrained environments results in the challenges on data. The irregular data format and requirements for both accuracy and efficiency pose the problems that DL models need to solve.

1) *Challenges on LiDAR point clouds:* Changes in sensing conditions and unconstrained environments have dramatic impacts on object appearance. In particular, the objects captured at different scenes or instances exist a set of variations. Even for the same scene, the scanning times, locations, weather conditions, sensor types, sensing distances and backgrounds are all brought about intra-class differences. All these conditions produce significant variations for both intra- and extra-class objects in LiDAR point cloud data:

- **Diversified point density and reflective intensity.** Due to the scanning mode of LiDAR, the density and intensity

	<ul style="list-style-type: none"> <li>Diversified point density and reflective intensity (Depending on the distance between objects and LiDAR sensors)</li> <li>Noisy (Originated from sensors, such as point perturbations and outliers)</li> <li>Incompleteness (Caused by occlusion between objects and cluttered background)</li> <li>Confusion categories (Caused by shape-similar or reflectance similar objects)</li> </ul>
	<p>(d) Challenges on data</p> <ul style="list-style-type: none"> <li>➤ Permutation and orientation invariance (Caused by unordered and unoriented data structure)</li> <li>➤ Rigid transformation challenge (e.g., 3D rotations and translations)</li> <li>➤ Big data challenge (LiDAR collects millions to billions of points in different scenes)</li> <li>➤ Accuracy challenge (Caused by vast range of intra- and extra-class variations and the quality of data)</li> <li>➤ Efficiency challenge (Caused by limited computational capabilities and storage space)</li> </ul>
	<p>(e) Problems for DL models</p>

Fig. 2. Tasks and challenges related to DL-based applications on 3D point clouds: (a) Point cloud segmentation [10], (b) 3D object detection [41], (c) 3D object classification [10], (d) challenges on LiDAR point clouds, (e) Problems for DL models

for objects vary a lot. The distribution of these two characteristics highly depends on the distance between objects and LiDAR sensors [46–48]. Besides, the ability of the LiDAR sensors, the time constraints of scanning and needed resolution also affect their distribution and intensity.

- **Noisy.** All sensors are noisy. There are a few types of noise that include point perturbations and outliers [49]. It means that a point has some probability to be within a sphere of a certain radius around the place it was sampled (perturbations), or it may appear in a random position in space [50].
- **Incompleteness.** Point cloud data obtained by LiDAR are commonly incomplete [51]. This mainly results from the occlusion between objects [50], cluttered background in urban scenes [46, 49], and unsatisfactory material surface reflectivity. Such problems are severe in real-time capturing of moving objects, which exist large gaping holes and severe under-sampling.
- **Confusion categories.** In a natural environment, shape-similar or reflectance similar objects have interference in object detection and classification. For example, some manmade objects such as commercial billboards have similar shapes and reflectance with traffic signs.

2) *Problems for 3D DL models:* The irregular data format and the requirements for accuracy and efficiency from tasks bring some new challenges for DL models. A discriminative and general-purpose 3D DL model should solve the following problems when designing and constructing its framework:

- **Permutation and orientation invariance.** Compared with 2D grid pixels, the LiDAR point clouds are a set of points with irregular order and no specific orientation

[52]. Within the same group of  $N$  points, the network should feed  $N!$  permutations in an order to be invariant. Besides, the orientation of point sets is missing, which poses a great challenge for object pattern recognition [53].

- **Rigid transformation challenge.** There exist various rigid transformations among point sets, such as 3D rotations and 3D translations. These transformations should not affect the performance of networks [12, 52].
- **Big data challenge.** LiDAR collects millions to billions of points in different urban or rural environments with nature scenes [49]. For example, in Kitti dataset [54], each frame captured by 3D Velodyne laser scanners contains 100k points. The smallest collected scene has 114 frames, which has more than 10 million points. Such amounts of data bring difficulties in data storage.
- **Accuracy challenge.** Accurate perception of road objects is crucial for AVs. However, the variation for both intra-class and extra-class objects and the quality of data pose challenges for accuracy. For example, objects in the same category have a set of different instances, in terms of various material, shape, and size. Besides, the model should be robust to the unevenly distributed, sparse, and missing data.
- **Efficiency challenge.** Compared with 2D images, processing a large quantity number of point clouds produces high computation complexity and time costs. Besides, the computation devices on AVs have limited computational capabilities and storage space [55]. Thus, an efficient and scalable deep network model is critical.

### III. DATASETS AND EVALUATION METRICS

#### A. Datasets

Datasets pave the way towards the rapid development of 3D data application and exploitation using DL networks. There are two roles of reliable datasets: one for providing a comparison for competing algorithms, another for pushing the fields towards more complex and challenging tasks [23]. With the increasing application of LiDAR in multiple fields, such as autonomous driving, remote sensing, photogrammetry, there is a rise of large scale datasets with more than millions of points. These datasets accelerate the crucial breakthroughs and unpredicted performance in point cloud segmentation, 3D object detection, and classification. Apart from the mobile LiDAR data, some discriminative datasets [56] acquired by terrestrial laser scanning (TLS) by static LiDAR are also employed due to they provide high-quality point cloud data.

As shown in Table I, we classify those existing datasets related to our topic into three types: segmentation-based datasets, detection-based datasets, classification-based datasets. Besides, long-term autonomy dataset is also summarized.

- Segmentation-based datasets

**Semantic3D** [56]. Semantic3D is the existing largest LiDAR dataset for outdoor scene segmentation tasks with more than 4 billion points and around  $110,000m^2$  covering area. This dataset is labeled with 8 classes and split into training and test sets with nearly equal size. These data are acquired

TABLE I  
SURVEY OF EXISTING LiDAR DATASET

Dataset	Format	Primary Fields	Points / Objects	# Classes	Sparsity	Highlight
<b>Segmentation</b>						
Semantic3D [56]	ASCII	X, Y, Z, Intensity, R, G, B	4 billion points	8	Dense	training & testing; competing with the most algorithms
Oakland [57]	ASCII	X, Y, Z, Class	1.6 million points	5	Sparse	training to tune model architecture
iQmulus [58]	PLY	X, Y, Z, Intensity, GPS time, Scan origin, # echoes, Object ID, Class	300.0 million points	22	Moderate	training & testing
Paris-Lille-3D [59]	PLY	X, Y, Z, Intensity, Class	143.1 million points	50	Moderate	training & testing; competing with limited algorithms
<b>Localization/Detection</b>						
KITTI Object Detection/ Bird's Eye View Benchmark [60]	-	3D bounding boxes	80,256 objects	3	Sparse	training & testing; competing with the most algorithms
<b>Classification/Recognition</b>						
Sydney Urban Objects Dataset [61]	ASCII	Timestamp, Intensity, Laser id, X, Y, Z, Azimuth, Range, id	588 objects	14	Sparse	training & testing; competing with limited algorithms
ModelNet [30]	ASCII	X, Y, Z, number of Vertices, edges, faces	12,311 objects (ModelNet40) 4,899 objects (ModelNet10)	40 (ModelNet40) 10 (ModelNet10)	Dense	training & testing; competing with most algorithms

by a static LiDAR with high measurement resolution and covered long measurement distance. The challenges for this dataset mainly stems from the massive point clouds, unevenly distributed point density, and severe occlusions. In order to fit the high computation algorithms, a reduced-8 dataset is introduced for training and testing, which share the same training data but fewer test data compared with Semantic3D.

**Oakland 3-D Point Cloud Dataset** [57]. This dataset is acquired in an early year compared with the above two datasets. A mobile platform equipped with LiDAR is used to scan the urban environment and generated around 1.3 million points, while 100,000 points are split into a validation set. The whole dataset is labeled with 5 classes such as wire, vegetation, ground, pole/tree-trunk, and facade. This dataset is small and thus suitable for lightweight networks. Besides, this dataset can be used to test and tune the network architectures without a lot of training time before final training on other datasets.

**iQmulus & TerraMobilita Contest** [58]. This dataset is also acquired by a mobile LiDAR system in the urban environment in Paris. There are more than 300 million points in this dataset, which covered 10km street. The data is split into 10 separate zones and labeled with more than 20 fine classes. However, this dataset also has severe occlusion.

**Paris-Lille-3D** [59]. Compared with Semantic3D [56], Paris-Lille-3D contains fewer points (140 million points) and covering area ( $55,000m^2$ ). The main difference of this dataset is that its data are acquired by a Mobile LiDAR system in two cities: Paris and Lille. Thus, the points in this dataset are sparse and comparatively low measurement resolution compared with Semantic3D [56]. But this dataset is more similar to the LiDAR data acquired by AVs. The whole dataset

is fully annotated into 50 classes unequally distributed in three scenes:Lille1, Lille2, and Paris. For simplicity, these 50 classes are combined into 10 coarse classes for challenging.

- Detection-based datasets

**KITTI Object Detection/Birds Eye View Benchmark** [60]. Different from the above LiDAR datasets which are specific for segmentation task, KITTI dataset is acquired from an autonomous driving platform and records six hours driving using digital cameras, LiDAR, GPS/IMU inertial navigation system. Thus, apart from the LiDAR data, the corresponding imagery data are also provided. Both the Object Detection and Birds Eye View Benchmark contains 7481 training images and 7518 test images as well as the corresponding point clouds. Due to the moving scanning mode, the LiDAR data in this benchmark is highly sparse. Thus, only three objects are labeled with bounding box: cars, pedestrians, and cyclists.

- Classification-based datasets

**Sydney Urban Objects Dataset** [61]. This dataset contains a set of general urban road objects scanned with a LiDAR in the CBD of Sydney, Australia. There are 588 labeled objects and classified in 14 categories, such as vehicles, pedestrians, signs, and trees. The whole dataset is split into four folds for training and testing. Similar to other LiDAR datasets, the collected objects in this dataset are sparse with incomplete shape. Although it is small and not ideal for the classification task, it the most commonly used benchmark due to the limitation of the tedious labeling process.

**ModelNet** [30]. This dataset is the existing largest 3D benchmark for 3D object recognition. Different from Sydney Urban Objects Dataset [61], which contains road objects collected by LiDAR sensors, this dataset is composed of general objects in CAD models with evenly distributed point density

TABLE II  
EVALUATION METRICS FOR 3D POINT CLOUD SEGMENTATION, DETECTION/LOCALIZATION, AND CLASSIFICATION

Metric	Equation	Description
$IoU$	$IoU_i = \frac{c_{ii}}{c_{ii} + \sum_{j \neq i} c_{ij} + \sum_{k \neq i} c_{ki}}$	Intersection over Union, where $c_{ij}$ is the number of points from ground-truth class $i$ predicted as class $j$ [62]
$\overline{IoU}$	$\overline{IoU} = \frac{\sum_{i=1}^N IoU_i}{N}$	Mean IoU, where N is the number of classes
OA	$OA = \frac{\sum_{i=1}^N c_{ii}}{\sum_{j=1}^N \sum_{k=1}^N c_{jk}}$	Overall accuracy
Precision	$Precision = \frac{TP}{TP+FP}$	The ratio of correctly detected objects in the whole detection results, where $TP, TN, FP$ , and $FN$ are the number of true positives, true negatives, false positives, and false negatives, respectively [63]
Recall	$Recall = \frac{TP}{(TP+FN)}$	The ratio of correctly detected objects in the ground truth
$F_1$	$F_1 = \frac{2TP}{(2TP+FP+FN)}$	The balance between precision and recall
MCC	$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))}}$	The combined ratio of detected and undetected objects as well as non-objects
AP	$AP = \frac{1}{11} \sum_{r \in \{0,1,\dots,11\}} \max_{\tilde{r}; \tilde{r} \geq r} p(\tilde{r})$	Average Precision, where $r$ represents the recall, $p(r)$ represents the precision
AOS	$AOS = \frac{1}{11} \sum_{r \in \{0,0.1,\dots,1\}} \max_{\tilde{r}; \tilde{r} \geq r} s(\tilde{r}),$	Average Orientation Similarity
$s(r)$	$s(r) = s(r) = \frac{1}{ \mathcal{D}(r) } \sum_{i \in \mathcal{D}(r)} \frac{1 + \cos \Delta_\theta^{(i)}}{2} \delta_i$	Orientation similarity, where $\mathcal{D}(r)$ represents the whole object detection at recall rate $r$ and $\Delta_\theta^{(i)}$ is the angle difference between predicted and ground truth orientation of detection $i$ , $\delta_i$ is the penalty value when multiple detection tasks describe one object

and complete shape. There are approximately 130K labeled models in a total of 660 categories (e.g., car, chair, clock). The most commonly used benchmarks are ModelNet40 that contains 40 general objects and ModelNet10 with 10 general objects. The milestone 3D deep architectures are commonly trained and tested on these two datasets due to the affordable computation burden and time.

**Long-Term Autonomy:** To address challenges of long-term autonomy, a novel dataset for autonomous driving has been presented by Maddern et al. [64]. They collected images, LiDAR, and GPS data while traversing 1,000 km in central Oxford in the UK for one year. This allowed them to capture different scene appearances under various illumination, weather, and season with dynamic objects and constructions. Such long-term datasets allow for in-depth investigation of problems that detain the realization of autonomous vehicles such as localization at different times of the year.

