

Deep Learning based 3D Segmentation: A Survey

YONG HE, HONGSHAN YU, XIAOYAN LIU, ZHENGENG YANG, WEI SUN, YAONAN WANG,
 QIANG FU, YANMEI ZOU, Hunan University, China
 AJMAL MIAN, University of Western Australia, Australia

3D object segmentation is a fundamental and challenging problem in computer vision with applications in autonomous driving, robotics, augmented reality and medical image analysis. It has received significant attention from the computer vision, graphics and machine learning communities. Traditionally, 3D segmentation was performed with hand-crafted features and engineered methods which failed to achieve acceptable accuracy and could not generalize to large-scale data. Driven by their great success in 2D computer vision, deep learning techniques have recently become the tool of choice for 3D segmentation tasks as well. This has led to an influx of a large number of methods in the literature that have been evaluated on different benchmark datasets. This paper provides a comprehensive survey of recent progress in deep learning based 3D segmentation covering over 150 papers. It summarizes the most commonly used pipelines, discusses their highlights and shortcomings, and analyzes the competitive results of these segmentation methods. Based on the analysis, it also provides promising research directions for the future.

CCS Concepts: • Computing methodologies → Computer vision tasks; Computer vision problems.

Additional Key Words and Phrases: 3D data, 3D semantic segmentation, 3D instance segmentation, 3D part segmentation, deep learning, unmanned system, medical diagnosis

ACM Reference Format:

Yong He, Hongshan Yu, Xiaoyan Liu, Zhengeng Yang, Wei Sun, Yaonan Wang, Qiang Fu, Yanmei Zou and Ajmal Mian. 2021. Deep Learning based 3D Segmentation: A Survey. In *Proceedings of . ACM*, New York, NY, USA, 36 pages. <https://doi.org/10.1145/nnnnnnn>.

1 INTRODUCTION

Segmentation of 3D scenes is a fundamental and challenging problem in computer vision as well as graphics. The objective of 3D segmentation is to build computational techniques that predict the fine-grained labels of objects in a 3D scene for a wide range of applications such as autonomous driving, mobile robots, industrial control, augmented reality and medical image analysis. As illustrated in Fig. 1 second row, 3D segmentation can be divided into three types: semantic, instance and part segmentation. Semantic segmentation aims to predict object class labels such as table and chair. Instance segmentation additionally distinguishes between different instances of the same class labels e.g. table

This research was partially supported by the National Natural Science Foundation of China under Grant 61973106 and Grant U2013203, the Australian Research Council Discovery Grant DP190102443, and the China Scholarship Council(CSC).

Authors Addresses: Yong He, Hongshan Yu (corresponding author), Xiaoyan Liu, Zhengeng Yang, Wei Sun, Yaonan Wang, Qiang Fu, Yanmei Zou, The Hunan University, 2 Lushan South Road, Yuelu District, Changsha, Hunan, China; emails: {h.yong, Yuhongshan, xiaoyan.liu, yzg050215, wei_sun, yaonan, fu_qiang_hnu, zouyanmei}@hnu.edu.cn; Ajmal Mian, The University of Western Australia, 35 Stirling Hwy, WA, 6009, Perth, Australia; emails: ajmal.mian@uwa.edu.au.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

Manuscript submitted to ACM

one/two and chair one/two. Part segmentation aims to decompose instances further into their different components such as armrests, legs and backrest of the same chair. Compared to 2D segmentation, 3D segmentation gives a more comprehensive understanding of a scene because 3D data (e.g. RGB-D, point clouds, projected images, voxels, and mesh) contain richer geometric, shape, and scale information with less background noise.

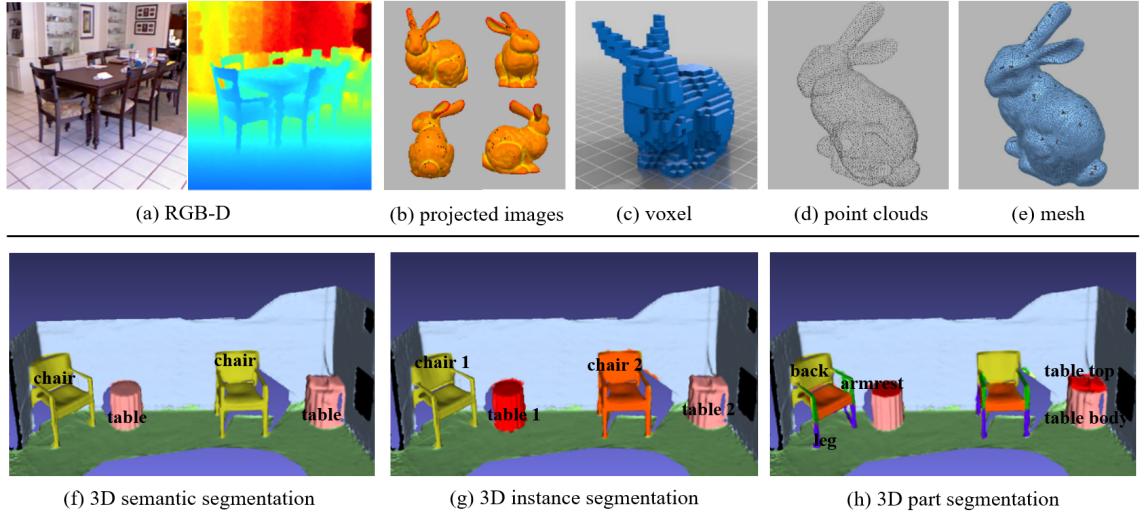


Fig. 1. **Top:** The main five types of 3D data: (a) RGB-D image, (b) projected images, (c) voxels, (d) point clouds, and (e) mesh. **Bottom:** 3D segmentation Types: (f) 3D semantic segmentation, (g) 3D instance segmentation, and (h) 3D part segmentation.

Recently, deep learning techniques have dominated many research areas including computer vision, speech recognition, and natural language processing. Motivated by its success in learning powerful features, deep learning for 3D segmentation has also attracted a growing interest from the research community over the past decade. However, 3D deep learning methods still face many unsolved challenges. For example, features from RGB and depth channels are difficult to fuse. The irregularity of point clouds makes it difficult to exploit local features and converting them to high-resolution voxels brings a huge computational burden.

This paper provides a comprehensive survey of recent progress in deep learning methods for 3D segmentation. It focuses on analyzing commonly used building blocks, convolution kernels and complete architectures pointing out the pros and cons in each case. The survey covers over 150 representative papers published in the last five years. Although some notable 3D segmentation surveys have been released including RGB-D semantic segmentation [31], point clouds segmentation [158],[37],[88],[4],[103],[54], these surveys do not comprehensively cover all 3D data types and typical application domains. Most importantly, these surveys do not focus on 3D segmentation but give a general survey of deep learning from point clouds [37],[88],[4],[103],[54]. Given the importance of the three segmentation tasks, this paper focuses exclusively on deep learning techniques for 3D segmentation. The contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the first survey paper to comprehensively cover deep learning methods on 3D segmentation using different 3D data representations, including RGB-D, projected images, voxels, point clouds, mesh, and 3D video.

- We provide an in-depth analysis of the relative advantages and disadvantages of different types of 3D data segmentation methods.
- Unlike existing reviews, we focus on deep learning methods designed specifically for 3D segmentation and also discuss typical application domains.
- We provide comprehensive comparisons of existing methods on several public benchmark 3D datasets, draw interesting conclusions and identify promising future research directions.

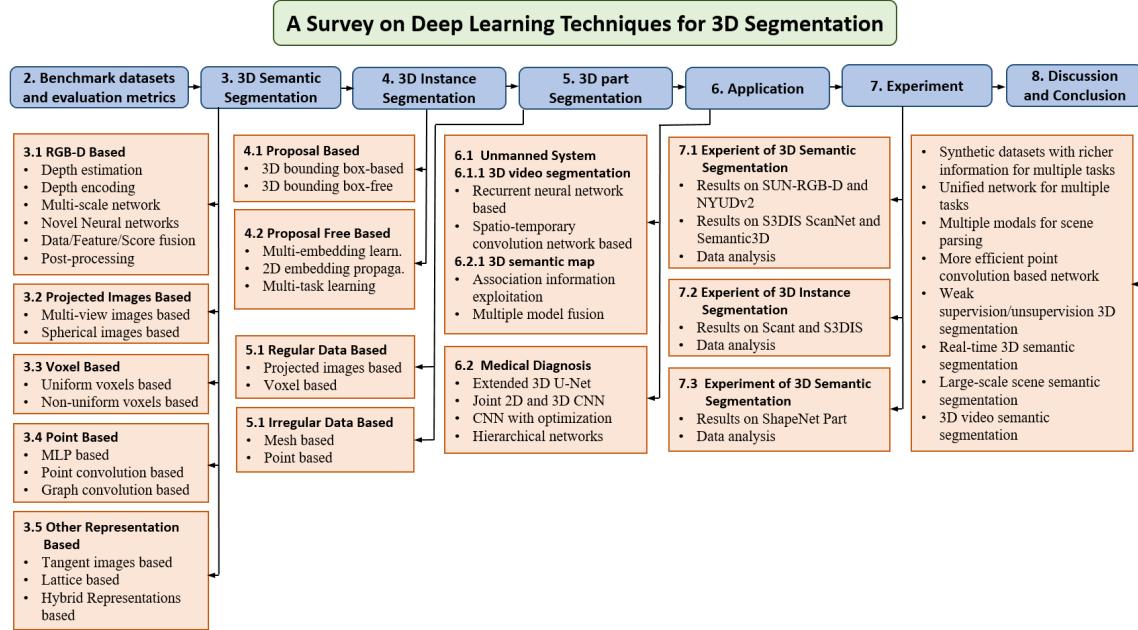


Fig. 2. Complete overview of the paper.

Figure 2 shows a snapshot of how the rest of the paper is organized. Section 2 introduces some basic knowledge and background concepts, including popular 3D datasets and evaluation metrics for 3D segmentation. Section 3 reviews methods for 3D semantic segmentation whereas Section 4 reviews methods for 3D instance segmentation. Section 5 provides a survey of existing methods for 3D part segmentation. Section 6 reviews the 3D segmentation methods used in some common application areas including unmanned systems and medical diagnosis. Section 7 presents performance comparison between 3D segmentation methods on several popular datasets, and gives corresponding data analysis. Finally, Section 8 identifies promising future research directions and concludes the paper.

2 BENCHMARK DATASETS AND EVALUATION METRICS

This section introduces some terminologies and background concepts, including popular 3D segmentation datasets and evaluation metrics to help the reader easily navigate through the field of 3D segmentation.

2.1 3D Segmentation Datasets

Datasets are critical to train and test 3D segmentation algorithms using deep learning. However, it is cumbersome and expensive to privately gather and annotate datasets as it needs domain expertise, high quality sensors and processing equipment. Thus, building on public datasets is an ideal way to reduce the cost. Following this way has another advantage for the community that it provides a fair comparison between algorithms. Table 1 summarizes some of the most popular and typical datasets with respect to the sensor type, data size and format, scene class and annotation method.

These datasets are acquired for *3D semantic segmentation* by different type of sensors, including RGB-D cameras [123],[124],[127],[49],[20], mobile laser scanner [120],[3], static terrestrial scanner [39] and unreal engine [7],[155] and other 3D scanners [1],[10]. Among these, the ones obtained from unreal engine are synthetic datasets [7][155] that do not require expensive equipment or annotation time. These are also rich in categories and quantities of objects. Synthetic datasets have complete 360 degree 3D objects with no occlusion effects or noise compared to the real-world datasets which are noisy and contain occlusions [123],[124],[127],[49],[20],[120],[3],[1],[39],[10]. For *3D instance segmentation*, there are limited 3D datasets, such as ScanNet[20] and S3DIS[1]. These two datasets contain scans of real-world indoor scenes obtained by RGB-D cameras or Matterport separately. For *3D part segmentation*, the Princeton Segmentation Benchmark (PSB)[12], COSEG [147] and ShapeNet [169] are three of the most popular datasets. Below, we introduce five famous segmentation datasets in detail, including S3DIS [1], ScanNet [20], Semantic3D [39], SemanticKITTI [10] and ShapeNet [169]. Some examples with annotation from these datasets are shown in Fig. 3.

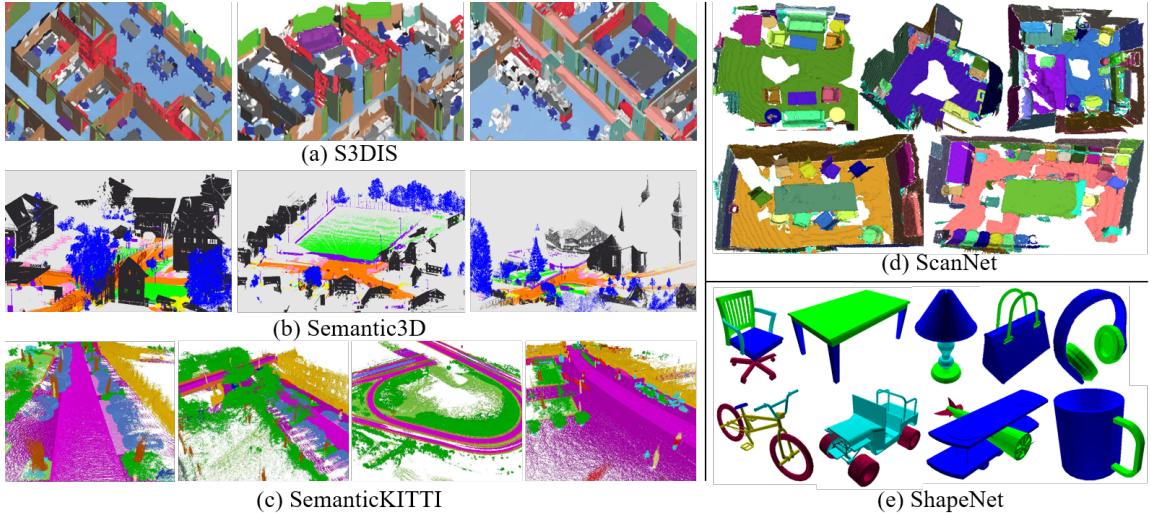


Fig. 3. Annotated examples from (a) S3DIS, (b) Semantic3D, (c) SemanticKITTI for 3D semantic segmentation, (d) ScanNet for 3D instance segmentation, and (e) ShapeNet for 3D part segmentation. See Table 1 for a summary of these datasets.