### B. Evaluation Metrics

To evaluate those proposed methods performance, several metrics, as summarized in Table II, are proposed for those tasks: segmentation, detection, and classification. The detail of these metrics is given as follows.

For the segmentation task, the most commonly used evaluation metrics are the Intersection over Union (IoU) metric,  $\overline{IoU}$ , and overall accuracy (OA) [62]. IoU defines the quantify the percent overlap between the target mask and the prediction output [56].

For detection and classification tasks, the results are commonly analyzed region-wise. Precision, recall,  $F_1$ -score and Matthews correlation coefficient (MCC) [65] are commonly used to evaluate the performance. The precision represents the ratio of correctly detected objects in the whole detection results, while the recall means the percentage of the correctly detected objects in the ground truth, the  $F_1$ -score conveys the balance between the precision and the recall, the MCC is the

combined ratio of detected and undetected objects and non-objects.

For 3D object localization and detection task, the most frequently used metrics are: Average Precision ( $AP_{3D}$ ) [66], and Average Orientation Similarity (AOS) [36]. The average precision is used to evaluate the localization and detection performance by calculating the averaged valid bounding box overlaps, which exceed predefined values. For orientation estimation, the orientation similarities with different thresholded valid bounding box overlaps are averaged to report the performance.

## IV. GENERAL 3D DEEP LEARNING FRAMEWORKS

In this section, we review the milestone DL frameworks on 3D data. These frameworks are pioneers in solving the problems defined in section II. Besides, their stable and efficient performance makes them suitable for use as the backbone framework in detection, segmentation and classification tasks. Although 3D data acquired by LiDAR is often in the form of point clouds, how to represent point cloud and what DL models to use for detection, segmentation and classifications remains an open problem [41]. Most existing 3D DL models process point clouds mainly in form of voxel grids [30, 67–69], point clouds [10, 12, 70, 71], graphs [72–75] and 2D images [15, 76–78]. In this section, we analyze the frameworks, attributes and problems of these models in detail.

### A. Voxel-based models

Conventionally, CNNs are mainly applied to data with regular structures, such as the 2D pixel array [79]. Thus, in order to apply CNNs to unordered 3D point cloud data, such data are divided into regular grids with a certain size to describe the distribution of data in 3D space. Typically, the size of the grid is related to the resolution of data [80]. The advantage of voxel-based representation is that it can encode

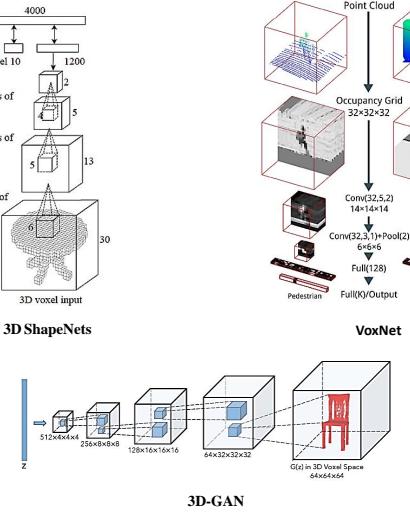


Fig. 3. Deep architectures of 3D ShapeNet [30], VoxNet [67], 3D-GAN [68].

the 3D shape and viewpoint information by classifying the occupied voxels into several types such as visible, occluded, or self-occluded. Besides, 3D convolution (Conv) and pooling operations can be directly applied in voxel grids [69].

**3D ShapeNet** [30], proposed by Wu et al. and shown in Fig.3, is the pioneer in exploiting 3D volumetric data using a convolutional deep belief network. The probability distribution of binary variables is used to represent the geometric shape of a 3D voxel grid. Then these distributions are input to the network which is mainly composed of three Conv layers. This network is initially pre-trained in a layer-wise fashion and then trained with a generative fine-tuning procedure. The input and Conv layers are modeled based on the Contrastive Divergence, where the output layer was trained based on the Fast-Persistent Contrastive Divergence. After training, the input test data is output with a single depth map and then transformed to represent the voxel grid. ShapeNet has notable results in low-resolution voxels. However, the computation cost increases cubically with the increment of input data size or resolution, which limit the models performance in large-scale or dense point clouds data. Besides, multi-scale and multi-view information from the data is not fully exploited, which hinder the output performance.

**VoxNet** [67] is proposed by Maturana et al. to conduct 3D object recognition using 3D convolution filters based on volumetric data representation, as shown in Fig.3. Occupancy grids represented by a 3D lattice of random variables are employed to show the state of the environment. Then a probabilistic estimate is used to estimate the occupancy of these grids which is maintained as the prior knowledge. Three different occupancy grid models, such as binary occupancy grid, density grid, and hit grid are experimented to select the best model. This network framework is mainly composed of Conv, pooling layer, and fully connected (FC) layers. Both ShapeNet [30] and VoxNet employ rotation augmentation for training. Compared with ShapeNet [30], VoxNet has a smaller architecture that has less than 1 million parameters. However,

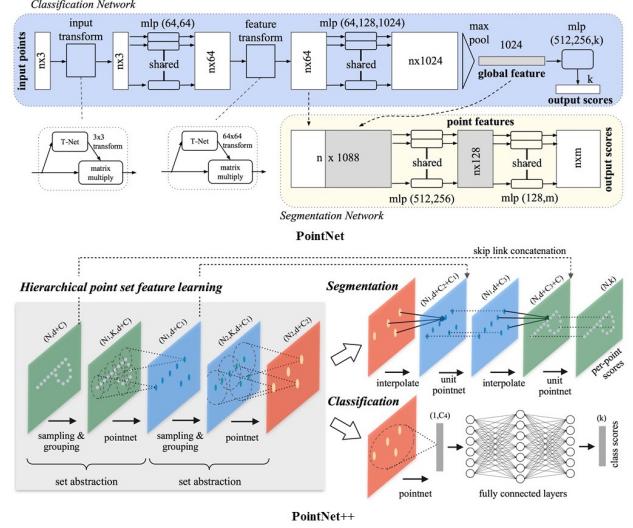


Fig. 4. PointNet [10] and PointNet++ [12] architectures.

not all occupancy grids contain useful information but only increase the computation cost.

**3D-GAN** [68] combines the merits of both general-adversarial network (GAN) [81] and volumetric convolutional networks [67] to learn the features of 3D objects. This network is composed of a generator and a discriminator as shown in Fig.3. The adversarial discriminator is conducted to classify objects into synthesized and real categories due to the generative-adversarial criterion has the advantage in capturing the structural variation between two 3D objects. And the employment of generative-adversarial loss is helpful to avoid possible criterion-dependent over-fitting. The generator attempts to confuse the discriminator. Both generator and discriminator consist of five volumetric fully Conv layers. This network provides a powerful 3D shape descriptor with unsupervised training in 3D object recognition. But the density of data affects the performance of adversarial discriminator for finest feature capturing. Consequently, this adaptive method is suitable for evenly distributed point cloud data.

In conclusion, there are some limitations of this general volumetric 3D data representation:

- Firstly, not all voxel representations are useful because they contain occupied and non-occupied parts of the scanning environment. Thus, the high demand for computer storage is actually unnecessary within this ineffective data representation [69].
- Secondly, the size of the grid is hard to set, which affects the scale of input data and may disrupt the spatial relationship between points.
- Thirdly, computational and memory requirements grow cubically with the resolution [69]. Thus, existing voxel-based models are maintained at low 3D resolutions, and the most commonly used size is  $30^3$  for each grid.[69].

A more advanced voxel-based data representation is the octree-based grids [69, 82], which use adaptive size to divides the 3D point cloud into cubes. It is a hierarchical data structure

that recursively decomposes the root voxels into multiple leaf voxels.

**OctNet** [69] is proposed by Riegler et al., which exploits the sparsity of the input data. Motivated by the observation that the object boundaries have the highest probability in producing the maximum responses across all feature maps generated by the network at different layers, they partitioned the 3D space hierarchically into a set of unbalanced octrees [83] based on the density of the input data. Specifically, the octree nodes that have point clouds are split recursively in its domain, ending at the finest resolution of the tree. Thus, the size of leaf nodes varies. For each leaf node, those features that activate their comprised voxel is pooled and stored. Then the convolution filters are conducted in these trees. In [82], the deep model is constructed by learning the structure of the octree and the represented occupancy value for each grid. This octree-based data representation largely reduces the computation and memory resources for DL architectures, which achieves better performance in high-resolution 3D data compared with voxel-based models. However, the disadvantage of octree data is similar to voxels, both of them fail to exploit the geometry feature of 3D objects, especially the intrinsic characteristics of patterns and surfaces [29].

### B. Point clouds based models

Different from volumetric 3D data representation, point cloud data can preserve the 3D geospatial information and internal local structure. Besides, the voxel-based models that scan the space with fixed strides are constrained by the local receptive fields. But for point clouds, the input data and the metric decide the range of receptive fields, which has high efficiency and accuracy.

**PointNet** [10], as a pioneer in consuming 3D point clouds directly for deep models, learns the spatial feature of each point independently via MLP layers and then accumulates their features by max-pooling. The point cloud data are input directly to the PointNet, which predicts per-point label or per-object label, its framework is illustrated in Fig.4. In PointNet, spatial transform network and a symmetric function are designed to improve the invariance to permutation. The spatial feature of each input point was learned through the networks. Then, the learned features are assembled across the whole region of point clouds. The outstanding performance of PointNet has achieved in 3D objects classification and segmentation tasks. However, the individual point features are grouped and pooled by max-pooling, which fails to preserve the local structure. As a result, PointNet is not robust to fine-grained patterns and complex scenes.

**PointNet++** was proposed later by Qi et al. [12], which compensates the local feature extraction problems in PointNet. Within the raw unordered point clouds as input, these points are initially divided into overlapping local regions using the Euclidean distance metric. These partitions are defined as a neighborhood ball in this metric space and labeled with the centroid location and scale. In order to sample the points evenly over the whole point set, the farthest point sampling (FPS) algorithm is applied. Local features are extracted from

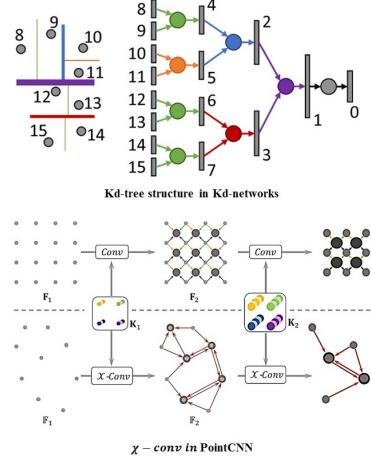


Fig. 5. Kd-tree structure in Kd-networks [70] and  $\chi$ -Conv in PointCNN [71].

the small neighborhoods around the selected points using K-nearest-neighbor (KNN) or query-ball searching methods. These neighborhoods are gathered into larger clusters and leveraged to extract high-level features via PointNet [10] network. The sampling and grouping module are repeated until the local and global features of the whole points are learned, as shown in Fig.4. This network, which outperforms the PointNet [10] network in classification and segmentation tasks, extracts the local feature for points in different scales. However, features from the local neighborhood points in different sampling layers are learned in an isolated fashion. Besides, max-pooling operation based on PointNet [10] for high-level feature extraction in PointNet++ fails to preserve the spatial information between the local neighborhood points.

**Kd-networks** [70] uses the kd-tree to create the order of the input points, which is different from PointNet [10] and PointNet++ [12] as both of them use the symmetric function to solve the permutation problem. Klokov et al. used the maximum range of point coordinates along the coordinate axis to recursively split the certain size point clouds  $N = 2^D$  into subsets with a top-down fashion to construct a kd-tree. As shown in Fig.5, this kd-tree is ending with a fixed depth. Within this balanced tree structure, vectorial representations in each node, which represents a subdivision along certain axis, is computed using kd-networks. These representations are then exploited to train a linear classifier. This network has better performance than PointNet [10] and PointNet++ [12] in small objects classification. However, it is not robust to rotations and noise, since these variations can lead to the change of tree structure. Besides, it lacks the overlapped receptive field which reduces the spatial-correlation between leaf nodes.

**PointCNN**, proposed by Li et al. [71], solves the input points permutation and transformation problems based on an  $\chi$ -Conv operation, as shown in Fig.5. They proposed the  $\chi$ -transformation which is learned from the input points by weighting the input point features and permutating the points into a latent and potentially canonical order. Then the traditional convolution operators are applied in the learned  $\chi$ -transformation features. These spatially-local correlation

features in each local range are aggregated to construct a hierarchical CNN network architecture. However, this model still has not exploited the correlations of different geometric features and their discriminate information toward results, which limits the performance.

Point cloud based deep models are mostly focused on solving permutation problems. Although they treat points independently at local scales to maintain permutation invariance. This independence, however, neglects the geometric relationships among points and their neighbors, presenting a fundamental limitation that leads to local features' missing.