S3DIS: In this dataset, the complete point clouds are obtained without any manual intervention using the Matterport scanner. The dataset consists of 271 rooms belonging to 6 large-scale indoor scenes from 3 different buildings (total of 6020 square meters). These areas mainly include offices, educational and exhibition spaces, and conference rooms etc.

Semantic3D comprises a total of around 4 billion 3D points acquired with static terrestrial laser scanners, covering up to $160 \times 240 \times 30$ meters in real-world 3D space. Point clouds belong to 8 classes (e.g. urban and rural) and contain 3D coordinates, RGB information, and intensity. Unlike 2D annotation strategies, 3D data labeling is easily amenable to over-segmentation where each point is individually assigned to a class label.

SemanticKITTI is a large outdoor dataset containing detailed point-wise annotation of 28 classed. Building on the KITTI vision benchmark [32], SemanticKITTI contains annotations of all 22 sequences of this benchmark consisting of 43K scans. Moreover, the dataset contains labels for the complete horizontal 360 field-of-view of the rotating laser sensor.

ScanNet dataset is particularly valuable for research in scene understanding as its annotations contain estimated calibration parameters, camera poses, 3D surface reconstruction, textured meshes, dense object level semantic segmentation, and CAD models. The dataset comprises annotated RGB-D scans of real-world environments. There are 2.5M RGB-D images in 1513 scans acquired in 707 distinct places. After RGB-D image processing, annotation HITs (Human Intelligence Tasks) were performed using the Amazon Mechanical Turk.

ShapeNet dataset has a novel scalable method for efficient and accurate geometric annotation of massive 3D shape collections. The novel technical innovations explicitly model and lessen the human cost of the annotation effort. Researchers create detailed point-wise labeling of 31963 models in shape categories in ShapeNetCore and combine feature-based classifiers, point-to-point correspondences, and shape-to-shape similarities into a single CRF optimization over the network of shapes.

Table 1. Summary of popular datasets for 3D segmentation datasets including the sensor, type, size, object class, number of classes (shown in brackets), and annotation method. S←synthetic environment. R←real-world environment. Kf←thousand frames. s←scan. Mp←million points. the symbol ‘-’ means information unavailable.

Method[Reference]	Sensors	Type	Size	Scene class(Number)	Annotation method
datasets for 3D semantic segmentation					
NYUv1[123]	Microsoft Kinect v1	R	2347f	bedroom, cafe, kitchen, etc. (7)	Condition Random Field-based model
NYUv2[124]	Microsoft Kinect v1	R	1449f	bedroom, cafe, kitchen, etc. (26)	2D LabelMe-style annotation from AMK
SUN RGB-D[127]	RealSense, Xtion LIVE PRO, MKv1/2	R	10355f	Objects, room layouts, etc.(47)	2D/3D polygons +3D bounding box
SceneNN[49]	Asus Xtion PRO, MK v2	R	100s	Bedroom, office, apartment, etc.(-)	3D Labels project to 2D frames
RueMonge2014[114]	-	R	428s	window, wall, balcony, door, etc(7)	Multi-view semantic labelling + CRF
ScanNet[20]	Occipital structure sensor	R	2.5Mf	Office, apartment, bathroom, etc(19)	3D labels project to 2D frames
S3DIS[1]	Matterport camera	R	70496f	Conference rooms, offices, etc(11)	Hierarchical labeling
Semantic3D[39]	Terrestrial laser scanner	R	1660Mp	Farms, town hall, sport fields, etc (8)	Three baseline methods
PL3D [120]	Velodyne HDL-32E LiDAR	R	143.1Mp	Ground, vehicle, humman, etc (50)	Human labeling
SemanticKITTI[3]	Velodyne HDL-64E	R	43Ks	Ground, vehicle, humman, etc(28)	Multi-scans semantic labelling
Matterport3D[10]	Matterport camera	R	194.4Kf	various rooms (90)	Hierarchical labeling
HoME[7]	Planner5D platform	S	45622f	rooms, object and etc.(84)	SSCNet+ a short text description
House3D[155]	Planner5D platform	S	45622f	rooms, object and etc.(84)	SSCNet+3 ways
datasets for 3D instance segmentation					
ScanNet[20]	Occipital structure sensor	R	2.5Mf	Office, apartment, bathroom, etc(19)	3D labels project to 2D frames
S3DIS[1]	Matterport camera	R	70496f	Conference rooms, offices, etc(11)	Active learning method
datasets for 3D part segmentation					
ShapeNet[169]	-	R	31963s	Transportation, tool, etc.(16)	Propagating human label to shapes
PSB[12]	Amazon's Mechanical Turk	R	380s	Human,cup, glasses airplane,etc(19)	Interactive segmentation tool
COSEG[147]	-	R	1090s	Vase, lamp, guitar, etc (11)	semi-supervised learning method

2.2 Evaluation Metrics

Different evaluation metrics can assert the validity and superiority of segmentation methods including the execution time, memory footprint and accuracy. However, few authors provide detailed information about the execution time and memory footprint of their method. This paper introduces the accuracy metrics mainly.

For *3D semantic segmentation*, mean class Accuracy (mAcc) and mean class Intersection over Union (mIoU) are the most frequently used metrics to measure the accuracy of segmentation methods. For the sake of explanation, we assume that there are a total of $K + 1$ classes, and p_{ij} is the minimum unit (e.g. pixel, voxel, mesh, point) of class i implied to belong to class j . In other words, p_{ii} represents true positives, while p_{ij} and p_{ji} represent false positives and false negatives respectively.

Overall Accuracy (OAcc) is a simple metric that computes the ratio between the number of truly classified samples and the total number of samples.

$$OAcc = \sum_{i=0}^K \frac{p_{ii}}{\sum_{j=0}^K p_{ij}}$$

Mean Accuracy (mAcc): It is extension of OAcc, computing OAcc in a per-class and then averaging over the total number of classes K .

$$mAcc = \frac{1}{K+1} \sum_{i=0}^K \frac{p_{ii}}{\sum_{j=0}^K p_{ij}}$$

Mean Intersection over Union (mIoU) is a standard metric for semantic segmentation. It computes the intersection ratio between ground truth and predicted value averaged over the total number of classes K .

$$mIoU = \frac{1}{K+1} \sum_{i=0}^K \frac{p_{ii}}{\sum_{j=0}^K p_{ij} + \sum_{i=0}^K p_{ji} - p_{ii}}$$

For *3D instance segmentation*, Average Precision (AP) and mean class Average Precision (mAP) are also frequently used. Assuming $L_I, I \in [0, K]$ instance in every class, and c_{ij} is the amount of point of instance i inferred to belong to instance j ($i = j$ represents correct and $i \neq j$ represents incorrect segmentations). Average Precision (AP) is another simple metric for segmentation that computes the ratio between true positives and the total number of positive samples.

$$AP = \sum_{I=0}^K \sum_{i=0}^{L_I} \frac{c_{ii}}{c_{ii} + \sum_{j=0}^{L_I} c_{ij}}$$

Mean Average precision (mAP) is an extension of AP which computes per-class AP and then averages over the total number of classes K .

$$mAP = \frac{1}{K+1} \sum_{I=0}^K \sum_{i=0}^{L_I} \frac{c_{ii}}{c_{ii} + \sum_{j=0}^{L_I} c_{ij}}$$

For *3D part segmentation*, overall average category Intersection over Union (Cat.mIoU) and overall average instance Intersection over Union (Ins.mIoU) are most frequently used. For the sake of explanation, we assume $M_J, J \in [0, L_I]$ parts in every instance, and q_{ij} as the total number of points in part i inferred to belong to part j . Hence, q_{ii} represents the number of true positive, while q_{ij} and q_{ji} are false positives and false negative respectively. Overall average category Intersection over Union (Cat. mIoU) is an evaluation metric for part segmentation that measures the mean IoU averaged across K classes.

$$Cat.mIoU = \frac{1}{K+1} \sum_{I=0}^K \sum_{J=0}^{L_I} \sum_{i=0}^{M_J} \frac{q_{ii}}{\sum_{j=0}^{M_J} q_{ij} + \sum_{i=0}^{M_J} q_{ji} - q_{ii}}$$

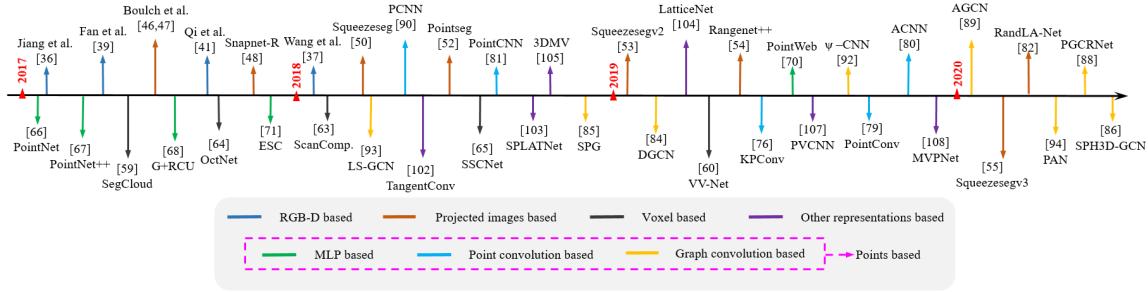


Fig. 4. Milestones of deep learning based 3D semantic segmentation methods.

Overall average instance Intersection over Union (Ins. mIoU), for part segmentation, measures the mean IoU across all instances.

$$\text{Ins.} \text{mIoU} = \frac{1}{\sum_{I=0}^K L_I + 1} \sum_{I=0}^K \sum_{J=0}^{L_I} \sum_{i=0}^{M_J} \frac{q_{ii}}{\sum_{j=0}^{M_j} q_{ij} + \sum_{i=0}^{M_j} q_{ji} - q_{ii}}$$

3 3D SEMANTIC SEGMENTATION

Many deep learning methods on 3D semantic segmentation have been proposed in the literature. These methods can be divided into five categories according to the data representation used, namely, RGB-D image based, projected images based, voxel based, point based and other representations based. Point based methods can be further categorized, based on the network architecture, into multiple layer perceptron (MLP) based, point convolution based and graph convolution based methods. Figure 4 shows the milestones of deep learning on 3D semantic segmentation in recent years.

3.1 RGB-D Based

The depth map in an RGB-D image contains geometric information about the real-world which is useful to distinguish foreground objects from background, hence providing opportunities to improve the segmentation accuracy. In this category, generally the classical two-channel network is used to extract features from RGB and depth images separately. However, this simple framework is not powerful enough to extract rich and refined features. To this end, researchers have integrated several additional modules into the above simple two-channel framework to improve the performance by learning rich *context* and *geometric* information that are crucial for semantic segmentation. These modules can be roughly divided into six categories: multi-task learning, depth encoding, multi-scale network, novel neural network architectures, data/feature/score level fusion and post-processing (see Fig.5). RGB-D image based semantic segmentation methods are summarized in Table 2.

Multi-tasks learning: *Depth estimation* and semantic segmentation are two fundamental challenging tasks in computer vision. These tasks are also somewhat related as depth variation within an object is small compared to depth variation between different objects. Hence, many researchers choose to unite depth estimation task and semantic segmentation task. From the view of relationship of the two tasks, there are two main types of multi-task learning framework, cascade and parallel framework.

As for the cascade framework, depth estimation task provides depth images for semantic segmentation task. For example, Cao et al. [8] used the deep convolutional neural fields (DCNF) introduced by Liu et al. [85] for depth estimation. The estimated depth images and RGB images are fed into a two-channel FCN for semantic segmentation. Similarly, Guo

et al. [36] adopted the deep network proposed by Ivanecky [55] for automatic generating depth images from single RGB images, and then proposed a two-channel FCN model on the image pair of RGB and predicted depth map for pixel labeling.

The cascade framework performs depth estimation and semantic segmentation separately, which is simultaneously unable to perform end-to-end training for two tasks. Consequently, depth estimation task does not get any benefit from semantic segmentation task. In contrast, the *parallel* framework performs these two tasks in an unify network, which allows two tasks get benefits each other. For instance, Wang et al. [141] used Joint Global CNN to exploit pixel-wise depth values and semantic labels from RGB images to provide accurate global scale and semantic guidance. As well as, they use Joint Region CNN to extract region-wise depth values and semantic map from RGB to learn detailed depth and semantic boundaries. Mousavian et al. [101] presented a multi-scale FCN comprising five streams that simultaneously explore depth and semantic features at different scales, where the two tasks share the underlying feature representation. Liu et al. [87] proposed a collaborative deconvolutional neural network(C-DCNN) to jointly model the two tasks. However, the quality of depth maps estimated from RGB images is not as good as the one acquired directly from depth sensors. This multi-task learning pipeline has been gradually abandoned in RGB-D semantic segmentation.