### C. Graph-based models

Graphs are a type of non-Euclidean data structure that can be used to represent point cloud data. Their node corresponds to each input point and the edges represent the relationship between each point neighbors. Graph neural networks propagate the node states until equilibrium in an iterative manner [75]. With the advancement of CNNs, there is an increment graph convolutional networks applied to 3D data. Those graph CNNs define convolutions directly on the graph in the spectral and non-spectral (spatial) domain, operating on groups of spatially close neighbors [84]. The advantage of graph-based models is that the geometric relationships among points and their neighbors are exploited. Thus, more spatially-local correlation features are extracted from the grouped edge relationships on each node. But there are two challenges for constructing graph-based deep models:

- Firstly, defining an operator that is suitable for dynamically sized neighborhoods and maintaining the weight sharing scheme of CNNs [75].
- Secondly, exploiting the spatial and geometric relationships among each node's neighbors.

**SyncSpecCNN** [72] exploited the spectral eigen-decomposition of the graph Laplacian to generate a convolution filter applied in point clouds. Yi et al. constructed SyncSpecCNN based on that two considerations: the first is the coefficients sharing and multi-scale graph analyzing; the second is information sharing across related but different graphs. They solved these two problems by constructing the convolution operation in the spectral domain: the signal of point sets in the Euclidean domain is defined by the metrics on the graph nodes, and the convolution operation in the Euclidean domain is related to the scaling signals based on eigenvalues. Actually, such operation is linear and only applicable to the graph weights generated from eigenvectors of the graph Laplacian. Despite SyncSpecCNN achieved excellent performance in 3D shape part segmentation, it has several limitations:

- Basis-dependent. The learned spectral filters coefficients are not suitable for another domain with a different basis.
- Computationally expensive. The spectral filtering is calculated based on the whole input data, which requires high computation capability.
- Missing local edge features. The local graph neighborhood contains useful and distinctive local structural information, which is not exploited.

**Edge-conditioned convolution** (ECC) [73] considers the edge information in constructing the convolution filters based on the graph signal in the spatial domain. The edge labels in a vertex neighborhood are conditioned to generate the Conv filter weights. Besides, in order to solve the basis-dependent problem, they dynamic generalized the convolution operator for arbitrary graphs with varying size and connectivity. The whole network follows the common structure of feedforward network with interlaced convolutions and pooling followed by global pooling and FC layers. Thus, features from local neighborhoods are extracted continually from these stacked layers, which increase the receptive field. Although the edge labels are fixed for a specific graph, the learned interpretation networks may vary in different layers. ECC learns the dynamic pattern of local neighborhoods, which is scalable and effective. However, the computation cost remains high, and it is not applicable for large-scale graphs with continuous edge labels.

**DGCNN** [74] also constructed a local neighborhood graph to extract the local geometric features and applied Conv-like operations, named EdgeConv which is shown in Fig.6, on the edges connecting neighboring pairs of each point. Different from ECC [73], EdgeConv dynamically updates the given fixed graph with Conv-like operations for each layer output. Thus, DGCNN can learn how to extract local geometric structures and group point clouds. This model takes  $n$  points as input, and then find the K neighborhoods of each point to calculate the edge feature between the point and its K neighborhoods in each EdgeConv layer. Similar to PointNet[34] architecture, the features convolved in the last EdgeConv layer are aggregated globally to construct a global feature, while all the EdgeConv outputs are treated as local features. Local and global features are concatenated to generate results score. This model extracts distinctive edge features from point neighborhoods, which can be applied in different point clouds related tasks. However, the fixed size of edge features limits the performance of the model when facing different scales and resolution point clouds.

ECC [73] and DGCNN [74] propose general convolutions on graph nodes and their edge information, which is isotropy about input features. However, not all the input features contribute equally to its nodes. Thus, attention mechanisms are introduced to deal with variable sized inputs and focus on the most relevant parts of the nodes' neighbors to make decisions [75].

**Graph Attention Networks** (GAT) [75]. The core insight behind GAT is to calculate the hidden representations of each node in the graph, by assigning different attentional weights to different neighbors, following a self-attention strategy. Within a set of node features as input, a shared linear transformation, parametrized by a weight matrix is applied to each node. Then a self-attention, a shared attentional mechanism which is shown in Fig.6, is applied on the nodes to computes attention coefficients. These coefficients indicate the importance of corresponding nodes' neighbor features, respectively, and are further normalized to make them comparable across different nodes. These local features are combined according to the attentional weights to form the output features for each node. In order to improve the stability of the self-attention

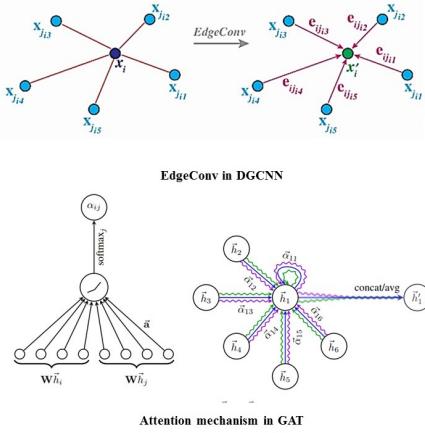


Fig. 6. EdgeConv in DGCNN [74] and attention mechanism in GAT [75].

mechanism, multi-head attention is employed to conduct  $k$  independent attention schemes, which are then concatenated together to form the final output features for each node. This attention architecture is efficient and can extract fine-grained representations for each graph node by assigning different weights to the neighbors. However, local spatial relationship between neighbors are not considered in calculating the attentional weights. To further improve its performance, Wang et al. [85] proposed graph attention convolution (GAC) to generate attentional weights by considering different neighboring points and feature channels.

#### D. View-based models

The last type of MLS data representation is 2D views obtained from 3D point clouds from different directions. With the projected 2D views, traditional well-established convolutional neural networks (CNN) and pre-trained networks on image datasets, such as AlexNet [86], VGG [87], GoogLeNet [88], ResNet [89] can be exploited. Compared with voxel-based models, these methods can improve the performance for different 3D tasks by taking multi-view of the interest object or scenes and then fusing or voting the outputs for final prediction. Compared with the above three different 3D data representations, view-based models can achieve near-optimal results, as shown in Table III. Su et al. [90] experimented that multiview methods have the optimal generalization ability even without using pre-trained models compared with point cloud and voxel data representation models. The advantages of view-based models compared with 3D models can be concluded as:

- Efficiency. Compared with 3D data representations such as point clouds or voxel grids, the reduced one dimension information can greatly reduce the computation cost but with increased resolution [76].
- Exploiting established 2D deep architectures and datasets. The well-developed 2D DL architectures can better exploit the local and global information from projected 2D view images [91]. Besides, existing 2D image databases

(such as ImageNet [92]) can be used to train 2D DL architectures.

**Multi-View CNN (MVCNN)** [76] is the pioneer in exploiting 2D DL models to learn 3D representation. Multiple views of 3D objects are extracted without specific order using a view pooling layer. Two different CNNs models are proposed and tested in this paper. The first CNN model takes 12 views rendered from the object via placing 12 virtual cameras with equal distance around the objects as the input, while the second CNN model takes 80 views rendered in the same way as input. These views are first learned separately and then fused through max-pooling operation the extract the most representative feature among all views for the whole 3D shape. This network is effective and efficient compared with volumetric data representation. However, the max-pooling operation only considers the most important views and discards information from other views, which fails to preserve comprehensive visual information.

**MVCNN-MultiRes** was proposed by Qi et al [15] to improve multi-view CNNs. Different from traditional view rendering methods, the 3D shape is projected to 2D via a convolution operation based on an anisotropic probing kernel applied to the 3D volume. Multi-orientation pooling is combined together to improve the 3D structure capturing capability. Then the MVCNN [76] is applied to classify the 2D projects. Compared with MVCNN [76], multi-resolution 3D filtering is introduced to capture multi-scale information. Sphere rendering is performed at different volume resolutions to achieve view-invariant and improve the robust to potential noise and irregularities. This model achieves better results in 3D object classification task compared with MVCNN [76].

**3DMV** [77] combines the geometry and imagery data as input to train a joint 3D deep architecture. Feature maps extracted from imagery data are first extracted and then mapped into the 3D feature extracted from the volumetric grid data derived from a differentiable back-projection layer. Because there exists redundant information among multiple views, a multiview pooling approach is applied to extract useful information from these views. This network achieved remarkable results in 3D objects classification. However, compared with models using one source of data such as LiDAR point or RGB images solely, the computation cost of this method is higher.

**RotationNet** [78] is proposed following the assumption that when the object is observed by a viewer from a partial set of full multiview images, the observation direction should be recognized to correctly infer the objects category. Thus, the multiview images of an object are input to the RotationNet, which outputs its pose and category. The most representative characteristic of RotationNet is that it treats viewpoints which are the observation of training images as latent variables. Then unsupervised learning of object poses is conducted based on an unaligned object dataset, which can eliminate the process of pose normalization to reduce noise and individual variations in shape. The whole network is constructed as a differentiable MLP network with softmax layers as the final layer. The outputs are the viewpoint category probabilities, which correspond to the predefined discrete viewpoints for

TABLE III  
SUMMARIZING MILESTONE DL ARCHITECTURES BASED ON FOUR POINT CLOUD DATA REPRESENTATIONS

Model	Input Size	Highlights	Disadvantages	Model size(MB)	Acc (%)
<b>Voxel</b>					
3dShapeNet [30]	voxels	Pioneer in exploiting 3D volumetric data; Permutation and orientation invariance.	Computation and memory requirement grows cubically; Use one view in a fixed voxel size.	12	84.7
VoxNet [67]	voxels	Occupancy grids are employed to represent the distribution of the scene as a 3D lattice of random variables for each grid; Permutation and orientation invariance; Improved efficiency.	Not all occupancies are useful.	1.0	85.9
3D-GAN [68]	voxels	Combines the adversarial modeling and volumetric convolutional networks to learn features; Permutation and orientation invariance; Rigid transformation invariance.	Not invariance to data density variation	7	83.3
OctNet [69]	hybrid grid octree	Hierarchically divide the data into a series of unbalanced octrees according to data density; Permutation and orientation invariance; Efficient.	Fail to preserve the geometry relationship among points.	0.4	86.5
<b>Point Clouds</b>					
PointNet [10]	1024 points	Pioneer in applying DL using 3D point clouds and solving the permutation problem via maxpooling.	Not capture local structure induced by the metric; Hard to generalize to unseen point configurations.	40	89.2
PointNet++ [12]	5000 points +normal	Hierarchically learn multi-scale local geometric features and aggregate them for inference; Permutation and rigid transformation invariance and efficient.	Local spatial relationship among point neighborhoods is not exploited.	12	90.7
Kd-networks [70]	1024 points	Use the kd-tree to create the order of the input points and hierarchically extract features from the leaves to root; Permutation and rigid transformation invariance.	Non-invariance to rotations and noises; Computation grows linearly with increasing resolution; Low spatial-correlation between leaf nodes.	120	91.8
PointCNN [71]	1024 points	Propose X-Conv operator that permutes and weights input points and features; Permutation and rigid transformation invariance.	Not exploit the correlations of different geometric features and their discriminative information toward final results.	4.5	92.2
<b>Graph</b>					
Spectral -CNN [72]	graphs	Exploit the spectral eigen-decomposition of the graph Laplacian to generate a Conv-like operator.	Basis-dependent; Computationally expensive; Missing local edge features.	0.8	-
ECC [73]	graphs	The edge labels in a vertex neighborhood are conditioned to generate the Conv filter weights; Permutation invariance.	High computation cost; Not suitable for large-scale graphs with continuous labels; Isotropic about input features.	-	87.4
DGCNN [74]	graphs	Extract edge features and dynamically update the graph for each layer; Permutation invariance.	Fixed size edge features are not invariance to points with different resolution and scale; Isotropic about input features.	21	92.2
GAT [75]	graphs	Compute the hidden representations of each node's neighbors, following a self-attention strategy; Permutation and invariance, improved accuracy.	Apply the attention mechanism only to input points not to their local features.	-	-
<b>2D View</b>					
MVCNN [76]	12 views	Pioneer in applying CNN to each view and then aggregate the features by a view pooling procedure; Permutation and orientation invariance; Efficient.	Multi-resolution features are not considered.	99	90.1
MVCNN -MultiRes [15]	20 views	Propose multi-resolution 3D filtering to capture comprehensive information at multi-scales; Permutation and orientation invariance; Efficient.	Geometric information are not exploited.	16.6	91.4
3DMV [77]	20 views	Extract RGB and geometric features and aggregate them via a joint 2D-3D network; Permutation and orientation invariance; Efficient.	2D occlusion and background clutter affects the 3D network performance.	-	-
RotationNet [78]	12 views	Treat the viewpoints of the observed training images as latent variables; Permutation and orientation invariance; High accuracy.	Not suitable for per-point processing tasks.	59	<b>97.37</b>

each input image. These likelihoods are optimized by the selected object pose.

However, there some limitation of 2D view-based models:

- The first is that the projection from 3D space to 2D views can lose some geometrically-related spatial information.
- The second is the redundant information among multiple views.

#### E. 3D Data Processing and Augmentation

Due to the massive amount of data and the tedious labeling process, there exist limited reliable 3D datasets. To better

exploit the architecture of deep networks and improve the model generalization ability, data augmentation is commonly conducted. Augmentation can be applied to both data space and feature space, while the most common augmentation is conducted in the first space. This type of augmentation can not only enrich the variations of data but also can generate new samples by conducting transformations to the existing 3D data. There are several types of transformations, such as translation, rotation, and scaling. Several requirements for data augmentation are summarised as:

- There must exist similar features between original aug-

mented data, such as shape;

- There must exist different features between original and augmented data such as orientation.