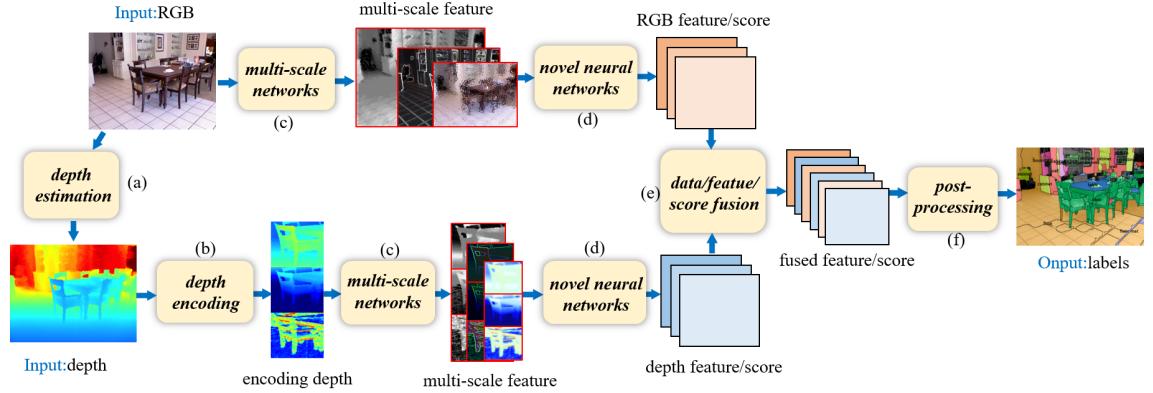


Fig. 5. Typical two-channel framework with six improvement modules, including (a) multi-tasks learning, (b) depth encoding, (c) multi-scale network, (d) novel neural network architecture, (e) feature/score level fusion, and (f) post-processing.

Depth Encoding: Conventional 2D-CNNs are unable to exploit rich geometric features from raw depth images. An alternative way is to encode raw depth images into other representations that are suitable to 2D-CNN. Hoft et al.[46] used a simplified version of the histogram of oriented gradients (HOG) to represent depth channel from RGB-D scenes. Gupta et al.[38] and Aman et al.[82] calculated three new channels named horizontal disparity, height above ground and angle with gravity (HHA) from the raw depth images. Liu et al.[86] point out a limitation of HHA that some scenes may not be enough horizontal and vertical planes. Hence, they propose a novel gravity direction detection method with vertical lines fitted to learn better representation. Hazirbas et al.[42] also argue that HHA representation has a high computational cost and contains less information than the raw depth images. They propose an architecture called FuseNet that consists of two encoder-decoder branches, including a depth branch and an RGB branch, which directly encodes depth information with a lower computational load.

Multi-scale Network: The context information learned by multi-scale networks is useful for small objects and detailed region segmentation. Couprie et al.[19] applied a multi-scale convolutional network to learn features directly

from the RGB images and the depth images. Aman et al.[111] proposed a multi-scale deep ConvNet for segmentation where the coarse predictions of VGG16-FC net are up sampled in a Scale-2 module and then concatenated with the low-level predictions of VGG-M net in Scale-1 module to get both high and low level features. However, this method is sensitive to clutter in the scene resulting in output errors. Lin et al.[82] exploit the fact that lower scene-resolution regions have higher depth, and higher scene-resolution regions have lower depth. They use depth maps to split the corresponding color images into multiple scene-resolution regions, and introduce context-aware receptive field (CaRF) which focuses on semantic segmentation of certain scene-resolution regions. This makes their pipeline a multi-scale network.

Novel Neural Network Architectures: Given the fixed grid computation of CNNs, their ability to process and exploit geometric information is limited. Therefore, researchers have proposed other novel neural network architectures to better exploit geometric features and the relationships between RGB and depth images. These architectures can be divided into four main categories.

Improved 2D Convolutional Neural Networks (2D-CNNs) Inspired from cascaded feature networks [82], Jiang et al.[61] proposed a novel Dense-Sensitive Fully Convolutional Neural Network (DFCN) which incorporates depth information into the early layers of the network using feature fusion tactics. This is followed by several dilated convolutional layers for context information exploitation. Similarly, Wang et al.[144] proposed a depth-aware 2D-CNN by introducing two novel layers, depth aware convolution layer and depth-aware pooling layer, which are based on the prior that pixels with the same semantic label and similar depth should have more impact on one another.

Deconvolutional Neural Networks(DeconvNets) are a simple yet effective and efficient solution for the refinement of segmentation map. Liu et al.[87] and Wang et al.[139] all adopt the DeconvNet for RGB-D semantic segmentation because of good performance. However, the potential of DeconvNet is limited since the high-level prediction map aggregates large context for dense prediction. To this end, Cheng et al.[14] proposed a locality-sensitive DeconvNet (LS-DeconvNet) to refine the boundary segmentation over depth and color images. LS-DeconvNet incorporates local visual and geometric cues from the raw RGB-D data into each DeconvNet, which is able to up sample the coarse convolutional maps with large context while recovering sharp object boundaries.

Recurrent Neural Networks (RNNs) can capture long-range dependencies between pixels but are mainly suited to a single data channel (e.g. RGB). Fan et al.[29] extended the single-modal RNNs to multimodal RNNs (MM-RNNs) for application to RGB-D scene labeling. The MM-RNNs allow ‘memory’ sharing across depth and color channels. Each channel not only possess its own features but also has the attributes of other channel making the learned features more discriminative for semantic segmentation. Li et al.[79] proposed a novel Long Short-Term Memorized Context Fusion (LSTM-CF) model to capture and fuse contextual information from multiple channels of RGB and depth images.

Graph Neural Networks (GNNs) were first used for RGB-D semantic segmentation by Qi et al.[110] who cast the 2D RGB pixels into 3D space based on depth information and associated the 3D points with semantic information. Next, they built a k-nearest neighbor graph from the 3D points and applied a 3D graph neural network (3DGNN) to perform pixelwise predictions.

Data/Feature/Score Fusion: Optimal fusion of the texture (RGB channels) and geometric (depth channel) information is important for accurate semantic segmentation. There are three fusion tactics: data level, feature level and score level, referring to early, middle and late fusion respectively. A simple *data level fusion* strategy is to concatenate the RGB and depth images into four channels for direct input to a CNN model e.g. as performed by Couprie et al.[19]. However, such a data level fusion does not exploit the strong correlations between depth and photometric channels. *Feature level fusion*, on the other hand, captures these correlations. For example, Li et al.[79] proposed a memorized fusion

layer to adaptively fuse vertical depth and RGB contexts in a data-driven manner. Their method performs bidirectional propagation along the horizontal direction to hold true 2D global contexts. Similarly, Wang et al.[139] proposed a feature transformation network that correlates the depth and color channels, and bridges the convolutional networks and deconvolutional networks in a single channel. The feature transformation network can discover specific features in a single channel as well as common features between two channels, allowing the two branches to share features to improve the representation power of shared information. The above complex feature level fusion models are inserted in a specific same layer between RGB and depth channels, which is difficult to train and ignores other same layer feature fusion. To this end, Hazirbas et al.[42] and Jiang et al.[61] carry out fusion as an element-wise summation to fuse feature of multiple same layers between the two channels.

Score level fusion is commonly performed using the simple averaging strategy. However, the contributions of RGB model and depth model for semantic segmentation are different. Liu et al.[86] proposed a score level fusion layer with weighted summation that uses a convolution layer to learn the weights from the two channels. Similarly, Cheng et al.[14] proposed a gated fusion layer to learn the varying performance of RGB and depth channels for different class recognition in different scenes. Both techniques improved the results over the simple averaging strategy at the cost of additional learnable parameters.

Post-Processing: The results of CNN or DCNN used for RGB-D semantic segmentation are generally very coarse resulting in rough boundaries and the vanishing of small objects. A common method to address this problem is to couple the CNN with a Conditional Random Field (CRF). Wang et al.[141] further boost the mutual interactions between the two channels by the joint inference of Hierarchical CRF (HCRF). It enforces synergy between global and local predictions, where the global layouts are used to guide the local predictions and reduce local ambiguities, as well as local results provide detailed regional structures and boundaries. Mousavian et al.[101], Liu et al.[87], and Long et al.[86] adopt a Fully Connected CRF (FC-CRF) for post-processing, where the pixel-wise label prediction jointly considers geometric constraint, such as pixel-wise normal information, pixel position, intensity and depth, to promote the consistency of pixel-wise labeling. Similarly, Jiang et al.[61] proposed Dense-sensitive CRF (DCRF) that integrates the depth information with FC-CRF.

Table 2. Summary of RGB-D based methods with deep learning. Est.←depth estimation. Enc.←depth encoding. Mul.←multi-scale network. Nov.←novel neural network. Fus.←data/feature/score fusion. Pos.←post-processing.

Method[Reference]	Est.	Enc.	Mul.	Nov.	Fus.	Pos.	Architecture(2-stream)	Contribution
Cao et al.[8]	✓	✓	✗	✗	✓	✗	FCNs	Estimating depth images+a unified network for two tasks
Guo et al.[36]	✓	✗	✗	✗	✓	✗	FCNs	Incorporating depth & gradient for depth estim. + a network for two tasks
Wang et al.[141]	✓	✗	✗	✗	✗	✓	region/global CNN	HCRF for fusion and refining + two tasks by a network
Mous. et al.[101]	✓	✗	✓	✗	✓	✓	FCN	FC-CRF for refining + Mutual improvement for two tasks
Liu et al.[87]	✓	✗	✗	✓	✗	✓	S/D-DCNN	PBL for two feature maps integration+FC-CRF for fusion and refining
Hof et al.[46]	✗	✓	✗	✗	✗	✗	CNNs	A embedding for depth images
Gupta et al.[38]	✗	✓	✗	✗	✗	✗	CNNs	HHA for depth images
Hong et al.[86]	✗	✓	✗	✗	✓	✓	DCNNs	A new depth encoding+ FC-CRF for refining
Hazir. et al.[42]	✗	✓	✗	✗	✓	✗	Encoder-decoder	Semantic and depth feature fusion at each layer
Coup. et al.[19]	✗	✗	✓	✗	✓	✗	ConvNets	RGB laplacian pyramid for multi-scale features
Aman et al.[111]	✗	✓	✓	✗	✓	✗	VGG-M	A new multi-scale deep CNN
Lin et al.[82]	✗	✗	✓	✓	✓	✓	CFN	CaRF for multi-resolution features
Jiang et al.[61]	✗	✗	✗	✓	✓	✓	RGB-FCN	Semant. & depth feature fusion at each layer + DCRF for refining
Wang et al.[144]	✗	✗	✗	✓	✗	✗	Depth-aware CNN	Depth-aware Conv. and depth aware average pooling
Cheng et al.[14]	✗	✓	✗	✓	✓	✗	FCN + Deconv	LS-DeconvNet + novel gated fusion
Fan et al.[29]	✗	✗	✗	✓	✓	✗	MM-RNNs	Multimodal RNN
Li et al.[79]	✗	✓	✗	✓	✓	✗	LSTM-CF	LSTM-CF for capturing and fusing contextual inf.
Qi et al.[110]	✗	✗	✗	✓	✗	✗	3DGNN	GNN for RGB-D semantic segmentation
Wang et al.[139]	✗	✗	✗	✓	✓	✗	ConvNet - DeconvNet	MK-MMD for assessing the similarity between common features

3.2 Projected Images Based Segmentation

The core idea of projected images based semantic segmentation is to use 2D-CNNs to exploit features from projected images of 3D scenes/shapes and then fuse these features for label prediction. This pipeline not only exploits more semantic information from large-scale scenes compared to a single-view image, but also reduces the data size of a 3D scene compared to a point cloud. The projected images mainly include *multi-view images* or *spherical images*. Projected images based semantic segmentation methods are summarized in Table 3.

3.2.1 Multi-View Images Based Segmentation. MV-CNN[130] uses a unified network to combine features from multiple views of a 3D shape, formed by a virtual camera, into a single and compact shape descriptor to get improved classification performance. This inspired researchers to take the same idea into 3D semantic segmentation (see Fig.6). For example, Lawin et al.[70] project point clouds into multi-view synthetic images, including RGB, depth and surface normal images. The prediction score of all multi-view images is fused into a single representation and back-projected into each point. However, the snapshot can erroneously catch the points behind the observed structure if the density of the point cloud is low, which makes the deep network to misinterpret the multiple views. To this end, SnapNet[6],[5] preprocesses point clouds for computing point features (like normal or local noise) and generating a mesh, which is similar to point cloud densification. From the mesh and point clouds, they generate RGB and depth images by suitable snapshot. Then, they perform a pixel-wise labeling of 2D snapshots using FCN and fast back-project these labels into 3D points by efficient buffering. Above methods need obtain the whole point clouds of 3D scene in advance to provide a complete spatial structure for back-projection. However, the multi-view images directly obtained from real-world scene would lose much spatial information. Some works attempt to unite 3D scene reconstruction with semantic segmentation, where scene reconstruction could make up for spatial information. For example, Guerry et al.[35] reconstruct 3D scene with global multi-view RGB and Gray stereo images. Then, the labels of 2D snapshots are back-projected onto the reconstructed scene. But, simple back-projection can not optimally fuse semantic and spatial geometric features. Along the line, Pham et al.[106] proposed a novel Higher-order CRF, following back-projection, to further develop the initial segmentation.

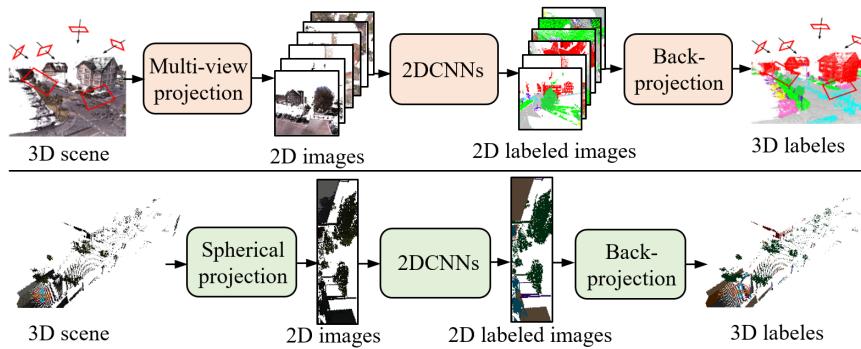


Fig. 6. Illustration of basic frameworks for projected images based segmentation methods. **Top:** Multi-view images based framework. **Bottom:** Spherical images based framework.