Based on those existing methods, classical data augmentation for point clouds can be concluded as:

- Mirror  $x$  and  $y$  axis with predefined probability [59, 93]
- Rotation around z-axis with certain times and angles[13, 59, 93, 94]
- Random (uniform) height or position jittering in certain range [67, 93, 95]
- Random scale with certain ratio [13, 59]
- Random occlusions or randomly down-sampling points within predefined ratio [59]
- Random artefacts or randomly down-sampling points within predefined ratio [59]
- Randomly adding noise, following certain distribution, to the points' coordinates and local features [45, 59, 96].

## V. DEEP LEARNING IN LiDAR POINT CLOUD FOR AVs

The application of LiDAR point clouds for AVs can be concluded into three types: 3D point cloud segmentation, 3D object detection and localization, and 3D objects classification and recognition. Targets for these tasks vary, for example, scene segmentation focus on per-point label prediction, while detection and classification concentrate on integrated point set labeling. But they all need to exploit the input point feature representations before feature embedding and network construction.

We first make a survey of input point cloud feature representations applied in DL architectures for all these three tasks, such as local density and curvature. These features are representations of a specific 3D point or position in 3D space, which describe the geometrical structures and features based on the extracted information around the point. These features can be grouped into two types: one is derived directly from the sensors such as coordinate and intensity, we term them as direct point feature representations; the second is extracted from the information provided by each points neighbors, we term them as geo-local point feature representations.

*1) Direct input point feature representations:* The direct input point feature representations are mainly provided by laser scanners, which include the  $x$ ,  $y$ , and  $z$  coordinates, and other characteristics (e.g., intensity, angle, and number of returns). Two most frequently used features applied in DL are selected:

- **XYZ coordinate.** The most direct point feature representation is the  $XYZ$  coordinate provided by the sensors, which means the position of a point in the real world coordinate.
- **Intensity.** The intensity represents the reflectance characteristics of the material surface, which is one common characteristic of laser scanners [97]. Different objects have different reflectance, thus produce different densities in point clouds. For example, traffic signs have a higher intensity than vegetation.

*2) Geo-local point feature representations:* Local input point feature embeds the spatial relationship of points and their neighborhoods, which plays a significant role in point cloud segmentation [12], object detection [42], and classification [74]. Besides, the searched local region can be exploited by some operations such as CNNs [98]. Two most representative and widely-used neighborhood searching methods are k-nearest neighbors (KNN) [12, 96, 99] and spherical neighborhood [100].

The geo-local feature representations are usually generated from the searched region using the above two neighborhood searching algorithms. They are composed of eigenvalues (e.g.,  $\eta_0$ ,  $\eta_1$  and  $\eta_2$  ( $\eta_0 > \eta_1 > \eta_2$ )) or eigenvectors (e.g.,  $\vec{v}_0^1$ ,  $\vec{v}_1^1$ , and  $\vec{v}_2^1$ ) by decomposing the covariance matrix defined in the searched region. We list five most commonly used 3D local feature descriptors applied in DL:

- **Local density.** The local density is typically determined by the quantity of points in a selected area [101]. Typically, the point density decreases when the distance of objects to the LiDAR sensor increases. In voxel-based models, the local density of points is related to the setting of voxel sizes [102].
- **Local normal.** It infers the direction of the normal at a certain point on the surface. The equation about normal extraction can be found in [65]. In [103], the eigenvector  $\vec{v}_2^1$  of  $\eta_2$  in  $C_i$  is selected as the normal vector for each point. However, in [10], the eigenvectors of  $\eta_0$ ,  $\eta_1$  and  $\eta_2$  are all chose as the normal vectors of point  $p_i$ .
- **Local curvature.** The local curvature is defined to be the rate at which the unit tangent vector changes direction. Similar to local normal calculation in [65], the surface curvature change in [103] can be estimated from the eigenvalues derived from the Eigen decomposition:  $curvature = \eta_0 / (\eta_0 + \eta_1 + \eta_2)$
- **Local linearity.** It is a local geometric characteristic for each point to indicate the linearity of its local geometry [104]:  $linearity = (\eta_1 - \eta_2) / \eta_1$ .
- **Local planarity.** It describes the flatness of a given point neighbors. for example, group points have higher planarity compared with tree points [104]:  $planarity = (\eta_2 - \eta_3) / \eta_1$

### A. LiDAR point cloud semantic segmentation

The goal of semantic segmentation is to label each point as belonging to a specific semantic class. For AVs segmentation tasks, these classes could be a street, buildings, cars, pedestrians, trees or traffic lights. When applying DL for point cloud segmentation, classification of small features is required [38]. However, the LiDAR 3D point clouds are usually acquired in large scale, and they are irregularly shaped with changeable spatial contents. In the review of the recent five years papers related in this region, we group these papers into three schemes according to the types of data representation: point cloud based, voxel-based, and multi-view based models. There is limited research focusing on graph-based models, thus we combine the graph-based and point cloud based models together to illustrate their paradigms. Each type of model is

represented by a compelling deep architecture as shown in Fig.7.

1) *Point cloud based networks*: For point cloud based networks, they are mainly composed of two parts: feature embedding and network construction. For the discriminative feature representing, both local and global features have demonstrated to be crucial for the success of CNNs [12]. However, in order to apply conventional CNNs, the permutation and orientation problem for unordered and unoriented points requires a discriminative feature embedding network. Besides, lightweight, effective, and efficient deep network construction is another key module that affects the segmentation performance.

Local feature is commonly extracted from points neighborhoods [104]. The most frequently used local features are local normal and curvature [10, 12]. To improve the receptive field, PointNet [10] has been proved to be a compelling architecture to extract semantic feature from unordered point sets. Thus, in [12, 105, 108, 109], a simplified PointNet is exploited to abstract local features from sampled point sets into high-level representations. Landrieu et al. [105] proposed superpoint graph (SPG) to represent large 3D point clouds as a set of interconnected simple shapes coined superpoints, then PointNet is operated on these superpoints to embed features.

To solve the permutation problem and extract local features, Huang et al. [40] proposed a novel slice pooling layer to extract the local context layer from the input point features and outputs an ordered sequence of aggregated features. To this end, the input points are first grouped into slices and then a global representation for each slice is generated via concatenating points features within the slice. The advantage of this slice pooling layer is the low computation cost compared with point-based local features. However, the slice size is sensitive to the density of data. In [110], bilateral Conv layers (BCL) are applied to perform convolutions on occupied parts of the lattice for hierarchical and spatially-aware feature learning. BCL first maps input points onto a sparse lattice and applies convolutional operations on the sparse lattice and then the filtered signal are interpolated smoothly to recover the original input points.

To reduce the computation cost, in [108], an encoding-decoding framework is adopted. Features extracted from the same scale of abstraction are combined and then upsampled by 3D deconvolutions to generate the desired output sampling density, which is finally interpolated by Latent nearest-neighbor interpolation to output per-point label. However, the down-sampling and up-sampling operations are hard to preserve the edge information, thus cannot extract the fine-grained features. In [40], RNNs are applied to model dependencies of the ordered global representation derived from slice pooling. Similar to sequence data, each slice is viewed as one timestamp and the interaction information with other slices also follows the timestamps in RNN units. This operation enables the model to generate dependencies between slices.

Although Zhang et al. [65] proposed the ReLu-NN to learn embedded point features, which is a four-layer MLP architecture. However, for objects without discriminative features, such as shrubs or trees, their local spatial relationship is not fully exploited. To better leverage the rich spatial information

of objects, Wang et al. constructed a lightweight and effective deep neural network with spatial pooling (DNNSP) [111] to learn point features. They clustered the input data into groups and then applied distance minimum spanning tree-based pooling to extract the spatial information among the points in the clustered point sets. Finally, an MLP is used for classification with these features. In order to achieve multiple tasks, such as instance segmentation and object detection with simple architecture, Wang et al. [109] proposed a similarity group proposal network SGPN. Within the extracted local and global point features by PointNet, feature extraction network generates a matrix which is then diverged into three subsets that each pass through a single PointNet layer to obtain three similarity matrices. These three matrices are used to produce a similarity matrix, a confidence map and a semantic segmentation map.

2) *Voxel-based networks*: In voxel-based networks, the point clouds are first voxelized into grids and then learn features from these grids. The deep network is finally constructed to map these features into segmentation masks.

Wang et al. [106] conducted a multi-scale voxelization method to extract objects spatial information at different scales to form a comprehensive description. At each scale, a neighboring cubic with selected length is constructed for a given point [112]. After that, the cube is divided into grid voxels with different size as a patch. The smaller the size is, the finer the scale. The point density and occupancy are selected to represent each voxel. The advantage of this kind voxelization is that it can accommodate objects with different sizes without losing their spatial space information. In [113], the class probabilities for each voxel are predicted using 3D-FCNN, which are then transferred back to the raw 3D points based on trilinear interpolation. In [106], after the multi-scale voxelization of point clouds, features at different scales and spatial resolutions are learned by a set of CNNs with shared weights which are finally fused together for final prediction.

In voxel-based point cloud segmentation task, there are two ways to label each point: (1) Using the voxel label derived from the argmax of the predicted probabilities; (2) Further globally optimizing the class label of the point cloud based on spatial consistency. The first method is simple, but the result is provided at the voxel level and inevitably influenced by noise. The second one is more accurate but complex with additional computation. Because the inherent invariance of CNN networks to spatial transformations affects the segmentation accuracy [25]. In order to extract the fine-grained details for volumetric data representations, the Conditional Random Field (CRF) [106, 113, 114] is commonly adopted in a post-processing stage. The CRFs have the advantage in combining low-level information such as the interactions between points to output multi-class inference for multi-class per-point labeling tasks, which compensates the fine local details that CNNs fail to capture.

3) *Multiview-based networks*: As for multi-view based models, view rendering and deep architecture construction are two key modules for segmentation task. The first one is used to generate structural and well-organized 2D grids that can exploit existing CNN-based deep architectures. The second

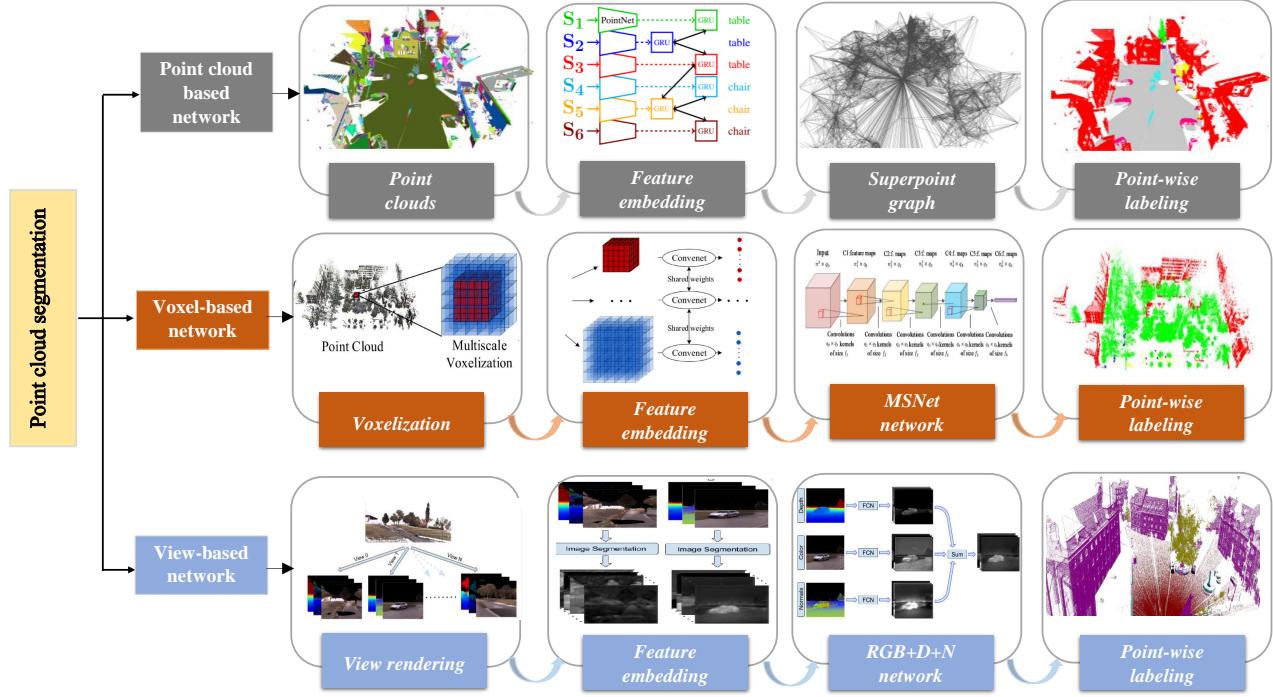


Fig. 7. DL architectures on LiDAR point cloud segmentation with three different data representations: point cloud based networks represented by SPG [105], voxel-based networks represented by MSNet [106], view-based networks represented by DeePr3SS [107]

one is proposed to construct the most suitable and generative models for different data.