3.2.2 Spherical Images Based Segmentation. Selecting snapshots from a 3D scene is not straight forward. Snapshots must be taken after giving due consideration to the number of viewpoints, viewing distance and angle of the virtual

cameras to get an optimal representation of the complete scene. To avoid these complexities, researchers project the complete point cloud onto a sphere (see Fig.6.Bottom). For example, Wu et al.[152] proposed an end-to-end pipeline called SqueezeSeg, inspired from SqueezeNet [53], to learn features from spherical images which are then refined by CRF implemented as a recurrent layer. Similarly, PointSeg [148] extends the SqueezeNet by integrating the feature-wise and channel-wise attention to learn robust representation. SqueezeSegv2 [153] improves the structure of SqueezeSeg with Context Aggregation Module (CAM), adding LiDAR mask as a channel to increase robustness to noise. RangNet++ [98] transfers the semantic labels to 3D point clouds, avoiding discarding points regardless of the level of discretization used in CNN. Despite the likeness between regular RGB and LiDAR images, the feature distribution of LiDAR images changes at different locations. SqueezeSegv3 [160] has a spatially-adaptive and context-aware convolution, termed Spatially-Adaptive Convolution (SAC) to adopt different filters for different locations.

Table 3. Summary of projected images/voxel/other representation based methods with deep learning. M←multi-view image. S←spherical image. V←voxel. T←tangent images. L←lattice. P←point clouds.

Type	Method[Reference]	Input	Architecture	Feature extractor	Contribution
projected images	Lawin et al.[70]	M	multi-stream	VGG-16	Investigate the impact of different input modalities
	Boulch et al.[6][5]	M	SegNet/U-Net	VGG-16	An new and efficient framework SnapNet
	Guerry et al.[35]	M	SegNet/U-Net	VGG-16	An improved MVCNN+3D consistent data augmentation/sampling
	Pham et al.[106]	M	Two-stream	2DConv	High-order CRF+ real-time reconstruction pipeline
	Wu et al.[152]	S	AlexNet	firemodules	End-to-end pipeline SqueezeSeg + real time
	Wang et al.[148]	S	AlexNet	firemodules	Quite light-weight framework PointSeg + real time
	Wu et al.[153]	S	AlexNet	firemodules	A robust framework SqueezeSegV2
	Milioto et al.[98]	S	DarkNet	residual block	GPU-accelerated post-processing +RangNet++
voxel	Xu et al.[160]	S	RangeNet	SAC	Adopting different filters for different locations
	Huang et al.[51]	V	3DCNN	3DConv	Efficiently handling large data
	Tchapmi et al.[132]	V	3DFCNN	3DConv	Combining 3DFCNN with fine-representation using TI and CRF
	Meng et al.[96]	V	VAE	RBF	A novel voxel-based representation + RBF
	Liu et al.[84]	V	3DCNN+DQN+RNN	3DConv	Integrating object localization,segmen. and classifi. into one frame.
	Rethage et al.[112]	V	3DFCNN	FPCov	The first fully-convolutional network operating on raw point sets
	Dai et al.[22]	V	3DFCNN	3DConv	Combing scene completion and semantic labeling
	Riegler et al.[113]	V	Octree	3DConv	Making DL with high-resolution voxels
others	Graham et al.[33]	V	FCN/U-Net	SSConv	SSConv with less computation
	TangentConv[131]	T	U-Net	TConv	Tangent convolution + Parsing large scenes
	SPLATNet[129]	L	DeepLab	BConv	Hierarchical and spatially-aware feature learning
	LatticeNet[116]	L	U-Net	PN+3DConv	Hybrid architecture + novel slicing operator
	3DMV[21]	M+V	Cascade frame.	ENet+3DConv	Inferring 3D semantics from both 3D geometry and 2D RGB input
	Hung et al.[15]	V+M+P	Parallel frame.	SSCNet/DeepLab/PN	Leveraging 2D and 3D features
	PVCNN[90]	V+P	PN	PVConv	Both memory and computation efficient
	MVPNet[58]	M+P	Cascade frame.	U-Net+PN++	Leveraging 2D and 3D features
	LaserNet++[97]	M+P	Cascade frame.	ResNet+LNet	Unified network for two tasks

3.3 Voxel Based Segmentation

Similar to pixels, voxels divide the 3D space into many volumetric grids with a specific size and discrete coordinates. It contains more geometric information of the scene compared to projected images. 3D ShapeNets [156] and VoxNet [94] take volumetric occupancy grid representation as input to a 3D convolutional neural network for object recognition, which guides 3D semantic segmentation based on voxels. According to the unity of voxels size, voxel based methods can be divided into *uniform voxel* based and *nonuniform voxel* based methods. Voxel based semantic segmentation methods are summarized in Table 3.

3.3.1 Uniform Voxels. 3D CNN is a common architecture used to process uniform voxels for label prediction. Huang et al.[51] presented a 3D FCN for coarse voxel level predictions. Their method is limited by spatial inconsistency between predictions and provide a coarse labeling. Tchapmi et al.[132] introduce a novel network SEGCloud to produce

fine-grained predictions. It up samples the coarse voxel-wise prediction obtained from a 3D FCN to the original 3D point space resolution by trilinear interpolation.

With fixed resolution voxels, the computational complexity grows linearly with the increase of the scene scale. Large voxels can lower the computational cost of large-scale scene parsing. Liu et al.[84] introduced a novel network called 3DCNN-DQN-RNN. Like the sliding windows in 2D semantic segmentation, this network proposes eye window that traverses the whole data for fast localizing and segmenting class objects under the control of 3D-CNN and deep Q-Network (DQN). The 3D-CNN and Residual RNN further refine features in the eye window. The pipeline learns key features of interesting regions efficiently to enhance the accuracy of large-scale scene parsing with less computational cost. Rethage et al.[112] present a novel fully convolutional point network (FCPN), sensitive to multi-scale input, to parse large-scale scene without pro- or post-process steps. Particularly, FCPN is able to learn memory efficient representations that scale well to larger volumes. Similarly, Dai et al.[22] design a novel 3D-CNN to train on scene subvolumes but deploy on arbitrarily large scenes at test time, as it is able to handle large scenes with varying spatial extent. Additionally, their network adopts a coarse-to-fine tactic to predict multiple resolution scenes to handle the resolution growth in data size as the scene increases in size. Traditionally, the voxel representation only comprises Boolean occupancy information which loses much geometric information. Meng et al.[96] develop a novel information-rich voxel representation by using a variational auto-encoder(VAE) taking radial basis function(RBF) to capture the distribution of points within each voxel. Further, they proposed a group equivariant convolution to exploit feature.

3.3.2 Nonuniform Voxels. In fixed scale scenes, the computational complexity grows cubically as the voxel resolution increases. However, the volumetric representation is naturally sparse, resulting in unnecessary computations when applying 3D dense convolution on the sparse data. To alleviate this problem, OcNet [113] divides the space hierarchically into nonuniform voxels using a series of unbalanced octrees. Tree structure allows memory allocation and computation to focus on relevant dense voxels without sacrificing resolution. However, empty space still imposes computational and memory burden in OctNet. In contrast, Graham et al.[33] proposed a novel submanifold sparse convolution (SSC) that does not perform computations in empty regions, making up for the drawback of OcNet.

3.4 Point Based Segmentation

Point clouds are scattered irregularly in 3D space, lacking any canonical order and translation invariance, which restricts the use of conventional 2D/3D convolutional neural networks. Recently, a series of point-based semantic segmentation networks have been proposed. These methods can be roughly subdivided into three categories: multiple layer perceptron (MLP) based, point convolution based and graph convolution based. These methods are summarized in Table 4.

3.4.1 MLP Based. These methods apply a Multi Layer Perceptron directly on the points to learn features. Based on their framework, these methods can be further divided into two categories: *PN framework based* and *PN++ framework based*, as illustrated in Fig.7 (a) and (b).

PN framework based methods: The PointNet [108] (PN) is a pioneering work that directly processes point clouds. It uses shared MLP to exploit points-wise features and adopts a symmetric function such as max-pooling to collect these features into a global feature representation. Because the max-pooling layer only captures the maximum activation across global points, PN cannot learn to exploit local features.

Based on the framework of PN, some networks take on to define local regions to enhance local feature learning and utilize Recurrent Neural Networks (RNN) to increase the context feature exploitation. For example, Engelmann et al.[28] define local regions by KNN clustering and K-means clustering and use a simplified PN to extract local features.

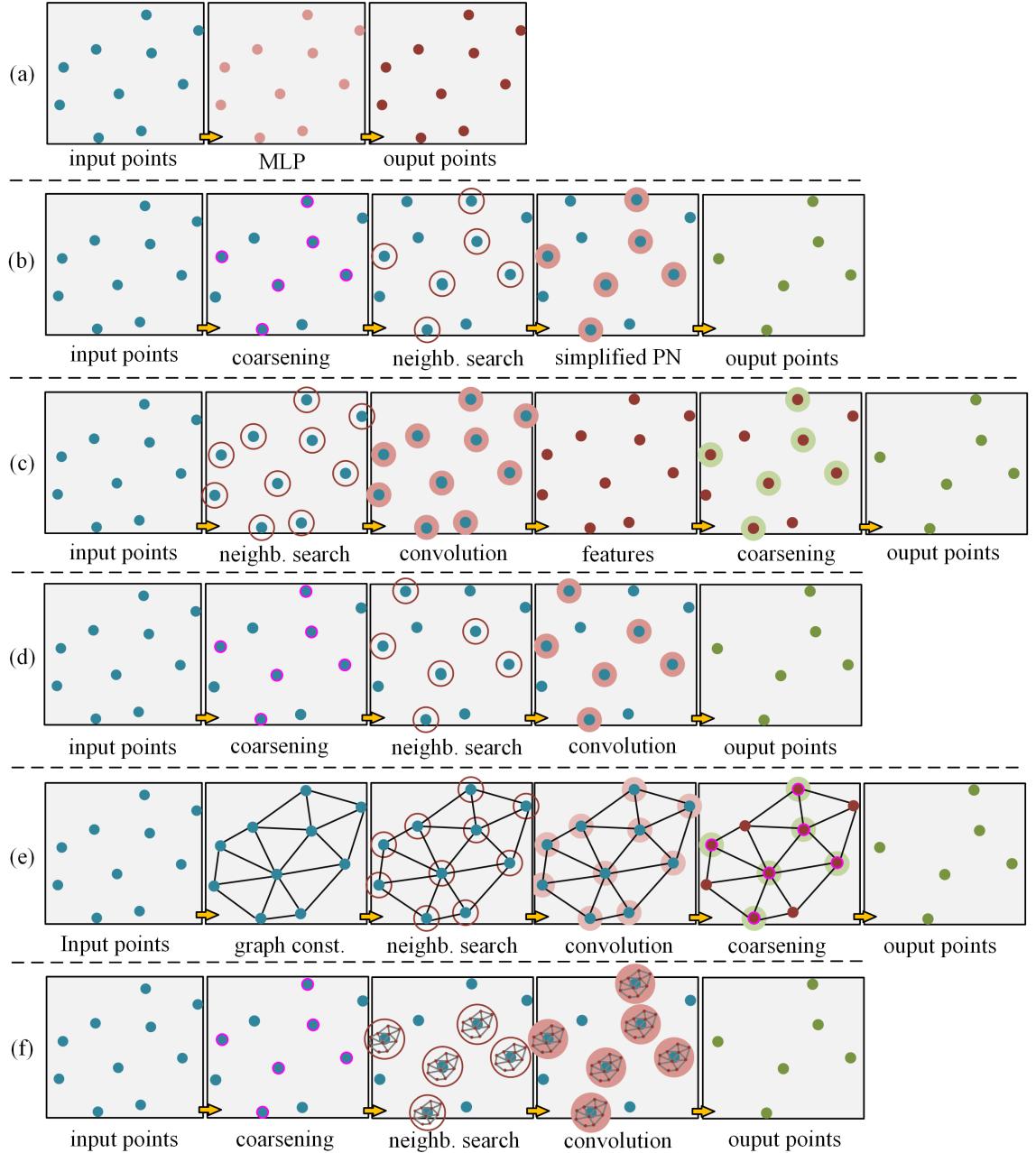


Fig. 7. An illustration of basic frameworks for point based methods, including (a) PN framework for MLP based methods, (b) PN++ framework for MLP based methods, (c) PN framework for point convolution based methods, (d) PN++ framework for point convolution based methods, (e) PN framework for graph convolution based methods, and (f) PN++ framework for graph convolution based methods.

ESC [26] divides global region points into multi-scale/grid blocks. The concatenated (local) block features are appended to the point-wise features and passed through Recurrent Consolidation Units (RCUs) to further learn global context features. Similarly, HRNN [168] uses Pointwise Pyramid Pooling (3P) to extract local features on the multi-size local regions. Point-wise features and local features are concatenated and a two-direction hierarchical RNN explores context features on these concatenated features. However, the local features learned are not sufficient because the deeper layer features do not cover a larger spatial extent.

PN++ framework based methods: Building on PointNet, PointNet++ [109] (PN++) defines a hierarchical learning architecture. It hierarchically samples points using farthest point sampling (FPS) and groups local regions using k nearest neighbor search as well as ball search. Progressively, a simplified PointNet exploits features in local regions at multiple scales or multiple resolutions. PN++ framework expands the receptive field to exploit more local features collectively. Inspired by SIFT [91], PointSIFT [63] inserts a PointSIFT module layer before the sampling layer to learn local shape information. This module transforms each point into a new shape representation by encoding information of different orientations. Similarly, PointWeb [177] inserts Adaptive Feature Adjustment (AFA) module layer after a grouping layer to embed the interactive information between points into each point. These tactics enhance the representation ability of learned point-wise features. However, MLP still treats each local point individually and does not pay attention to the geometric connections between local points. Moreover, the MLP is effective but lacks the complexity to capture wider and finer local features.

3.4.2 Point Convolution Based. Point convolution based methods perform convolution operations directly on the points. Similar to the MLP based segmentation, these networks can also be subdivided into *PN framework based* and *PN++ framework based* methods as illustrated in Fig. 7(c)(d).

PN framework based methods perform convolution on the neighboring points of each point. For example, RSNet[52] exploit point-wise features using 1x1 convolution and then pass them through the local dependency module (LDM) to exploit local context features. However, it does not define the neighborhood for each point in order to learn local features. On the other hand, PointwiseCNN [50] sorts points in a specific order, e.g. XYZ coordinate or Morton cureve[100], and queries nearest neighbors dynamically and bins them into 3x3x3 kernel cells before convolving with the same kernel weights. DPC [27] adapts point convolution [154] on neighborhood points of each point where the neighborhood points are determined though a dilated KNN search. That method integrates dilated mechanism into KNN search to expand the receptive yield. PCNN [143] performs Parametric CNN, where the kernel is estimated as an MLP, on KD-tree neighborhood to learn local features. However, the fixed resolution of the feature map makes the network difficult to suit deeper architectures.

Based on the framework, some methods insert coarsening layer after convolution layers to progressively decrease the point resolution. This approach highly resembles the spatial 2D convolution neural network framework. KPCConv [133] uses strided convolution to reduce the number of points. The convolution kernels are defined as a set of kernel points with weights and the influenced area of kernel points is defined by a correlation function proposed in [121]. Flex-Convolution [34] uses a linear function with fewer parameters to model a convolution kernel and adapts inverse density importance sub-sampling (IDISS) to coarsen the points. PCNN [77] coarsens the input points with farthest point sampling. The convolution layer learns an χ -transformation from local points by MLP to simultaneously weight and permute the features, subsequently applying a standard convolution on these transformed features.

PN++ framework based methods have the convolution layer as their key layer. For instance, an extension of Monte Carlo approximation for convolution called PointConv [154] takes the point density into account. It uses MLP to

approximate a weight function of the convolution kernel, and applies an inverse density scale to reweight the learned weight function. Similarly, MCC [45] phrases convolution as a Monte Carlo integration problem by relying on point probability density function (PDF), where the convolution kernel is also represented by an MLP. Moreover, it introduces Possion Disk Sampling (PDS)[151] to construct a point hierarchy instead of FPS, which provides an opportunity to get the maximal number of samples in a receptive field. A-CNN [67] defines a new local ring-shaped region by dilated KNN, and projects points on a tangent plane to further order neighbor points in local regions. Then, the standard point convolutions are performed on these ordered neighbors represented as a closed loop array.

In the large-scale point clouds semantic segmentation area, RandLA-Net [48] uses random point sampling instead of the more complex point selection approach. It introduces a novel local feature aggregation module (LFAM) to progressively increase the receptive field and effectively preserve geometric details. Another technology, PolarNet [175] first partitions a large point cloud into smaller grids (local regions) along their polar bird's-eye-view (BEV) coordinates. It then abstracts local region points into a fixed-length representation by a simplified PointNet and these representations are passed through a standard convolution.

3.4.3 Graph Convolution Based. The graph convolution based methods perform convolution on points connected with a graph structure. Here, the graph construction (definition) and convolution design are becoming the two main challenges. The same categorization of *PN framework* and *PN++ framework* also applies to graph convolution methods as illustrated in Fig. 7(e) and (f).

PN framework based methods construct the graph from the points globally and perform convolution on neighborhoods points of each point. For example, ECC [125] is among of the pioneer methods to apply spatial graph network to extract features from point clouds. It dynamically generates edge-conditioned filters to learn edge features that describe the relationships between a point and its neighbors. Based on PN architecture, DGCN [149] implements dynamic edge convolution called EdgeConv on the neighborhood of each point. The convolution is approximated by a simplified PN. SPG [69] parts the point clouds into a number of simple geometrical shapes (termed super-points) and builts super graph on global super-points. Furthermore, this network adopts PointNet to embed these points and refine the embedding by Gated Recurrent Unit (GRU).

The dynamic generation of edge convolution filters comes with a significant computational overhead. Moreover, spatial graph networks generally suffer from the vanishing gradient problem. These limitations pose a major bottleneck in designing deeper GCN architectures. DeepGCNs [74] borrows some concepts from 2D-CNN such as residual connections between different layers (ResNet) to alleviate the vanishing gradient problem, and dilation mechanism to allow the GCN to go deeper. Lei et al. [73] propose a discrete spherical convolution kernel (SPH3D kernel) that consists of the spherical convolution learning depth-wise features and point-wise convolution learning point-wise features.

Attention mechanism has recently become popular for improving point cloud segmentation accuracy. For example, Ma et al. [93] use the channel self-attention mechanism to learn independence between any two point-wise feature channels, and further define a Channel Graph where the channel maps are presented as nodes and the independencies are represented as graph edges. AGCN [159] integrates attention mechanism with GCN for analyzing the relationships between local features of points and introduces a global point graph to compensate for the relative information of individual points.

PN++ framework based methods perform convolution on local points with graph structure. Graphs are either spectral or spatial graphs. In the former case, LS-GCN [137] adopts the basic architecture of PointNet++, replaces MLPs with a spectral graph convolution using standard unparametrized Fourier kernels, as well as a novel recursive spectral

cluster pooling substitute for max-pooling. However, transformation from spatial to spectral domain incurs a high computational cost. Besides that, spectral graph networks are usually defined on a fixed graph structure and are thus unable to directly process data with varying graph structures.

In the spatial graph category, based on the basic architecture of PoinNet++, Feng et al. [30] constructed a local graph on neighborhood points searched along multi-directions and explore local features by a local attention-edge convolution (LAE-Conv). These features are imported into a point-wise spatial attention module to capture accurate and robust local geometric details. Similarly, Li et al. [78] proposed Geometric Graph Convolution (TGConv), its filters defined as products of local point-wise features with local geometric connection features expressed by Gaussian weighted Taylor kernels. Continuous graph convolution also incurs a high computational cost. Inspired by the separable convolution strategy in Xception [16] that significantly reduces parameters and computation burden, HDGCN [80] designed a DGConv that composes depth-wise graph convolution followed by a point-wise convolution, and add DGConv into the hierarchical structure to extract local and global features.

Tree structures such as KD-tree and Octree can be viewed as a special type of graph, allowing to share convolution layers depending on the tree splitting orientation. 3DContextNet [174] adopts a KD-tree structure to hierarchically represent points where the nodes of different tree layers represent local regions at different scales, and employs a simplified PointNet with a gating function on nodes to explore local features. However, their performance depends heavily on the randomization of the tree construction. Lei et al. [72] built an Octree based hierarchical structure on global points to guide the spherical convolution computation in per layer of the network. The spherical convolution kernel systematically partitions a 3D spherical region into multiple bins that specifies learnable parameters to weight the points falling within the corresponding bin.

3.5 Other Representation Based

Some methods transform the original point cloud to representations other than projected images, voxels and points. Examples of such representations include *tangent images*[131] and *lattice*[129],[116]. In the former case, Tatarchenko et al.[131] project local surfaces around each-point to a series of 2D tangent images and develop a tangent convolution based U-Net to extract features. In the latter case, SPLATNet [129] adapts the bilateral convolution layers (BCLs) proposed by Jampani et al.[56] to smoothly map disordered points onto a sparse lattice. Similarly, LatticeNet [116] uses a hybrid architecture that combines PointNet, which obtains low-level features, with sparse 3D convolution, which explores global context features. These features are embedded into a sparse lattice that allows the application of standard 2D convolutions.

Although the above methods have achieved significant progress in 3D semantic segmentation, each has its own drawbacks. For instance, multi-view images have more semantic information but less geometric information of the scene. On the other hand, voxels have more geometric information but less semantic information. To get the best of both worlds, some methods adopt *hybrid representations* as input to learn comprehensive features of a scene. Dai et al.[21] map 2D semantic features obtained by multi-view networks into 3D grids of scene. These pipelines make 3D grids attach rich 2D semantic as well as 3D geometric information so that the scene can get better segmentation by a 3D-CNN. Similarly, Hung et al.[15] back-project 2D multi-view image features on to the 3D point cloud space and use a unified network to extract local details and global context from sub-volumes and the global scene respectively. Liu et al.[90] argue that voxel-based and point-based NN are computationally inefficient in high-resolution and data structuring respectively. To overcome these challenges, they propose Point-Voxel CNN (PVCNN) that represents the 3D input data as point clouds to take advantage of the sparsity to lower the memory footprint, and leverage the voxel-based

Table 4. Summary of point based semantic segmentation methods with deep learning.

PN frame.	Method[Refer.]	Neighb. search	Feature extractor	Coarsening	Contribution
MLP	PointNet[108]	None	MLP	None	Pioneering processing points directly
	G+RCU[28]	None	MLP	None	Two local definition+local/global pathway
	ESC[26]	None	MLP	None	MC/Grid Block for local defini.+RCUs for context exploitation
	HRNN[168]	None	MLP	None	3P for local feature exploi.+HRNN for local context exploitation
Point Conv.	RSNet[52]	None	1x1 Conv	None	LDM for local context exploitation
	DPC[27]	DKNN	PointConv	None	Dilated KNN for expanding the receptive field
	PWCNN[50]	Grid	PWConv.	None	Novel point convolution
	PCNN[143]	KD index	PCCov.	None	KD-tree index for neighb. search+novel point Conv.
	KPConv[133]	KNN	KPConv.	strided Conv	Novel point convolution
	FlexConv[34]	KD index	flexConv.	IDISS	novel point Conv.+flex-maxpooling without subsampling
	PointCNN[77]	DKNN	χ -Conv	FPS	Novel point convolution
Graph Conv.	DGCN[149]	KNN	EdgeConv	None	Novel graph convolution+updating graph
	SPG[69]	partition	PN	None	Superpoint graph + parsing large-scale scene
	DeepGCNs[74]	DKNN	DGConv	RPS	Adapting residual connections between layers
	SPH3D-GCN[73]	FPS	SPH3D-GConv	Ball	Novel graph convolution+pooling+uppooling
	PGCRNet[93]	None	Conv1D	None	PointGCR to model context dependencies between different categories
	AGCN[159]	KNN	MLP	None	Point attention layer for aggregating local features
PN++ frame.	Method[Refer.]	Coarsening	Neighb. search	Feature extractor	Contribution
MLP	PointNet++[109]	FPS	Ball/KNN	PN	Proposing hierarchical learning framework
	PointWeb[177]	FPS	KNN	PN	AFA for interactive feature exploitation
	PointSIFT[63]	FPS	KNN	PN	PointSIFT module for local shape information
Point Conv.	MCC[45]	PDS	Ball	MCConv.	Novel coarsening layer+point convolution
	PointConv[154]	FPS	KNN	PointConv	Novel point convolution considering point density
	A-CNN[67]	FPS	Dilated KNN	ACovn	Novel neighborhood search+point convolution
	RandLA-Net[48]	RPS	KNN	LocSE	LFAM with large receptive field and keeping geometric details
	PolarNet[175]	PolarGrid	None	PN	Novel local regions definition + RingConv
Graph Conv.	LS-GCN[137]	FPS	KNN	Spec.Conv.	Local spectral graph + Novel graph convolution
	PAN[30]	PFS	Multi-direct.	LAE-Conv	Point-wise spatial attention+local graph Conv.
	TGNet[78]	FPS	Ball	TGConv	Novel graph Conv.+multi-scale features explo.
	HGDGCN[80]	FPS	KNN	DGConv	Depthwise graph Conv.+ Pointwise Conv.
	3DCoN-Net[174]	Tree layer	KNN	PN	KD tree structure
	ψ -CNN[72]	Tree layer	Octree neig.	ψ -Conv	Octree structure+ Novel graph convolution

convolution to obtain a contiguous memory access pattern. Jaritz et al.[58] proposed MVPNet that collect 2D multi-view dense image features into 3D sparse point clouds and then use a unified network to fuse the semantic and geometric features. Also, Meyer et al.[97] fuse 2D image and point clouds to address 3D object detection and semantic segmentation by a unifying network. The other representations based semantic segmentation methods are summarized in Table 3.

4 3D INSTANCE SEGMENTATION

3D instance segmentation methods additionally distinguish between different instances of the same class. Being a more informative task for scene understanding, 3D instance segmentation is receiving increased interest from the research community. 3D instance segmentation methods are roughly divided into two directions: *proposal-based* and *proposal-free*.

4.1 Proposal Based

Proposal-based methods first predict object proposals and then refine them to generate final instance masks (see Fig.8), breaking down the task into two main challenges. Hence, from the proposal generation point of view, these methods can be grouped into *detection-based* and *detection-free* methods.

Detection-based methods sometimes define object proposals as a 3D bounding box regression problem. 3D-SIS[47] incorporates high-resolution RGB images with voxels, based on the pose alignment of the 3D reconstruction, and jointly learns color and geometric features by a 3D detection backbone to predict 3D bounding box proposals. In

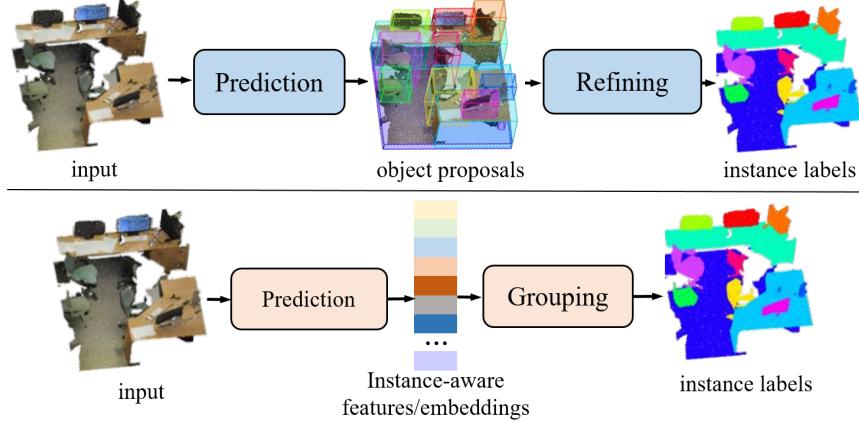


Fig. 8. Illustration of two different approaches for 3D instance segmentation. **Top:** proposal-based framework. **Bottom:** proposal-free framework.

these proposals, a 3D mask backbone predicts the final instance masks. Similarly, GPSN[171] introduces a 3D object proposal network termed Generative Shape Proposal Network (GPSN) that reconstructs object shapes from shape noisy observations to enforce geometric understanding. GPSN is further embedded into a 3D instance segmentation network named Region-based PointNet (R-PointNet) to reject, receive and refine proposals. Training of these networks needs to be performed step-by-step and the object proposal refinement requires expensive suppression operation. To this end, Yang et al.[164] introduced a novel end-to-end network named 3D-BoNet to directly learn a fixed number of 3D bounding boxes without any rejection, and then estimate an instance mask in each bounding box.