In order to extract local and global features simultaneously, some hand-designed feature descriptors are employed for representative information extraction. In [65, 111], the spin image descriptor is employed to represent point-based local features, which contains the global description of objects from partial views and clutters of local shape description. In [107], point splatting was applied to generate view images by projecting the points with a spread function into the image plane. The point is first projected into image coordinates of a virtual camera. For each projected point, its corresponding depth value and feature vectors such as normal are stored.

Once the points are projected into multi-view 2D images, some discriminative 2D deep networks can be exploited, such as VGG16 [87], AlexNet [86], GoogLeNet [88], and ResNet [89]. In [25], these deep networks have been detailed analyzed in 2D semantic segmentation. Among these methods, VGG16 [87], composed of 16 layers, is the most frequently used. Its main advantage is the use of stacked Conv layers with small receptive fields, which produces a lightweight network with limited parameters and increasing nonlinearity [25, 107, 115].

*4) Evaluation on Point cloud segmentation:* Due to the high volume of point clouds, which pose a great challenge for computation capability. We choose the models tested on Reduced-8 Semantic3D dataset to compare their performance, as shown in Table IV. Reduced-8 shares the same training data as semantic-8 but only use a small part of test data, which can also suit the high computation cost algorithm for competing. The metrics used to compare these models are  $IoU_i$ ,  $\overline{IoU}$ , and OA. The computation efficiency for these

algorithms are not reported and compared due to the difference between computation capacity, selected training dataset, model architecture.

### B. 3D objects detection (localization)

The detection(& localization) of 3D objects in LiDAR point clouds can be summarised as bounding box prediction and objectness prediction [14]. In this paper, we mainly survey the LiDAR-only paradigm, which takes advantage from accurate geo-referenced information. Overall, there are two ways for data representation in this paradigm: one detects and locates 3D objects directly from point clouds [118]; another first converts 3D points into regular grids, such as voxel grids or birds eye view images as well as front views, and then utilizes architectures in 2D detectors to extract object from images, the 2D detection results are finally back-projected into 3D space for final 3D object location estimation [50]. Fig.8 shows the representative network frameworks of the above-listed data representations.

*1) 3D objects detection (localization) from point clouds:* The challenges for 3D object detection from sparse and large-scale point clouds are concluded as:

- The detected objects only occupy a very limited amount of the whole input data.
- The 3D object centroid can be far from any surface point thus hard to regress accurately in one step [42].
- The missing of 3D object center points. As LiDAR sensors only capture surfaces of objects, 3D object centers are likely to be in empty space, far away from any point.

Thus, a common procedure of 3D object detection and localization from large-scale point clouds is composed of

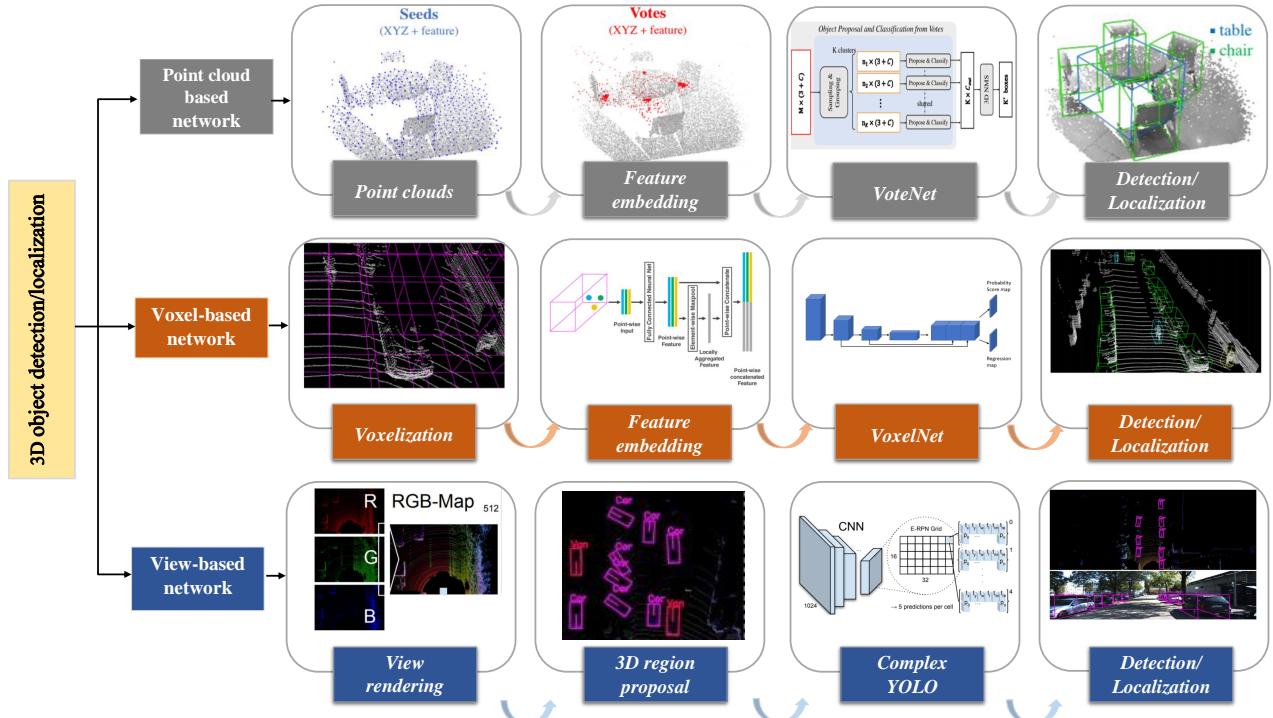


Fig. 8. DL architectures on 3D object detection/localization with three different data representations: point cloud based networks represented by VoteNet [42], voxel-based networks represented by VoxelNet [13], view-based networks represented by ComplexYOLO [116].

the following processes: firstly, the whole scene is roughly segmented, and then the coarse location of interest object is approximately proposed; secondly, the feature for each proposed region is extracted; finally, the localization and object class is predicted through a Bounding-Box Prediction Network [118, 119].

In [119], the PointNet++ [12] is applied to generate per-point feature within the whole input point clouds. Different from [118], each point is viewed as an effective proposal, which preserves the localization information. Then the localization and detection prediction is conducted based on the extracted point-based proposal features as well as local neighbor context information captured by increasing receptive field and input point features. This network preserves more accurate localization information but has higher computation cost for operating directly on point sets.

In [118], 3D CNN with three Conv layers and multiple FC layers is applied to learn the discriminate and robust features of objects. Then an intelligent eye window (EW) algorithm is applied to the scene. The label of point belong to the EW is predicted using the pre-trained 3D CNN. The evaluation result is then input to the deep Q-network (DQN) to adjust the size and position of EW. Then the new EW is evaluated by 3D CNN and DQN until the EW only contains one object. Different from the traditional bounding box of the region of interest (RoI), the EW can reshape its size and change the window center automatically, which is suitable for objects with different scales. Once the position of the object is located, the object in the input window is predicted with learned features. In [118], the object features are extracted based on 3D CNN

models and then fed into the residual RNN [120] for category labeling.

Qi et al. [42] proposed VoteNet a 3D object detection deep network based on Hough voting. The raw point clouds are input to PointNet++ [12] to learn point features. Based on these features, a group of seed points is sampled and generate votes from their neighbor features. These seeds are then gathered to cluster the object centers and generate bounding box proposals for a final decision. Compared with the above two architectures, VoteNet is robust to sparse and large-scale point clouds. Besides, it can localize the object center with high accuracy.

*2) 3D objects detection (localization) from regular voxel grid:* To better exploit CNNs, some approaches voxelize the 3D space into a voxel grid, which is represented by a scalar value such as occupancy or vector data extracted from voxels [8]. In [121, 122], the 3D space is first discretized into grids with a fixed size and then converted each occupied cell into a fixed-dimensional feature vector. Non-occupied cells without any points are represented with zero feature vectors. A binary occupancy and the mean and variance of the reflectance, as well as three shape factors are used to describe the feature vector. For simplicity, in [14], the voxelized grids are represented by length, width, height, and channels 4D array, and the binary value of one channel is used to represent the observation status of points in corresponding grids. Zhou et al. [13] voxelized the 3D point clouds along  $XYZ$  coordinates with predefined distance and grouped points in each grid. Then a voxel feature encoding (VFE) layer is proposed to achieve inter-point interaction within a voxel, by combining per-

TABLE IV  
SEGMENTATION RESULTS ON SEMANTIC3D REDUCED-8 DATASET

<b>Method</b>	<b>Input</b>	<b>Backbone</b>	<b>IoU</b>								<b>mIoU</b>	<b>OA (%)</b>	<b>Highlights</b>
			IoU1	IoU2	IoU3	IoU4	IoU5	IoU6	IoU7	IoU8			
SPGraph [105]	Graph & point cloud	PointNet [10]	0.974	0.926	0.879	0.44	0.932	0.31	0.635	0.762	0.732	94.0	3D point clouds are represented as a set of interconnected superpoints; Solve the big data challenge and the unevenly distributed point density problem.
MSDeepVoxNet [59]	voxels	VGG16 [87]	0.83	0.672	0.838	0.367	0.924	0.313	0.500	0.782	0.653	88.4	Extract multi-scale local features in a multi-scale neighborhood.
RF_MSSF [104]	point cloud	Random forest [117]	0.876	0.803	0.818	0.364	0.922	0.241	0.426	0.566	0.627	90.3	Define multi-scale neighborhoods in point clouds to extract features.
SEGCloud [113]	voxels	FCNN	0.839	0.66	0.86	0.405	0.911	0.309	0.275	0.643	0.613	88.1	Refine the labels generated at the voxel level for each point using Trilinear Interpolation; a FC CRF is connected with the FCNN to improve the segmentation result.
SnapNet [11]	images	CNN	0.82	0.773	0.797	0.229	0.911	0.184	0.373	0.644	0.591	88.6	Generate RGB and depth images from point cloud; Use CNN to conduct a pixel-wise labeling of each pair of 2D snapshot; Back-projection of the label predictions in the 3D space.
DeePr3SS [107]	images	VGG16 [87]	0.856	0.832	0.742	0.324	0.897	0.185	0.251	0.592	0.585	88.9	Generate view images from point clouds with a spread function into the image plane. Depth value and feature vectors of each projected point are stored and input to VGG16 for segmentation

point features and local neighbor features. The combination of multi-scale VFE layers enables this architecture to learn discriminative features from local shape information.

The voting scheme is adopted in [121, 122] to perform a sparse convolution on the voxelized grids. These grids, weighted by the convolution kernels as well as their surrounding cells in the receptive field, accumulate the votes from their neighbors by flipping the CNN kernel along each dimension and finally outputs the voting scores for potential interest objects. Based on that voting scheme, Engelcke et al. [122] then used a ReLU non-linearity to produce a novel sparse 3D representation of these grids. This process is iterated and stacked in conventional CNN operations and finally output the predicting scores for each proposal. However, the voting scheme has high computation during voting. Thus, modified region proposal networks (RPN) is employed by [13] in object detection to reduce computation. This RPN is composed of three blocks of Conv layers, which are used to downsample, filter features and upsample the input feature map and produce a probability score map, and a regression map for object detection and localization.

### 3) 3D objects detection (localization) from 2D views:

Some approaches also project LiDAR point clouds into 2D views. Such approaches are mainly composed of those two steps: first is the projection of 3D points; second is the object detection from projected images. There are several types of view generation methods to project 3D points into 2D images: BEV images [43, 116, 123, 124], front view images [123], spherical projections [50], and cylindrical projection [9].

Different from [50], in [43, 116, 123, 124], the point cloud data is split into grids with fixed size and then converted

to a birds eye view (BEV) image with corresponding three channels which encodes height, intensity, and density information. Considering the efficiency and performance, only the maximum height, the maximum intensity, and the normalized density among the grids are converted to a single birds-eye-view RGB-map [116]. In [125], only the maximum, median, and minimum height values are selected to represent the channels of the BEV image to exploit conventional 2D RGB deep models without modification. Dewan et al. [16] selected the range, intensity, and height values to represent three channels. In [8], the feature representation for each BEV pixel is composed of occupancy and reflectance value.

However, due to the sparsity of point clouds, the projection of point clouds to the 2D image plane produces a sparse 2D point map. Thus, Chen et al. [123] added front view representation to compensate for the missing information in BEV images. The point clouds are projected to a cylinder plane to produce dense front view images. In order to keep the 3D spatial information during projection, points are projected at multiview angles which are evenly selected on a sphere [50]. Pang et al. first discretized 3D points into cells with a fixed size. Then the scene is sampled to generate multiview images to construct positive and negative training samples. The benefits of this kind of dataset generation are that the spatial relationship and feature of the scene can be better exploited. However, this model is not robust to a new scene and cannot learn new features from a constructed dataset.

As for 2D object detectors, there exist enormous compelling deep models such as VGG-16 [87], Faster R-CNN [126]. In [23], a comprehensive survey of 2D detectors in object detection is concluded.

#### 4) Evaluation on 3D objects localization and detection:

In order to compare 3D objects localization and detection deep models, KITTI birds eye view benchmark and KITTI 3D object detection benchmark [60] are selected. As reported in [60], all non- and weakly-occluded (< 20%) objects which are neither truncated nor smaller than 40 px in height are evaluated. Truncated or occluded objects are not counted as false positives. Only a bounding box overlap of at least 50% results for pedestrian and cyclist, and 70% results for the car are considered for detection, localization, and orientation estimation measurements. Besides, this benchmark classified the difficulties of tasks into three types: easy, moderate, and hard.