Detection-free methods include SGPN [145] which assumes that the points belonging to the same object instance should have very similar features. Hence, it learns a similarity matrix to predict proposals. The proposals are pruned by confidence scores of the points to generate highly credible instance proposals. However, this simple distance similarity metric learning is not informative and is unable to segment adjacent objects of the same class. To this end, 3D-MPA [25] learns object proposals from sampled and grouped point features that vote for the same object center, and then consolidates the proposal features using a graph convolutional network enabling higher-level interactions between proposals which result in refined proposal features. AS-Net [60] uses an assignment module to assign proposal candidates and then eliminates redundant candidates by a suppression network.

4.2 Proposal Free

Proposal-free methods learn feature embedding for each point and then apply clustering to obtain definitive 3D instance labels (see Fig.8) breaking down the task into two main challenges. From the embedding learning point of a view, these methods can be roughly subdivided into three categories: *multi-embedding learning*, *2D embedding propagation* and *multi-task learning*.

Multi-embedding learning: methods like MASC [83] rely on high performance of the SSCN [33] to predict the similarity embedding between neighboring points at multiple scales and semantic topology. A simple yet effective clustering [89] is adapted to segment points into instances based on the two types of learned embeddings. MTML [68] learns two sets of feature embeddings, including the feature embedding unique to every instance and the direction embedding that orients the instance center, which provides a stronger grouping force. Similarly, PointGroup [62] groups

points into different clusters based on the original coordinate embedding space and the shifted coordinate embedding space. In addition, the proposed ScoreNet guides the proper cluster selection.

2D embedding propagation methods: An example of these methods is the 3D-BEVIS [23] that learns 2D global instance embedding with a bird’s-eye-view of the full scene. It then propagates the learned embedding onto point clouds by DGCN [149]. Another example is PanopticFusion [102] which predicts pixel-wise instance labels by 2D instance segmentation network Mask R-CNN [43] for RGB frames and integrates the learned labels into 3D volumes.

Multi-task jointly learning: 3D semantic segmentation and 3D instance segmentation can influence each other. For example objects with different classes must be different instances, and objects with the same instance label must be the same class. Based on this, ASIS [146] designs an encoder-decoder network, termed ASIS, to learn semantic-aware instance embeddings for boosting the performance of the two tasks. Similarly, JSIS3D [107] uses a unified network namely MT-PNet to predict the semantic labels of points and embedding the points into high-dimensional feature vectors, and further propose a MV-CRF to jointly optimize object classes and instance labels. Similarly Liu et al.[83] and 3D-GEL [81] adopt SSCN to generate semantic predictions and instance embeddings simultaneously, then use two GCNs to refine the instance labels. OccuSeg [40] uses a multi-task learning network to produce both occupancy signal and spatial embedding. The occupancy signal represents the number of voxel occupied by per voxel. 3D instance segmentation methods are summarized in Table 5.

Table 5. Summary of 3D instance segmentation methods with deep learning. M←multi-view image; Me←mesh; V←voxel; P←point clouds.

Type	Method[Refer.]	Input	Propo./Embed. Prediction	Refining/Grouping	Contribution
proposal based	GSPN[171]	P	GSPN	R-PointNet	A new proposal generation methods
	3D-SIS[47]	M+V	3D-RPN+3D-RoI	3DFCN	Learning on both geometry and RGB input for 3D bounding box
	3D-BoNet[164]	P	Bounding box regression	Point mask prediction	Directly regressing 3D bounding box
	SGPN[145]	P	SM + SCM + PN	Non-Maximum Suppression	A new group proposal
	3D-MPA[25]	P	SSCNet	Graph ConvNet	Multi proposal aggregation strategy
	AS-Net[60]	P	Four branches with MLPs	Candidate proposal Suppression	An algorithm mapping instance labels to instance candidates
proposal free	MASC[83]	Me	U-Net with SSConv	Clustering algorithm	A novel clustering based on multi-scale affinity and mesh topology
	MTML[68]	V	SSCNet	Mean-shift clustering	Multi-task learning
	PointGroup[62]	P	U-Net with SSConv	Point clustering + ScoreNet	A novel clustering algorithm based on dual coordinate sets
	3D-BEVIS[23]	M	U-Net/FCN + 3D prop.	Mean-shift clustering	Joint 2D-3D feature
	PanopticFus[102]	M	PSPNNet/Mask R-CNN	FC-CRF	Coopering with semantic mapping
	ASIS[146]	P	1 encoder+ 2 decoders	ASIS module	Simultaneously performing sem./ins. segmentation tasks
	JSIS3D[107]	P	MT-PNet	MV-CRF	Simultaneously performing sem./ins. segmentation tasks
	3D-GEL[81]	P	SSCNet	GCN	Structure-aware loss function + attention-based GCN
	OccuSeg[40]	P	3D-UNet	Graph-based clustering	Proposing a novel occupancy signal

5 3D PART SEGMENTATION

3D part segmentation is the next finer level, after instance segmentation, where the aim is to label different parts of an instance. The pipeline of part segmentation is quite similar to that of semantic segmentation except that the labels are now for individual parts. Therefore, some existing 3D semantic segmentation networks [96], [33], [108], [109], [174], [52], [133], [50], [45], [154], [77], [149], [73], [159], [143], [34], [72], [129], [116] can also be trained for part segmentation. However, these networks can not entirely tackle the difficulties of part segmentation. For example, various parts with the same semantic label might have diverse shapes, and the number of parts for an instance with the same semantic label may be different. We subdivide 3D part segmentation methods into two categories: *regular data* based and *irregular data* based as follows.

5.1 Regular Data Based

Regular data usually includes projected images [64], voxels [150],[71],[128]. As for projected images, Kalogerakis et al.[64] obtain a set of images from multiple views that optimally cover object surface, and then use multi-view Fully Convolutional Networks(FCNs) and surface-based Conditional Random Fields (CRFs) to predict and refine part labels separately. Voxel is a useful representation of geometric data. However, fine-grained tasks like part segmentation require high resolution voxels with more detailed structure information, which leads to high computation cost. Wang et al.[150] proposed VoxSegNet to exploit more detailed information from voxels with limited resolution. They use spatial dense extraction to preserve the spatial resolution during the sub-sampling process and an attention feature aggregation (AFA) module to adaptively select scale features. Le et al.[71] introduced a novel 3D-CNN called PointGrid, to incorporate a constant number of points with each cell allowing the network to learn better local geometry shape details. Furthermore, multiple model fusion can enhance the segmentation performance. Combining the advantages of images and voxels, Song et al.[128] proposed a two-stream FCN, termed AppNet and GeoNet, to explore 2D appearance and 3D geometric features from 2D images. In particular, their VolNet extracts 3D geometric features from 3D volumes guiding GeoNet to extract features from a single image.

Table 6. Summary of 3D part segmentation methods. M \leftarrow multi-view image; Me \leftarrow mesh; V \leftarrow voxel; P \leftarrow point clouds.

Type	Method[Reference]	Input	Architecture	Feature extractor	Contribution
reg. data	ShapePPCN[64]	M	Multi-stream FCN	2DConv	Per-label confidence maps + surface-based CRF
	VoxSegNet[150]	V	3DU-Net	AtrousConv	SDE for preserving the spatial resolution AFA for selecting multi-scales features
	Pointgrid[71]	V	Conv-deconv	3DConv	Learning higher order local geometry shape.
	SubvolumeSup [128]	M+V	2-stream FCN	2D/3DConv	GeoNet/AppNet for 3/2D features exploi. + DCT for aligning image and voxel
irregular data	DCN[161]	Me	2-tream DCN & NN	DirectionalConv	DCN/NN for local feature and global feature.
	MeshCNN[41]	Me	2D-CNN	MeshConv	Novel mesh convolution and pooling
	PartNet[172]	P	RNN	PN	part feature learning scheme for context and geometry feature exploitation
	SSCNN[170]	P	FCN	SpectralConv	STN for allowing weight sharing + spectral multi-scale kernel
	KCNet[121]	P	PN	MLP	KNN graph on points + kernel correlation for measuring geometric affinity
	SFCN[140]	P	FCN	SFCConv	Novel point convolution
	SpiderCNN[163]	P	PN	SpiderConv	Novel point convolution
	FeaStNet[136]	P	U-Net	GConv	Dynamic graph convolution filters
	Kd-Net[65]	P	Kd-tree	Affine Transformation	Using Kd-tree to build graphs and share learnable parameters
	O-CNN[142]	P	Octree	3DConv	Making 3D-CNN feasible for high-resolu. voxels
	PointCapsNet[178]	P	Encoder-decoder	PN	Semi-supervision learning
	SO-Net[75]	P	Encoder-decoder	FC layers	SOM for modeling the spatial distribution of points + un-supervision learning

5.2 Irregular Data Based

Irregular data representations usually includes meshes [161], [41] and point clouds [75], [121], [170], [136], [140], [172], [178]. Mesh provides an efficient approximation to a 3D shape because it captures the flat, sharp and intricate of surface shape surface and topology. Xu et al.[161] put the face normal and face distance histogram as the input of a two-stream framework and use the CRF to optimize the final labels. Inspired by traditional CNN, Hanocka et al.[41] design novel mesh convolution and pooling to operate on the mesh edges.

As for point clouds, the graph convolution is the most commonly used pipeline. In the spectral graph domain, SyncSpecCNN[170] introduces a Sychronized Spectral CNN to process irregular data. Specially, multichannel convolution and parametrized dilated convolution kernels are proposed to solve multi-scale analysis and information sharing across shapes respectively. In spatial graph domain, in analogy to a convolution kernel for images, KCNet[121] present point-set kernel and nearest-neighbor-graph to improve PointNet with an efficient local feature exploitation structure. Similarly, Wang et al.[140] design Shape Fully Convolutional Networks (SFCN) based on graph convolution and pooling operation, similar to FCN on images. SpiderCNN [163] applies a special family of convolutional filters that combine

simple step function with Taylor polynomial, making the filters to effectively capture intricate local geometric variations. Furthermore, FeastNet [136] uses dynamic graph convolution operator to build relationships between filter weights and graph neighborhoods instead of relying on static graph of the above network.

A special kind of graphs, the trees (e.g. Kd-tree and Oc-tree), work on 3D shapes with different representations and can support various CNN architectures. Kd-Net [65] uses a kd-tree data structure to represent point cloud connectivity. However, the networks have high computational cost. O-CNN [142] designs an Octree data structure from 3D shapes. However, the computational cost of the O-CNN grows quadratically as the depth of tree increases.

SO-Net [75] sets up a Self-Organization Map (SOM) from point clouds, and hierarchically learns node-wise features on this map using the PointNet architecture. However, it fails to fully exploit local features. PartNet [172] decomposes 3D shapes in a top-down fashion, and proposes a Recursive Neural Network (RvNN) for learning the hierarchy of fine-grained parts. Zhao et al.[178] introduce an encoder-decoder network, 3D-PointCapsNet, to tackle several common point cloud-related tasks. The dynamic routing scheme and the peculiar 2D latent space deployed by capsule networks, deployed in their model, bring improved performance. The 3D instance segmentation methods are summarized in Table 6.

6 APPLICATIONS OF 3D SEGMENTATION

We review 3D semantic segmentation methods for two main applications, unmanned systems and medical diagnosis.

6.1 Unmanned Systems

As LIDAR scanners and depth cameras become widely available and more affordable, they are increasingly being deployed in unmanned systems such as autonomous driving and mobile robots. These sensors provide realtime 3D video, generally at 30 frames per second (fps), as direct input to the system making *3D video semantic segmentation* as the primary task to understand the scene. Furthermore, in order to interact more effectively with the environment, unmanned systems generally build a *3D semantic map* of the scene. Below we review 3D video based semantic segmentation and 3D semantic map construction.

6.1.1 3D video semantic segmentation. Compared to the 3D single frame/scan semantic segmentation methods reviewed in Section 3.1, 3D video (continuous frames/scans) semantic segmentation methods take into account the connecting spatio-temporal information between frames which is more powerful at parsing the scene robustly and continuously. Conventional convolutional neural networks (CNNs) are not designed to exploit the temporal information between frames. A common strategy is to adapt Recurrent Neural Networks or Spatio-temporal convolutional network.

Recurrent Neural Network based Segmentation: RNNs generally work in combination with 2D-CNNs to process RGB-D videos. The 2D-CNN learns to extract the frame-wise spatial information and the RNN learns to extract the temporal information between the frames. Valipour et al.[134] proposed Recurrent Fully Neural Network to operate over a sliding window over the RGB-D video frames. Specifically, the convolutional gated recurrent unit preserves the spatial information and reduces the parameters. Similarly, Yurdakul et al.[24] combine fully convolutional and recurrent neural network to investigate the contribution of depth and temporal information separately in the synthetic RGB-D video.