Both the accuracy and execution time are compared to evaluate these algorithms because detection and localization in real-time are crucial for AVs [127]. For the localization task, the KITTI birds eye view benchmark is chosen as the evaluation benchmark, and the comparison results are shown in Table V. The 3D detection is evaluated on the KITTI 3D object detection benchmark. Table V shows the runtime and the average precision ( $AP_{3D}$ ) on the validation set. For each bounding box overlap, only 3D IoU exceeds 0.25 is considered as a valid localization/detection box [127].

### C. 3D object classification

Semantic object classification/recognition is crucial for safe and reliable driving of AVs in unstructured and uncontrolled real-world environments [67]. Existing 3D object detection are mainly focus on CAD data (e.g., ModelNet40 [30]) or RGBD data (e.g., NYUv2 [128]). However, these data have uniform point distribution, complete shapes, limited noise, occlusion and background clutter, which poses limit challenges for 3D classification compared with LiDAR point clouds [10, 12, 129]. Those compelling deep architectures applied on CAD data have been analyzed in the form of four types of data representations in section III. In this part, we mainly focus on the LiDAR data based deep models for the classification task.

1) *Volumetric architectures*: The voxelization of point clouds depends on the data spatial resolution, orientation, and the origin [67]. This operation which can provide enough recognizable information but not increase the computation cost is crucial for DL models. Thus, for LiDAR data, a voxel with spatial resolution such as  $(0.1m)^3$  is adopted in [67] to voxelize the input points. Then for each voxel, binary occupancy grid, density grid, hit grid are calculated to estimate its occupancy. The input layer, Conv layer, pooling layer, and FC layer are combined to construct the CNNs. Such architecture can exploit the spatial structure among data and extract global feature via pooling. However, the FC layer produces high computation cost and lose the spatial information between voxels. In [130], based on VoxNet [67], it takes a 3D voxel grid as input and contains two Conv layers with 3D filters followed by two FC layers. Different from other category-level classification tasks, they treated this task as a multi-task problem, where the orientation estimation and class label prediction are processed parallel.

For simplicity and efficiency, Zhi et al. [93, 131] adopted the binary grid of [67] to reduce the computation cost. However,

they only consider the voxels inside the surface, ignoring the difference between unknown and free space. Normal vectors, which contain geo-local position and orientation information, have been demonstrated stronger than binary grid in [132]. Similar to [130], the classification is treated as two tasks: voxel object class label predicting and its orientation prediction. To extract local and global features, there are two sub-tasks in the first task: the first sub-task is to predict the object label referencing the whole input shape while the second one predicts the object label with part of the shape. The orientation prediction is proposed to exploit the orientation augmentation scheme. The whole network is composed of three 3D Conv layers and two 3D max-pooling layers, which is lightweight and demonstrated robust to occlusion and clutter.

2) *Multi-view architectures*: The merit of view-based methods is their ability to exploit both local and global spatial relationships among points. Luo et al. [45] designed the three feature descriptors to extract local and global features from point clouds: the first one captures the horizontal geometric structure, the second one extracts vertical information, the last one provides complete spatial information. To better leverage the multi-view data representations, You et al. [91] integrated the merits of point cloud and multi-view data and achieved better results than MVCNN [76] in 3D classification. Besides, the high-level features extracted from view representations based on MVCNN [76] are embedded with an attention fusion scheme to compensate the local features extracted from point cloud data representations. Such attention-aware features are proved efficient in representing discriminative information of 3D data.

However, for different objects, the view generation process varies. Because the special attributes of objects can contribute to computation saving and accuracy improving. For example, in road marking extraction tasks, the elevation derived mainly from Z coordinate contributes little to the algorithm. But the road surface is actually a 2D structure. As a result, Wen et al. [47] directly projected 3D point clouds onto a horizontal plane and girded as a 2D image. Luo et al. [45] input the acquired three-view descriptors separately to capture low-level features to JointNet. Then this network learns high-level features by a convolutional operation based on the input features, and finally fuses the prediction scores. The whole framework is composed of five Conv layers, a spatial pyramid pooling (SPP) layer [133] and two FC layers and a reshape layer. The output results are fused through Conv layers and multi-view pooling layers. The well-designed view descriptors help the network achieve compelling results in object classification tasks.

Another representative architecture in 2D deep models is the encoder-decoder architecture. Due to the down-sampling and up-sampling can help to compress the information among pixels to extract the most representative features. In [47], Wen et al. proposed a modified U-net model to classify road markings. The point clouds data are first mapped into the intensity images. Then a hierarchical U-net module is applied to classify road markings by multi-scale clustering via CNNs. Due to such down-sampling and up-sampling is hard to preserve the fine-grained patterns, a GAN network is adopted to reshape small-size road markings, broken lane

TABLE V  
3D CAR LOCALIZATION PERFORMANCE ON KITTI BIRDS EYE VIEW BENCHMARK: AVERAGE PRECISION ( $AP_{loc}$  [%])

Method	Input	Times (s)	GPUs	Evaluation on AP (%)																Highlights		
				object detection						object localization												
				0.25			0.5			0.7			0.5			0.7						
E	M	H	E	M	H	E	M	H	E	M	H	E	M	H	E	M	H	E	M	H		
VeloFCN [14]	voxels	1	N/A	89.0	81.1	75.9	67.9	57.6	52.6	15.2	13.7	16.0	79.7	63.8	62.8	40.1	32.1	30.5	The voxelized grids are represented by length, width, height, and channels; The binary value of one channel is used to represent the observation status of points in corresponding grids.			
DOBEM [125]	images	0.6	Titan X	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	79.3	80.2	80.1	54.9	60.1	60.9	The maximum, median and minimum height values are selected to represent the BEV image channels to exploit conventional 2D RGB deep models without modification.			
MV3D [123]	images	0.36	Titan X	96.5	89.6	88.9	96.0	89.1	88.4	71.3	62.7	56.6	96.3	89.4	88.7	86.6	78.1	76.7	The point clouds are projected to a cylinder plane to produce dense front view images			
VoxelNet [13]	voxels	0.23	Titan X	N/A	N/A	N/A	N/A	N/A	N/A	82.0	65.5	62.9	N/A	N/A	N/A	89.6	84.8	78.6	Voxelize the 3D point clouds along XYZ coordinates with predefined distance and group points in each grid; Then a voxel feature encoding layer is proposed to achieve inter-point interaction within a voxel, by combining per-point features and local neighbor features.			
RT3D [127]	point cloud	0.09	Titan X	89.5	81.0	81.2	89.0	80.6	80.9	72.9	61.6	64.4	89.4	80.9	81.2	88.3	79.9	80.4	Propose a pre-RoI-pooling convolution technique that moves a majority of the convolution operations to the RoI pooling.			

lines and missing marking considering the expert context knowledge. This architecture exploits the efficiency of U-net and completeness of GAN to classify the road markings with high efficiency and accuracy.

3) *Evaluation on 3D objects classification:* There is limited published LiDAR point cloud benchmark specific for 3D objects classification task. Thus, the Sydney Urban Objects dataset is selected due to the performance of several state-of-the-art methods are available. The  $F_1$  score is used to evaluate these published algorithms [45], as shown in Table VI.

## VI. RESEARCH CHALLENGES AND OPPORTUNITIES

DL architectures developed in recent five years using LiDAR point clouds have made significant success in the field of autonomous driving detailing for 3D segmentation, detection, and classification tasks. However, there still exists a huge gap between cutting-edge results and human-level performance. Although there is much work to be done, we mainly summarize the remaining challenges specific for data, deep architectures, and tasks as follows:

1) *Multi-source Data Fusion:* To compensate the absence of 2D semantic, textual and incomplete information in 3D points, imagery, LiDAR point clouds, and radar data can be fused to provide accurate, geo-referenced, and information-rich cues for AVs' navigation and decision making [134]. Besides, there also exists a fusion between data acquired by low-end LiDAR (e.g., Velodyne HDL-16E) and high-end LiDAR (e.g., Velodyne HDL-64E) sensors. However, there exist several challenges in fusing these data: The first is the sparsity of point clouds causes the inconsistent and missing data when fusing multi-source data. The second is that the existing data fusion scheme using DL knowledge is processed in a separate line, which is not an end-to-end scheme. [41, 119, 135].

TABLE VI  
3D CLASSIFICATION PERFORMANCE ON THE SYDNEY URBAN OBJECTS DATASET [45]

Method	Input	$F_1$ score (%)	Highlights
VoxNet [67]	voxels	72.0	The input points are voxelized with spatial resolution; Binary occupancy grid, density grid, hit grid are calculated to estimate each voxel occupancy.
BV-CNNs [131]	voxels	75.5	Transform the inputs and weights in FC layers to binary values, which can potentially accelerate the networks by bit-wise
ORION [130]	voxels	77.8	Category level classification task is treated as a multitask problem, where the orientation estimation and class label prediction are processed parallel.
JointNet [45]	images	74.9	Three feature descriptors are proposed to extract local and global features from point clouds: the horizontal geometric structure, vertical information, complete spatial information.

2) *Robust Data Representation:* The unstructured and unordered data format [10, 12] poses a great challenge for robust 3D DL applications. Although there are several effective data representations such as voxels [67], point clouds [10, 12], graphs [74, 129], 2D views [78], or novel 3D data representations [136–138], there has not yet agreed on a robust and memory-efficient 3D data representation. For example, although voxels solve the ordering problem, the computation cost increases cubically with the increment of voxel resolution [30, 67]. As for point clouds and graphs, the permutation

invariance and the computation capability limit the processable quantity of points, which inevitably constrains the performance of the deep models [10, 74].

**3) Effective and More Efficient Deep Frameworks:** Due to the limitation of memory and computation facilities of the platform embedded in AVs, effective and efficient DL architectures are crucial for the wide application of automated AV systems. Although there are significant improvements in 3D DL models, such as PointNet [10], PointNet++ [12], PointCNN [71], DGCNN [74], RotationNet [78] and other work [52, 139–141]. Some limited models can achieve real-time segmentation, detection and classification tasks. Researches should focus on lightweight and compact architecture designing.

**4) Context Knowledge Extraction:** Due to the sparsity of point clouds and incompleteness of scanned objects, detailed context information for objects is not fully exploited. For example, the semantic contexts in traffic signs are crucial cues for AVs navigation, but existing deep models cannot extract such information completely from point clouds. Although multi-scale feature fusion approaches [142–144] have demonstrated significant improvements in context information extraction. Besides, GAN [47] can be utilized to improve the completeness of 3D point clouds. However, these frameworks cannot solve the sparsity and incompleteness problems for context information extraction in an end-to-end trainable way.

**5) Multi-task Learning:** The approaches related to LiDAR point clouds for AVs consist of several tasks, such as scene segmentation, object detection (e.g., cars, pedestrians, traffic lights, etc.) and classification (e.g., road markings, traffic signs). All these results are commonly fused together and reported to a decision system for final control [1]. However, there are few DL architectures combining these multiple LiDAR point cloud tasks together [15, 130]. Thus, the inherent information among them is not fully exploited and used to generalize better models with less computation.

**6) Weakly Supervised/Unsupervised Learning:** The existing state-of-art deep models are commonly constructed under supervised modes using labeled data with 3D objects bounding boxes or per-point segmentation masks [8, 74, 119]. However, there are some limitations for fully supervised models. First is the limited availability of high quality, large scale, and enormous general objects datasets and benchmarks. Second is the fully-supervised model generalization capability which is not robust to unseen or untrained objects. Weakly supervised [145] or unsupervised learning [146, 147] should be developed to increase the model’s generalization and solve the data absence problem.

## VII. CONCLUSION

In this paper, we have provided a systematic review of the state-of-the-art DL architectures using LiDAR point clouds in the field of autonomous driving for specific tasks such as segmentation, detection, and classification. Milestone 3D deep models and 3D DL applications on these three tasks have been summarized and evaluated with merits and demerits comparison. Research challenges and opportunities were listed to advance the potential development of DL in the field of autonomous driving.

## ACKNOWLEDGMENT

The authors would like to thank the Professors Jos Mancato Junior and Wesley Nunes Gonçalves for their carefully proofreading. Besides, we also would like to thank anonymous reviewers for their insightful comments and suggestions.

## REFERENCES

- [1] J. Janai, F. Güney, A. Behl, and A. Geiger, “Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art,” *arXiv:1704.05519*, 2017.
- [2] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt *et al.*, “Towards fully autonomous driving: Systems and algorithms,” in *IEEE Intell. Vehicles Symp.*, 2011, pp. 163–168.
- [3] J. Van Brummelen, M. O’Brien, D. Gruyer, and H. Najjaran, “Autonomous vehicle perception: The technology of today and tomorrow,” *Transp. Res. Part C Emerg. Technol.*, vol. 89, pp. 384–406, 2018.
- [4] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, “The apolloscape dataset for autonomous driving,” in *Proc. IEEE CVPR Workshops*, 2018, pp. 954–960.
- [5] R. P. D. Vivacqua, M. Bertozzi, P. Cerri, F. N. Martins, and R. F. Vassallo, “Self-localization based on visual lane marking maps: An accurate low-cost approach for autonomous driving,” *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 582–597, 2018.
- [6] F. Remondino, “Heritage recording and 3d modeling with photogrammetry and 3d scanning,” *Remote Sens.*, vol. 3, no. 6, pp. 1104–1138, 2011.
- [7] B. Wu, A. Wan, X. Yue, and K. Keutzer, “Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud,” in *IEEE ICRA*, 2018, pp. 1887–1893.
- [8] B. Yang, W. Luo, and R. Urtasun, “Pixor: Real-time 3d object detection from point clouds,” in *Proc. IEEE CVPR*, 2018, pp. 7652–7660.
- [9] B. Li, T. Zhang, and T. Xia, “Vehicle detection from 3d lidar using fully convolutional network,” *arXiv:1608.07916*, 2016.
- [10] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proc. IEEE CVPR*, 2017, pp. 652–660.
- [11] A. Boulch, B. Le Saux, and N. Audebert, “Unstructured point cloud semantic labeling using deep segmentation networks,” in *3DOR*, 2017.
- [12] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” in *Adv Neural Inf Process Syst*, 2017, pp. 5099–5108.
- [13] Y. Zhou and O. Tuzel, “Voxelpnet: End-to-end learning for point cloud based 3d object detection,” in *Proc. IEEE CVPR*, 2018, pp. 4490–4499.
- [14] B. Li, “3d fully convolutional network for vehicle detection in point cloud,” in *IEEE/RSJ IROS*, 2017, pp. 1513–1518.
- [15] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, “Volumetric and multi-view cnns for object classification on 3d data,” in *Proc. IEEE CVPR*, 2016, pp. 5648–5656.
- [16] A. Dewan, G. L. Oliveira, and W. Burgard, “Deep semantic classification for 3d lidar data,” in *IEEE/RSJ IROS*, 2017, pp. 3544–3549.
- [17] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, p. 436, 2015.
- [18] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, “Efficient processing of deep neural networks: A tutorial and survey,” *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.
- [19] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, “Deep learning for visual understanding: A review,” *Neurocomput*, vol. 187, pp. 27–48, 2016.
- [20] A. Voulovodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep learning for computer vision: A brief review,” *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, 2018.
- [21] L. Zhang, L. Zhang, and B. Du, “Deep learning for remote sensing data: A technical tutorial on the state of the art,” *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, 2016.
- [22] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep learning in remote sensing: A comprehensive review and list of resources,” *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, 2017.
- [23] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, “Deep learning for generic object detection: A survey,” *arXiv:1809.02165*, 2018.
- [24] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, “Object detection with deep learning: A review,” *IEEE Trans. Neural Netw. Learn. Syst.*, 2019.

- [25] 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*, 2017.
- [26] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomput.*, vol. 234, pp. 11–26, 2017.
- [27] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Process Mag.*, vol. 34, no. 4, pp. 18–42, 2017.
- [28] A. Ioannidou, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris, "Deep learning advances in computer vision with 3d data: A survey," *ACM CSUR*, vol. 50, no. 2, p. 20, 2017.
- [29] E. Ahmed, A. Saint, A. E. R. Shabayek, K. Cherenkova, R. Das, G. Gusev, D. Aouada, and B. Ottersten, "Deep learning advances on different 3d data representations: A survey," *arXiv:1808.01462*, 2018.
- [30] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proc. IEEE CVPR*, 2015, pp. 1912–1920.
- [31] L. Ma, Y. Li, J. Li, C. Wang, R. Wang, and M. Chapman, "Mobile laser scanned point-clouds for road object detection and extraction: A review," *Remote Sens.*, vol. 10, no. 10, p. 1531, 2018.
- [32] H. Guan, J. Li, S. Cao, and Y. Yu, "Use of mobile lidar in road information inventory: A review," *Int J Image Data Fusion*, vol. 7, no. 3, pp. 219–242, 2016.
- [33] E. Che, J. Jung, and M. J. Olsen, "Object recognition, segmentation, and classification of mobile laser scanning point clouds: A state of the art review," *Sensors*, vol. 19, no. 4, p. 810, 2019.
- [34] R. Wang, J. Peethambaran, and D. Chen, "Lidar point clouds to 3-d urban models : a review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 2, pp. 606–627, 2018.
- [35] X.-F. Hana, J. S. Jin, J. Xie, M.-J. Wang, and W. Jiang, "A comprehensive review of 3d point cloud descriptors," *arXiv:1802.02297*, 2018.
- [36] 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 Trans. Intell. Transp. Syst.*, 2019.
- [37] W. Liu, J. Sun, W. Li, T. Hu, and P. Wang, "Deep learning on point clouds and its application: A survey," *Sens.*, vol. 19, no. 19, p. 4188, 2019.
- [38] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schubert, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich *et al.*, "Speeding up semantic segmentation for autonomous driving," in *MLITS, NIPS Workshop*, vol. 1, 2016, p. 5.
- [39] A. Nguyen and B. Le, "3d point cloud segmentation: A survey," in *RAM*, 2013, pp. 225–230.
- [40] Q. Huang, W. Wang, and U. Neumann, "Recurrent slice networks for 3d segmentation of point clouds," in *Proc. IEEE CVPR*, 2018, pp. 2626–2635.
- [41] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proc. IEEE CVPR*, 2018, pp. 918–927.
- [42] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," *arXiv:1904.09664*, 2019.
- [43] J. Beltrán, C. Guindel, F. M. Moreno, D. Cruzado, F. García, and A. De La Escalera, "Birdnet: a 3d object detection framework from lidar information," in *ITSC*, 2018, pp. 3517–3523.
- [44] A. Kundu, Y. Li, and J. M. Rehg, "3d rfcn: Instance-level 3d object reconstruction via render-and-compare," in *Proc. IEEE CVPR*, 2018, pp. 3559–3568.
- [45] Z. Luo, J. Li, Z. Xiao, Z. G. Mou, X. Cai, and C. Wang, "Learning high-level features by fusing multi-view representation of mls point clouds for 3d object recognition in road environments," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 44–58, 2019.
- [46] Z. Wang, L. Zhang, T. Fang, P. T. Mathiopoulos, X. Tong, H. Qu, Z. Xiao, F. Li, and D. Chen, "A multiscale and hierarchical feature extraction method for terrestrial laser scanning point cloud classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2409–2425, 2015.
- [47] C. Wen, X. Sun, J. Li, C. Wang, Y. Guo, and A. Habib, "A deep learning framework for road marking extraction, classification and completion from mobile laser scanning point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 178–192, 2019.
- [48] T. Hackel, J. D. Wegner, and K. Schindler, "Joint classification and contour extraction of large 3d point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 231–245, 2017.
- [49] B. Kumar, G. Pandey, B. Lohani, and S. C. Misra, "A multi-faceted cnn architecture for automatic classification of mobile lidar data and an algorithm to reproduce point cloud samples for enhanced training," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 80–89, 2019.
- [50] G. Pang and U. Neumann, "3d point cloud object detection with multi-view convolutional neural network," in *IEEE ICPR*, 2016, pp. 585–590.
- [51] A. Tagliasacchi, H. Zhang, and D. Cohen-Or, "Curve skeleton extraction from incomplete point cloud," in *ACM Trans. Graph.*, vol. 28, no. 3, 2009, p. 71.
- [52] Y. Liu, B. Fan, S. Xiang, and C. Pan, "Relation-shape convolutional neural network for point cloud analysis," *arXiv:1904.07601*, 2019.
- [53] H. Huang, D. Li, H. Zhang, U. Ascher, and D. Cohen-Or, "Consolidation of unorganized point clouds for surface reconstruction," *ACM Trans. Graph.*, vol. 28, no. 5, p. 176, 2009.
- [54] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. J Rob Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [55] K. Jo, J. Kim, D. Kim, C. Jang, and M. Sunwoo, "Development of autonomous carpart ii: A case study on the implementation of an autonomous driving system based on distributed architecture," *IEEE Trans. Aerosp. Electron.*, vol. 62, no. 8, pp. 5119–5132, 2015.
- [56] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys, "Semantic3d. net: A new large-scale point cloud classification benchmark," *arXiv:1704.03847*, 2017.
- [57] D. Munoz, J. A. Bagnell, N. Vandapel, and M. Hebert, "Contextual classification with functional max-margin markov networks," in *Proc. IEEE CVPR*, 2009, pp. 975–982.
- [58] B. Vallet, M. Brédif, A. Serna, B. Marcotegui, and N. Paparoditis, "Terramobilita/igmus urban point cloud analysis benchmark," *Comput. Graph.*, vol. 49, pp. 126–133, 2015.
- [59] X. Roynard, J.-E. Deschaud, and F. Goulette, "Classification of point cloud scenes with multiscale voxel deep network," *arXiv:1804.03583*, 2018.
- [60] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE CVPR*, 2012, pp. 3354–3361.
- [61] M. De Deuge, A. Quadros, C. Hung, and B. Douillard, "Unsupervised feature learning for classification of outdoor 3d scans," in *ACRA*, vol. 2, 2013, p. 1.
- [62] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vision*, vol. 111, no. 1, pp. 98–136, 2015.
- [63] L. Yan, Z. Li, H. Liu, J. Tan, S. Zhao, and C. Chen, "Detection and classification of pole-like road objects from mobile lidar data in motorway environment," *Opt Laser Technol.*, vol. 97, pp. 272–283, 2017.
- [64] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *Int J Rob Res.*, vol. 36, no. 1, pp. 3–15, 2017.
- [65] L. Zhang, Z. Li, A. Li, and F. Liu, "Large-scale urban point cloud labeling and reconstruction," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 86–100, 2018.
- [66] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals using stereo imagery for accurate object class detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1259–1272, 2018.
- [67] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *IEEE/RSJ IROS*, 2015, pp. 922–928.
- [68] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *Adv Neural Inf Process Syst*, 2016, pp. 82–90.
- [69] G. Riegler, A. Osman Ulusoy, and A. Geiger, "Octnet: Learning deep 3d representations at high resolutions," in *Proc. IEEE CVPR*, 2017, pp. 3577–3586.
- [70] R. Klokov and V. Lempitsky, "Escape from cells: Deep kd-networks for the recognition of 3d point cloud models," in *Proc. IEEE ICCV*, 2017, pp. 863–872.
- [71] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "Pointcnn: Convolution on x-transformed points," in *NeurIPS*, 2018, pp. 820–830.
- [72] L. Yi, H. Su, X. Guo, and L. J. Guibas, "Syncspecnn: Synchronized spectral cnn for 3d shape segmentation," in *Proc. IEEE CVPR*, 2017, pp. 2282–2290.
- [73] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proc. IEEE CVPR*, 2017, pp. 3693–3702.
- [74] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *arXiv:1801.07829*, 2018.
- [75] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv:1710.10903*, 2017.