Spatio-temporal CNN based Segmentation: Nearby video frames provide diverse viewpoints and additional context of objects and scenes. STD2P[44] uses a novel spatio-temporal pooling layer to aggregate region correspondences computed by optical flow and image boundary-based super-pixels. Choy et al.[17] proposed 4D Spatio-Temporary

ConvNet, to directly process a 3D point cloud video. To overcome challenges in the high-dimensional 4D space (3D space and time), they introduced the 4D spatio-temporal convolution, a generalized sparse convolution, and the trilateral-stationary conditional random field that keeps spatio-temporal consistency. Similarly, based on 3D sparse convolution, Shi et al.[122] proposed SpSequenceNet that contains two novel modules, a cross frame-global attention module and a cross-frame local interpolation module to exploit spatial and temporal feature in 4D point clouds.

6.1.2 3D semantic map construction. Unmanned systems do not just need to avoid obstacles but also need to establish a deeper understanding of the scene such as object parsing, self localization etc. To facilitate such tasks, unmanned systems build a 3D semantic map of the scene which includes two key problems: geometric reconstruction and semantic segmentation. 3D scene reconstruction has conventionally relied on simultaneous localization and mapping system (SLAM) to obtain a 3D map without semantic information. This is followed by 2D semantic segmentation with a 2D-CNN and then the 2D labels are transferred to the 3D map following an optimization (e.g. conditional random field) to obtain a 3D semantic map [165]. This common pipeline does not guarantee high performance of 3D semantic maps in complex, large-scale, and dynamic scenes. Efforts have been made to enhance the robustness using association information exploitation from multiple frames, multi-model fusion and novel post-processing operations. These efforts are explained below.

Association information exploitation: mainly depends on SLAM trajectory, recurrent neural networks or scene flow. Ma et al.[92] enforce consistency by warping CNN feature maps from multi-views into a common reference view by using the SLAM trajectory and to supervise training at multiple scales. SemanticFusion [95] incorporates deconvolutional neural networks with a state-of-the-art dense SLAM system, ElasticFusion, which provides long-term correspondence between frames of a video. These correspondences allow label predictions from multi-views to be probabilistically fused into a map. Similarly, using the connection information between frames provided by a recurrent unit on RGB-D videos, Xiang et al.[157] proposed a data associated recurrent neural networks (DA-RNN) and integrated the output of the DA-RNN with KinectFusion, which provides a consistent semantic labeling of the 3D scene. Cheng et al.[13] use a CRF-RNN-based semantic segmentation to generate the corresponding labels. Specifically, the authors proposed an optical flow-based method to deal with the dynamic factors for accurate localization. Kochanov et al.[66] also use scene flow to propagate dynamic objects within the 3D semantic maps.

Multiple model fusion: Jeong et al.[59] build a 3D map by estimating odometry based on GPS and IMU, and use a 2D-CNN for semantic segmentation. They integrate the 3D map with semantic labels using a coordinate transformation and Bayes' update scheme. Zhao et al.[176] use PixelNet and VoxelNet to exploit global context information and local shape information separately and then fuse the score maps with a softmax weighted fusion that adaptively learns the contribution of different data streams. The final dense 3D semantic maps are generated with visual odometry and recursive Bayesian update.

6.2 Medical Diagnosis

The 2D U-Net[115] and 3D U-Net[18] are commonly used for medical image segmentation to achieve reasonable segmentation accuracy with very few training samples comparing to conventional FCN. Based on these basic ideas, many improved architectures have been designed which can mainly be divided into four categories: *Extended 3D U-Net*, *Joint 2D-3D CNN*, *CNN with optimization module* and *Hierarchical networks*.

Extended 3D U-Net: To exploit local and contextual information of 3D volumetric data, Adria et al.[9] use three different 3D-CNNs. The three 3D-CNNs (3DNet_1, 3DNet_2 and 3DNet_3) are based on 3D U-Net, and integrate fine

and coarse features to achieve the final segmentation. To accelerate the training, Zeng et al.[173] proposed a deeply supervised 3D U-Net-like fully convolutional network for segmentation, which can directly map a whole volumetric data to its volume-wise labels. An important aspect of this method is that the multi-level deep supervision alleviates the potential gradient vanishing problem during training. Furthermore, Roth et al.[117] firstly utilize a random forest regression to give an initial bounding box region of interest, then use a 3D U-Net type FCN that can be deployed on multiple GPUs.

Joint 2D/3D-CNN: Considering the rich of 2D semantic information in 2D-CNNs and 3D spatial information in 3D-CNNs, Jay et al.[105] developed a 2D-3D fully convolutional neural network. The 2D segmentation model is adapted from 2D U-Net and is trained slice-by-slice. The 3D-CNN model computes the volumetric segmentation of the full volume. A new dice loss function is also introduced to achieve accuracy scores in terms of geometry and clinical validity. Similarly, Christian et al.[2] investigate the suitability of 2D and 3D CNN for segmentation of three cardiac structures. Unlike the two approaches, Guotai Wang et al.[138] propose a novel deep learning-based framework to exploit 2D and 3D features for interactive 2D and 3D medical image segmentation. This network does not require annotations of all the organs for training and can, therefore, be applied to new organs or new segmentation protocols. To address the lack of 3D spatial information in 2D-CNNs and computational complexity of 3D-CNNs, Li et al.[76] proposed a hybrid densely connected UNet (H-DenseUNet), which consists of 2D DenseUNet for efficiently extracting intra-slice features and a 3D counterpart for hierarchically aggregating the volumetric context for liver and tumor segmentation.

CNN with optimization module: These architectures are state-of-the-art for medical image segmentation. A 3D-CNN optimization produces soft segmentation maps which is followed by a 3D FC-CRF for refinement of the maps. However, the 3D CNN and 3D FC-CNN must be trained separately. Miguel et al.[99] apply the CRF as RNN layers for semantic segmentation to 3D medical imaging segmentation. Similarly, Zhong et al.[179] use two 3D U-Net architectures to train on Positron Emission Tomography (PET) images and Computed Tomography (CT) images to capture implicit and informative high-level features followed by graph cut[126] to refine the segmentation. Furthermore, Guo et al.[104] report a Deep LOGISMOS approach that combines FCN and layered optimal graph image segmentation of multiple objects and surfaces(LOGISMOS). The LOGISMOS is based on the algorithmic incorporation of multiple spatial interrelationships in a single n-dimensional graph, followed by graph optimization that yields a globally optimal solution. Unlike FCN + graph methods, this framework directly constructs the graph based on the UNet-derived object boundaries and assigns UNet-derived probabilities as costs.

Hierarchical networks: The main idea of hierarchical networks is to realize coarse-to-fine segmentation by multiple stages. Chen et al.[11] propose a new method based on a cascaded 3D FCN, where the first stage is used to predict the region of the interest (ROI) of the target region, while the second stage is learned to predict the final segmentation. Rens et al.[57] introduced cascaded 3D FCN, which consists of a 3D localization FCN and a 3D segmentation FCN. The localization FCN is a regression 3D FCN and finds the bounding box of the ROI. The segmentation FCN is a 3D U-Net like FCN which performs volume-wise labelling. Rothet et al.[118] show a multi-class 3D FCN for competitive segmentation. The second stage FCN is fine-tuned from a first-stage FCN in a hierarchical manner and focused more on the boundary regions. Similarly, Sergi et al.[135] propose a cascade of two 3D patch-wise CNNs. The first network is trained to be more sensitive revealing possible candidate lesion voxels while the second network is trained to reduce the number of misclassified voxels coming from the first network. Yang et al.[166][167] combine 3D FCN and hierarchical deep supervision to segmentation on volumetric ultrasound and prenatal volumetric ultrasound. Furthermore, to address the problem that traditional 3DFCN cannot segment small ROIs, Holger et al.[119] proposed a two-stage, coarse-to-fine approach that first uses a 3D FCN to roughly define a candidate region, which is then used as input to a second 3D FCN.

7 EXPERIMENTAL RESULTS

Below we summarize the quantitative results of the segmentation methods discussed in Sections 3, 4 and 5 on some typical public datasets, as well as analyze these results qualitatively.

7.1 Results for 3D Semantic Segmentation

We report the results of RGB-D based semantic segmentation methods on SUN-RGB-D [127] and NYUDv2 [124] datasets using mAcc (mean Accuracy) and mIoU (mean Intersection over Union) as the evaluation metrics. These results of various methods are taken from the original papers and they are shown in Table 7. The Table shows below.

Table 7. Evaluation performance regarding for RGB-D semantic segmentation methods on the SUN-RGB-D and NYUDv2. Note that the ‘%’ after the value is omitted and the symbol ‘–’ means the results are unavailable.

Method	NYUDv2		SUN-RGB-D	
	mAcc	mIoU	mAcc	mIoU
Guo et al.[36]	46.3	34.8	45.7	33.7
Wang et al.[141]	–	44.2	–	–
Mousavian et al.[101]	52.3	39.2	–	–
Liu et al.[87]	50.8	39.8	50.0	39.4
Gupta et al.[38]	35.1	28.6	–	–
Hong et al.[86]	51.7	41.2	–	–
Hazirbas et al.[42]	–	–	48.3	37.3
Lin et al.[82]	–	47.7	–	48.1
Jiang et al.[61]	–	–	50.6	39.3
Wang et al.[144]	47.3	–	–	–
Cheng et al.[14]	60.7	45.9	58.0	–
Fan et al.[29]	50.2	–	–	–
Li et al.[79]	49.4	–	48.1	–
Qi et al.[110]	55.7	43.1	57.0	45.9
Wang et al.[139]	60.6	38.3	50.1	33.5

We report the results of projected images/voxel/point clouds/other representation semantic segmentation methods on S3DIS [1] (both Area 5 and 6-fold cross validation), ScanNet [20] (test sets), Semantic3D [39] (reduced-8 subsets) and SemanticKITTI [3] (only xyz without RGB). We use mAcc, oAcc (overall accuracy) and mIoU as the evaluation metrics. These results of various methods are taken from the original papers. Table 8 reports the results.

Sine our main interest is in point-based semantic segmentation methods, we focus on detailed analysis of performance of these methods. To capture wider context features and richer local features that are crucial for semantic segmentation performance, several dedicated strategies have been proposed on the basic framework. We present an analysis on these frameworks and strategies below.

(1) Basic framework: Basic networks are one of the main dirving forces behind the development of 3D segmentation. Generally, there are two main basic framework including PointNet and PointNet++ framework, and the shortcomings of them also point out the direction of improvement. We discuss as follow.

- *PN framework* PN employs shared MLP to exploit the point-wise feature and adopt maxpooling to collect these features to a global feature representation. But it cannot learn local features because of the lack of local neighborhood definition. On the other hand, the fixed resolution of the feature map makes the network difficult to suit deep architectures. Based on PN framework, a new version of framework, called spatial CNN-like framework, is proposed. It inserts coarsening layer after features exploitation layer to progressively decrease the point resolution.
- *PN++ framwork* PN++ creatively proposes a hierarchical learning architecture. It hierarchically defines local region and progressively extract feature in local regions. Yet it still lacks the dependency information of local

Table 8. Evaluation performance regarding for projected images, voxel, point clouds and other representation semantic segmentation methods on the S3DIS, ScanNet, Semantic3D and SemanticKITTI. Note: the ‘%’ after the value is omitted, the symbol ‘–’ means the results are unavailable, the dotted line means the subdivision of methods according to the type of architecture.

Method	Type	S3DIS		ScanNet		Semantic3D		SemanticKITTI	
		Area5 mAcc	6-fold mIoU	test set oAcc	mIoU	reduced-8 oAcc	mIoU	only xyz mAcc	mIoU
projected images		–	–	–	–	88.9	58.5	–	–
Boulch et al.[6][5]		–	–	–	–	91.0	67.4	–	–
Wu et al.[152]		–	–	–	–	–	–	–	37.2
Wang et al.[148]		–	–	–	–	–	–	–	39.8
Wu et al.[153]		–	–	–	–	–	–	–	44.9
Milioto et al.[98]		–	–	–	–	–	–	–	52.2
Xu et al.[160]		–	–	–	–	–	–	–	55.9
Tchapmi et al.[132]	voxel	57.35	48.92	48.92	–	88.1	61.30	–	–
Meng et al.[96]		–	78.22	–	–	–	–	–	–
Liu et al.[84]		–	70.76	–	–	–	–	–	–
PointNet[108]	point	48.98	41.09	47.71	–	14.69	–	29.9	17.9
G-RCU[28]		59.10	52.17	58.27	75.53	–	–	57.59	29.9
ESC[26]		54.06	45.14	49.7	63.4	–	–	40.9	26.4
HRNN[168]		71.3	53.4	–	76.5	–	–	49.2	34.5
PointNet++[109]		–	50.04	54.4	71.40	34.26	–	–	–
PointWeb[177]		66.64	60.28	66.7	85.9	–	–	–	–
PointSIFT[63]		–	70.23	70.2	–	41.5	–	–	–
RSNet[52]		59.42	56.5	56.47	–	39.35	–	–	–
DPC[27]		68.38	61.28	–	–	59.2	–	–	–
PointwiseCNN[50]		56.5	–	–	–	–	–	–	–
PCNN[143]		67.01	58.27	–	–	49.8	–	–	–
KPConv[133]		–	67.1	70.6	–	68.4	92.9	74.6	–
PointCNN[77]		63.86	57.26	65.3	85.1	45.8	–	–	–
PointConv[154]		–	50.34	–	–	55.6	–	–	–
A-CNN[67]		–	–	–	85.4	–	–	–	–
RandLA-Net[48]		–	–	70.0	–	–	94.8	77.4	–
PolarNet[175]		–	–	–	–	–	–	–	54.3
DGCN[149]		–	56.1	56.1	–	–	–	–	–
SPG[69]		66.50	58.04	62.1	–	–	94.0	73.2	–
SPH3D-GCN[73]		65.9	59.5	68.9	–	61.0	–	–	–
DeepGCNs[74]		–	60.0	–	–	–	–	–	–
PointGCRNet[93]		–	52.43	–	–	60.8	–	–	–
AGCN[159]		–	–	56.63	–	–	–	–	–
PAN[30]	others	–	66.3	–	86.7	42.1	–	–	–
TGNet[78]		–	58.7	–	66.2	–	–	–	–
HDGCN[80]		65.81	59.33	66.85	–	–	–	–	–
3DContextNet[174]		74.5	55.6	55.6	–	–	–	–	–
TangentConv[131]		62.2	52.8	–	80.1	40.9	89.3	66.4	–
SPLATNet[129]	–	–	–	–	39.3	–	–	–	–
LatticeNet[116]	–	–	–	–	64.0	–	–	–	52.9
Hung et al.[15]	–	–	–	–	63.4	–	–	–	–
PVCNN[90]	87.12	58.98	–	–	–	–	–	–	–
MVPNet[58]	–	–	–	–	66.4	–	–	–	–

points. To some extent, the PN++ framework is also similar to spatial CNN-like framework. Differently, the coarsening layer of PN++ framework is before feature exploitation layer.