- [76] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, “Multi-view convolutional neural networks for 3d shape recognition,” in *Proc. IEEE ICCV*, 2015, pp. 945–953.
- [77] A. Dai and M. Nießner, “3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation,” in *ECCV*, 2018, pp. 452–468.
- [78] A. Kanezaki, Y. Matsushita, and Y. Nishida, “Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints,” in *Proc. IEEE CVPR*, 2018, pp. 5010–5019.
- [79] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE CVPR*, 2015, pp. 3431–3440.
- [80] G. Vosselman, B. G. Gorte, G. Sithole, and T. Rabbani, “Recognising structure in laser scanner point clouds,” *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. 46, no. 8, pp. 33–38, 2004.
- [81] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Adv Neural Inf Process Syst*, 2014, pp. 2672–2680.
- [82] M. Tatarchenko, A. Dosovitskiy, and T. Brox, “Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs,” in *Proc. IEEE ICCV*, 2017, pp. 2088–2096.
- [83] A. Miller, V. Jain, and J. L. Mundy, “Real-time rendering and dynamic updating of 3-d volumetric data,” in *Proc. GPGPU*, 2011, p. 8.
- [84] C. Wang, B. Samari, and K. Siddiqi, “Local spectral graph convolution for point set feature learning,” in *ECCV*, 2018, pp. 52–66.
- [85] L. Wang, Y. Huang, Y. Hou, S. Zhang, and J. Shan, “Graph attention convolution for point cloud semantic segmentation,” in *Proc. IEEE CVPR*, 2019, pp. 10 296–10 305.
- [86] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Adv Neural Inf Process Syst*, 2012, pp. 1097–1105.
- [87] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv:1409.1556*, 2014.
- [88] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proc. IEEE CVPR*, 2015, pp. 1–9.
- [89] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE CVPR*, 2016, pp. 770–778.
- [90] J.-C. Su, M. Gadella, R. Wang, and S. Maji, “A deeper look at 3d shape classifiers,” in *ECCV*, 2018, pp. 0–0.
- [91] H. You, Y. Feng, R. Ji, and Y. Gao, “Pvnet: A joint convolutional network of point cloud and multi-view for 3d shape recognition,” in *2018 ACM Multimedia Conference on Multimedia Conference*, 2018, pp. 1310–1318.
- [92] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [93] S. Zhi, Y. Liu, X. Li, and Y. Guo, “Toward real-time 3d object recognition: a lightweight volumetric cnn framework using multitask learning,” *Comput Graph*, vol. 71, pp. 199–207, 2018.
- [94] ——, “Lightnet: A lightweight 3d convolutional neural network for real-time 3d object recognition,” in *3DOR*, 2017.
- [95] A. Dai, D. Ritchie, M. Bokeloh, S. Reed, J. Sturm, and M. Nießner, “Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans,” in *Proc. IEEE CVPR*, 2018, pp. 4578–4587.
- [96] J. Li, B. M. Chen, and G. Hee Lee, “So-net: Self-organizing network for point cloud analysis,” in *Proc. IEEE CVPR*, 2018, pp. 9397–9406.
- [97] P. Huang, M. Cheng, Y. Chen, H. Luo, C. Wang, and J. Li, “Traffic sign occlusion detection using mobile laser scanning point clouds,” *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2364–2376, 2017.
- [98] H. Lei, N. Akhtar, and A. Mian, “Spherical convolutional neural network for 3d point clouds,” *arXiv:1805.07872*, 2018.
- [99] F. Engelmann, T. Kontogianni, J. Schult, and B. Leibe, “Know what your neighbors do: 3d semantic segmentation of point clouds,” in *ECCV*, 2018, pp. 0–0.
- [100] M. Weinmann, B. Jutzi, S. Hinz, and C. Mallet, “Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers,” *ISPRS J. Photogramm. Remote Sens.*, vol. 105, pp. 286–304, 2015.
- [101] E. Che and M. J. Olsen, “Fast ground filtering for tls data via scanline density analysis,” *ISPRS J. Photogramm. Remote Sens.*, vol. 129, pp. 226–240, 2017.
- [102] A.-V. Vo, L. Truong-Hong, D. F. Laefer, and M. Bertolotto, “Octree-based region growing for point cloud segmentation,” *ISPRS J. Photogramm. Remote Sens.*, vol. 104, pp. 88–100, 2015.
- [103] R. B. Rusu and S. Cousins, “Point cloud library (pcl),” in *2011 IEEE ICRA*, 2011, pp. 1–4.
- [104] H. Thomas, F. Goulette, J.-E. Deschaud, and B. Marcotegui, “Semantic classification of 3d point clouds with multiscale spherical neighborhoods,” in *3DV*, 2018, pp. 390–398.
- [105] L. Landrieu and M. Simonovsky, “Large-scale point cloud semantic segmentation with superpoint graphs,” in *Proc. IEEE CVPR*, 2018, pp. 4558–4567.
- [106] L. Wang, Y. Huang, J. Shan, and L. He, “Msnet: Multi-scale convolutional network for point cloud classification,” *Remote Sens.*, vol. 10, no. 4, p. 612, 2018.
- [107] F. J. Lawin, M. Danelljan, P. Tostberg, G. Bhat, F. S. Khan, and M. Felsberg, “Deep projective 3d semantic segmentation,” in *CAIP*, 2017, pp. 95–107.
- [108] D. Rethage, J. Wald, J. Sturm, N. Navab, and F. Tombari, “Fully-convolutional point networks for large-scale point clouds,” in *ECCV*, 2018, pp. 596–611.
- [109] W. Wang, R. Yu, Q. Huang, and U. Neumann, “Sgpn: Similarity group proposal network for 3d point cloud instance segmentation,” in *Proc. IEEE CVPR*, 2018, pp. 2569–2578.
- [110] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M.-H. Yang, and J. Kautz, “Splatnet: Sparse lattice networks for point cloud processing,” in *Proc. IEEE CVPR*, 2018, pp. 2530–2539.
- [111] Z. Wang, L. Zhang, L. Zhang, R. Li, Y. Zheng, and Z. Zhu, “A deep neural network with spatial pooling (dnnspp) for 3-d point cloud classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4594–4604, 2018.
- [112] J. Huang and S. You, “Point cloud labeling using 3d convolutional neural network,” in *ICPR*, 2016, pp. 2670–2675.
- [113] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, and S. Savarese, “Segcloud: Semantic segmentation of 3d point clouds,” in *3DV*, 2017, pp. 537–547.
- [114] J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” 2001.
- [115] R. Zhang, G. Li, M. Li, and L. Wang, “Fusion of images and point clouds for the semantic segmentation of large-scale 3d scenes based on deep learning,” *ISPRS J. Photogramm. Remote Sens.*, vol. 143, pp. 85–96, 2018.
- [116] M. Simony, S. Milzy, K. Amendey, and H.-M. Gross, “Complex-yolo: an euler-region-proposal for real-time 3d object detection on point clouds,” in *ECCV*, 2018, pp. 0–0.
- [117] A. Liaw, M. Wiener *et al.*, “Classification and regression by random-forest,” *R news*, vol. 2, no. 3, pp. 18–22, 2002.
- [118] L. Zhang and L. Zhang, “Deep learning-based classification and reconstruction of residential scenes from large-scale point clouds,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 1887–1897, 2018.
- [119] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, “Ipod: Intensive point-based object detector for point cloud,” *arXiv:1812.05276*, 2018.
- [120] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning transferable architectures for scalable image recognition,” in *Proc. IEEE CVPR*, 2018, pp. 8697–8710.
- [121] D. Z. Wang and I. Posner, “Voting for voting in online point cloud object detection,” in *RSS*, vol. 1, no. 3, 2015, pp. 10–15 607.
- [122] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, “Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks,” in *IEEE ICRA*, 2017, pp. 1355–1361.
- [123] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3d object detection network for autonomous driving,” in *Proc. IEEE CVPR*, 2017, pp. 1907–1915.
- [124] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, “Joint 3d proposal generation and object detection from view aggregation,” in *IEEE/RSJ IROS*, 2018, pp. 1–8.
- [125] S.-L. Yu, T. Westfchel, R. Hamada, K. Ohno, and S. Tadokoro, “Vehicle detection and localization on bird’s eye view elevation images using convolutional neural network,” in *IEEE SSRR*, 2017, pp. 102–109.
- [126] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Adv Neural Inf Process Syst*, 2015, pp. 91–99.
- [127] Y. Zeng, Y. Hu, S. Liu, J. Ye, Y. Han, X. Li, and N. Sun, “Rt3d: Real-time 3-d vehicle detection in lidar point cloud for autonomous driving,” *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3434–3440, 2018.
- [128] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *ECCV*, 2012, pp. 746–760.
- [129] Y. Xu, T. Fan, M. Xu, L. Zeng, and Y. Qiao, “Spidercnn: Deep learning on point sets with parameterized convolutional filters,” in *ECCV*, 2018, pp. 87–102.
- [130] N. Sedaghat, M. Zolfaghari, E. Amiri, and T. Brox, “Orientation-

- boosted voxel nets for 3d object recognition,” *arXiv:1604.03351*, 2016.
- [131] C. Ma, Y. Guo, Y. Lei, and W. An, “Binary volumetric convolutional neural networks for 3-d object recognition,” *IEEE Trans. Instrum. Meas.*, no. 99, pp. 1–11, 2018.
- [132] C. Wang, M. Cheng, F. Sohel, M. Bennamoun, and J. Li, “Normalnet: A voxel-based cnn for 3d object classification and retrieval,” *Neurocomput.*, vol. 323, pp. 139–147, 2019.
- [133] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.
- [134] M. Liang, B. Yang, S. Wang, and R. Urtasun, “Deep continuous fusion for multi-sensor 3d object detection,” in *ECCV*, 2018, pp. 641–656.
- [135] D. Xu, D. Anguelov, and A. Jain, “Pointfusion: Deep sensor fusion for 3d bounding box estimation,” in *Proc. IEEE CVPR*, June 2018.
- [136] T. He, H. Huang, L. Yi, Y. Zhou, and S. Soatto, “Geonet: Deep geodesic networks for point cloud analysis,” *arXiv:1901.00680*, 2019.
- [137] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, “Occupancy networks: Learning 3d reconstruction in function space,” *arXiv:1812.03828*, 2018.
- [138] T. Le and Y. Duan, “PointGrid: A Deep Network for 3D Shape Understanding,” *Proc. IEEE CVPR*, June 2018.
- [139] J. Li, Y. Bi, and G. H. Lee, “Discrete rotation equivariance for point cloud recognition,” *arXiv:1904.00319*, 2019.
- [140] D. Worrall and G. Brostow, “Cubenet: Equivariance to 3d rotation and translation,” in *ECCV*, 2018, pp. 567–584.
- [141] K. Fujiwara, I. Sato, M. Ambai, Y. Yoshida, and Y. Sakakura, “Canonical and compact point cloud representation for shape classification,” *arXiv:1809.04820*, 2018.
- [142] Z. Dong, B. Yang, F. Liang, R. Huang, and S. Scherer, “Hierarchical registration of unordered tls point clouds based on binary shape context descriptor,” *ISPRS J. Photogramm. Remote Sens.*, vol. 144, pp. 61–79, 2018.
- [143] H. Deng, T. Birdal, and S. Ilic, “Ppfnet: Global context aware local features for robust 3d point matching,” in *Proc. IEEE CVPR*, 2018, pp. 195–205.
- [144] S. Xie, S. Liu, Z. Chen, and Z. Tu, “Attentional shapecontextnet for point cloud recognition,” in *Proc. IEEE CVPR*, 2018, pp. 4606–4615.
- [145] Z. J. Yew and G. H. Lee, “3dfeat-net: Weakly supervised local 3d features for point cloud registration,” in *ECCV*, 2018, pp. 630–646.
- [146] J. Sauder and B. Sievers, “Context prediction for unsupervised deep learning on point clouds,” *arXiv:1901.08396*, 2019.
- [147] M. Shoef, S. Fogel, and D. Cohen-Or, “Pointwise: An unsupervised point-wise feature learning network,” *arXiv:1901.04544*, 2019.



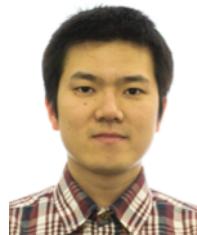
**Ying Li** received the M.Sc. degree in remote sensing from Wuhan University, China, in 2017. She is currently working toward the Ph.D. degree with the Mobile Sensing and Geodata Science Laboratory, Department of Geography and Environmental Management, University of Waterloo, ON, Canada.

Her research interests include autonomous driving, mobile laser scanning, intelligent processing of point clouds, geometric and semantic modeling, and augmented reality.



**Lingfei Ma** (S18) received the B.Sc. and M.Sc. degrees in geomatics engineering from the University of Waterloo, Waterloo, ON, Canada, in 2015 and 2017, respectively. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the Mobile Sensing and Geodata Science Laboratory, Department of Geography and Environmental Management, University of Waterloo.

His research interests include autonomous driving, mobile laser scanning, intelligent processing of point clouds, 3-D scene modeling, and machine learning.



**Zilong Zhong** (S15) Zilong Zhong received the Ph.D. in systems design engineering, specialized in machine learning and intelligence, from the University of Waterloo, Canada, in 2019. He is a postdoctoral fellow with the School of Data and Computer Science, Sun Yat-Sen University, China. His research interests include computer vision, deep learning, graph models and their applications involving large-scale image analysis.



**Fei Liu** received the B.Eng. degree from Yanshan University, China, in 2011. Since then, she has been working in the sectors of vehicular electronics and control, artificial intelligence, deep learning, FPGA, advanced driver assistance systems, and automated driving.

She is currently working at Xilinx Technology Beijing Limited, Beijing, China, focusing on development of automated driving technologies and data centers.



**Dongpu Cao** (M08) received the Ph.D. degree from Concordia University, Canada, in 2008.

He is the Canada Research Chair in Driver Cognition and Automated Driving, and currently an Associate Professor and Director of Waterloo Cognitive Autonomous Driving (CogDrive) Lab at University of Waterloo, Canada. His current research focuses on driver cognition, automated driving and cognitive autonomous driving. He has contributed more than 200 publications, 2 books and 1 patent. He received the SAE Arch T. Colwell Merit Award in 2012, and three Best Paper Awards from the ASME and IEEE conferences.



**Jonathan Li** (00’ M- 11’ SM) received the Ph.D. degree in geomatics engineering from the University of Cape Town, South Africa.

He is currently a Professor with the Departments of Geography and Environmental Management and Systems Design Engineering, University of Waterloo, Canada. He has coauthored more than 420 publications, more than 200 of which were published in refereed journals, including the *IEEE Transactions on Geoscience and Remote Sensing*, *IEEE Transactions on Intelligent Transportation Systems*, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *ISPRS Journal of Photogrammetry and Remote Sensing*, *Remote Sensing of Environment*, as well as leading artificial intelligence and remote sensing conferences including CVPR, AAAI, IJCAI, IGARSS and ISPRS. His research interests include information extraction from LiDAR point clouds and from earth observation images.



**Michael A. Chapman** received the Ph.D. degree in photogrammetry from Laval University, Canada.

He was a Professor with the Department of Geomatics Engineering, University of Calgary, Canada, for 18 years. Currently, he is a Professor of geomatics engineering with the Department of Civil Engineering, Ryerson University, Canada. He has authored or coauthored over 200 technical articles. His research interests include algorithms and processing methodologies for airborne sensors using GNSS/IMU, geometric processing of digital imagery in industrial environments, terrestrial imaging systems for transportation infrastructure mapping, and algorithms and processing strategies for biometrolgy applications.