(2) Local feature exploitation: Objects in their natural environment usually have various shapes. Local features can enhance detailed segmentation of objects as follows:

- *Neighborhood search*: Neighborhood definition is the premise of local feature extraction. Most existing methods adapt k-nearest neighbor search which is sensitive to variations in point density. Therefore, novel neighborhood search methods have recently been proposed.
- *Efficient local feature extractor*: There are two main feature extractors, MLPs (simplified PN) and convolution. Since MLPs lack the complexity to exploit robust local features, point convolution and graph convolution are becoming the more popular.

- *Attention-based aggregation*: The contribution of neighborhood points or local features is different because of variations in their distributions. Therefore, attention mechanism is used to learn their contributions.
- *Local-global concatenation*: Global features are sensitive to object layout but can guide local features to recover details. Hence, global features are concatenated with local ones to get the best of both worlds.

(3) **Context feature exploitation**: Objects in the 3D scene may be positioned according to some relationship with other objects in the environment. It has been established that the context features (referring to object dependency) can improve the accuracy of semantic segmentation, especially for small and similar objects. Here a summary on the main strategies.

- *Multi-scale/resolution grouping*: Single scale grouping is difficult to obtain context features of objects with fixed scale. Multi-scale/resolution grouping is simple, yet effectively helps to enlarge small objects and to narrow down large objects.
- *Dilated aggregation*: Dilated mechanism is usually combined with neighborhood search. It expands the broader context and maintains the feature map resolution.
- *Recurrent Neural Networks*: RNNs have powerful context features exploitation capability and is a popular candidate for context features extraction.
- *graph*: Graph construction on point clouds establishes the connection of points or objects. It is more effective to extract the dependency of key neighborhood points and the edge feature of objects.

7.2 Results for 3D Instance Segmentation

We report the results of 3D instance segmentation methods on ScanNet[20] datasets, and choose mAP as the evaluation metrics. These results of these methods are taken from the ScanNet Benchmark Challenge website, and they are shown in Table 9 and summarized in Fig.9. The table and figure shows that:

Table 9. Evaluation performance regarding for 3D instance segmentation methods on the ScanNet. Note: the ‘%’ after the value is omitted.

Method	mAP	bathtub	bed	bookshelf	cabinet	chair	counter	curtain	desk	door	other	picture	refrig.	showercur.	sink	sofa	table	toilet	window
GSPN[171]	30.6	50.0	40.5	31.1	34.8	58.9	5.4	6.8	12.6	28.3	29.0	2.8	21.9	21.4	33.1	39.6	27.5	82.1	24.5
3D-SIS[47]	38.2	100	43.2	24.5	19.0	57.7	1.3	26.3	3.3	32.0	24.0	7.5	42.2	85.7	11.7	69.9	27.1	83.3	23.5
3D-BoNet[164]	48.8	100	67.2	59.0	30.1	48.4	9.8	62.0	30.6	34.1	25.9	12.5	43.4	79.6	40.2	49.9	51.3	90.9	43.9
SGPN[145]	14.3	20.8	39.0	16.9	6.5	27.5	2.9	6.9	0	8.7	4.3	1.4	2.7	0	11.2	35.1	16.8	43.8	13.8
3D-MPA[25]	61.1	100	83.3	76.5	52.6	75.6	13.6	58.8	47.0	43.8	43.2	35.8	65.0	85.7	42.9	76.5	55.7	100	43.0
MASC[83]	44.7	52.8	55.5	38.1	38.2	63.3	0.2	50.9	26.0	36.1	43.2	32.7	45.1	57.1	36.7	63.9	38.6	98.0	27.6
MTML[68]	54.9	100	80.7	58.8	32.7	64.7	0.4	81.5	18.0	41.8	36.4	18.2	44.5	100	44.2	68.8	57.1	100	39.6
PointGroup[62]	63.6	100	76.5	62.4	50.5	79.7	11.6	69.6	38.4	44.1	55.9	47.6	59.6	100	66.6	75.6	55.6	99.7	51.3
3D-BEVIS[23]	24.8	66.7	56.6	7.6	3.5	39.4	2.7	3.5	9.8	9.8	3.0	2.5	9.8	37.5	12.6	60.4	18.1	85.4	17.1
PanopticFus.[102]	47.8	66.7	71.2	59.5	25.9	55.0	0	61.3	17.5	25.0	43.4	43.7	41.1	85.7	48.5	59.1	26.7	94.4	35.9
3D-GEL[81]	45.9	100	73.7	15.9	25.9	58.7	13.8	47.5	21.7	41.6	40.8	12.8	31.5	71.4	41.1	53.6	59.0	87.3	30.4
OccuSeg[40]	67.2	100	75.8	68.2	57.6	84.2	47.4	50.4	52.4	56.7	58.5	45.1	55.7	100	75.1	79.7	56.3	100	46.7

- OccuSeg[40] has the state-of-the-art performance, with 67.2% average precision on ScanNet dataset at the time of this view. It also achieves the best instance segmentation performance on most classes, including ‘bathtub’, ‘chair’, ‘shower curtain’, ‘sink’, ‘sofa’, ‘toilet’ and so on.
- Most methods have better segmentation performance on large scale classes such as ‘bathtub’ and ‘toilet’, and have poor segmentation performance on small scale classes such as ‘counter’, ‘desk’ and ‘picture’. Therefore, the instance segmentation of small objects is a prominent challenge.

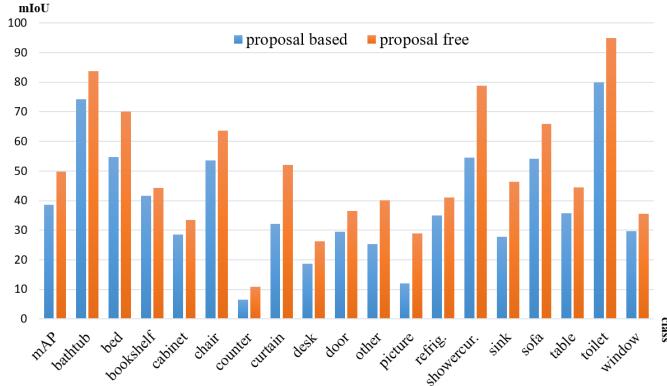


Fig. 9. Evaluation performance regarding for 3D instance segmentation architecture, including proposal based and proposal free, on the different class of ScanNet. For simplicity, we omit the ‘%’ after the value.

- Proposal free methods have better performance than proposal based methods on instance segmentation of all classes, especially for small objects such as ‘curtain’, ‘other’, ‘picture’, ‘shower curtain’ and ‘sink’.
- In the proposal based methods, the 2D embedding propagating based methods, including 3D-BEVIS[23], PanoticFusion[102], have poor performance compared to the other proposal free based methods. Simple embedding propagating is prone to error labels.

7.3 Results for 3D Part Segmentation

We report the results of 3D part segmentation methods on ShapeNet[169] datasets and use Ins. mIoU as the evaluation metric. These results of various methods are taken from the original papers and they are shown in Table 10. We can see that:

Table 10. Evaluation performance regarding for 3D part segmentation on the ShapeNet. Note: the ‘%’ after the value is omitted, the symbol ‘–’ means the results are unavailable.

Method	Ins. mIoU	Method	Ins. mIoU
VV-Net[96]	87.4	LatticeNet[129]	93.9
SSCNet[33]	86.0	SGPN[145]	85.8
PointNet[108]	83.7	ShapePFCN[64]	88.4
PointNet++[109]	85.1	VoxSegNet[150]	87.5
3DContextNet[174]	84.3	Pointgrid[71]	86.4
RSNet[52]	84.9	SubvolumeSup[128]	–
KPConv[133]	86.4	DCN[161]	–
PointwiseCNN[50]	–	MeshCNN[41]	–
MCC[45]	85.9	SO-Net[75]	84.9
PointConv[154]	85.7	PartNet[172]	87.4
PointCNN[77]	86.1	3DPointCapsNet[178]	–
DGCN[149]	85.1	SyncSpecCNN[170]	84.7
SPH3D-GCN[73]	86.8	KCNet[170]	84.7
AGCN[159]	85.4	SFCN[140]	–
PCNN[143]	85.9	SpiderCNN[163]	85.3
Flex-Conv[34]	85.0	FeaSTNet[136]	81.5
ψ -CNN[72]	86.8	Kd-Net[65]	82.3
SPLATNet[129]	84.6	O-CNN[142]	85.9

- LatticeNet[40] has the state-of-the-art performance, with 93.9% average precision on ShapeNet dataset at the time of this view.

- Part segmentation performance of all methods is quite similar.

8 DISCUSSION AND CONCLUSION

We provided a comprehensive survey of the recent development in 3D segmentation using deep learning techniques, including 3D semantic segmentation, 3D instance segmentation and 3D part segmentation. We presented a comprehensive performance comparison and merit of various methods in each category. 3D segmentation using deep learning techniques has made significant progress during recent years. However, this is just the beginning and significant developments lie ahead of us. Below, we present some outstanding issues and identify potential research directions.

- **Synthetic datasets with richer information for multiple tasks:** Synthetic datasets gradually play an important role on semantic segmentation due to the low cost and diverse scenes that can be generated [7],[155] compared to real datasets [20],[1],[39]. It is well known that the information contained in training data determine the upper limit of the scene parsing accuracy. Existing datasets lack important semantic information, such as material, and texture information, which is more crucial for segmentation with similar color or geometric information. Besides, most exiting datasets are generally designed for a single task. Currently, only a few semantic segmentation datasets also contain labels for instances [20] and scene layout [127] to meet the multi-task objective.
- **Unified network for multiple tasks:** It is expensive and impractical for a system to accomplish different computer vision tasks by various deep learning networks. Towards fundamental feature exploitation of scene, Semantic segmentation has strong consistency with some tasks, such as depth estimation [97],[85],[36],[141],[141],[87], scene completion [22], instance segmentation [146],[107],[81], and object detection [97]. These tasks could cooperate with each other to improve performance in a unified network. The semantic/instance segmentation can be further combined with part segmentation and other computer vision tasks for joint learning.
- **Multiple modals for scene parsing:** Semantic segmentation using multiple different representations, e.g. projected images, voxels and point clouds, can potentially achieve higher accuracy. However, single representation limits the segmentation accuracy because of the limitation of scene information, such as the less geometric information of images, the less semantic information of voxels. Multiple representations (multiple modals) would be an alternative way to enhance performance [21],[15],[90],[58],[97].
- **Efficient point convolution based network:** Point-based semantic segmentation networks are becoming the most investigated methods these days. These methods are devoted to fully explore the point-wise features and connections among points/features. However, they resort to neighborhood search mechanisms e.g. KNN, ball query [109], and hierarchical framework [154], which easily misses low-level features between local regions and further increases the difficulty of global context feature exploitation.
- **Weakly-supervised and unsupervised 3D segmentation:** Deep learning has gained significant success in 3D segmentation, but heavily hinges on large-scale labelled training samples. Weakly-supervised and unsupervised learning paradigms are considered as an alternative to relax the impractical requirement of large-scale labelled datasets. Currently, the work [162] proposes a weakly-supervised network that only needs labels for a small part of the training samples. The works [75],[178] propose a unsupervised network that generates supervision labels

from the data itself.

- **Real-time 3D semantic segmentation:** Real-time 3D scene parsing is crucial for some applications such as autonomous driving and mobile robots. However, most existing 3D semantic segmentation methods mainly focus on the improvement of segmentation accuracy but rarely focus on real-time performance. A few lightweight 3D semantic segmentation networks realize real-time by pre-processing point clouds into other presentations such as projected images [152],[148],[153],[98],[160]. However, such a pipeline is likely to miss a large part of the geometric information. Therefore, real-time 3D semantic segmentation methods based on point clouds requires more attention in the future.
- **Semantic segmentation of large-scale scenes** has always been a research hot spot. Existing approaches are limited to extremely small 3D point clouds [108],[69] (e.g. 4096 points or 1x1 meter blocks) and cannot be directly extended to larger scale point clouds (e.g. millions of points or hundreds of meters) without data preprocessing. Although RandLA-Net [48] can process one million points directly, this speed is still insufficient and there is a need to further investigate the problem of efficient semantic segmentation on large scale point clouds.
- **3D video semantic segmentation:** Like 2D video semantic segmentation, A handful of works try to exploit 4D spatio-temporal features on 3D videos (also call 4D point clouds) [17],[122]. From these works, it can be seen that the spatio-temporal features can help improve the robustness of 3D video or dynamic 3D scene semantic segmentation.

REFERENCES

- [1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1534–1543.
- [2] Christian F Baumgartner, Lisa M Koch, Marc Pollefeys, and Ender Konukoglu. 2017. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 111–119.
- [3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jürgen Gall. 2019. SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 9297–9307.
- [4] Saifullahi Aminu Bello, Shangshu Yu, Cheng Wang, Jibril Muhammad Adam, and Jonathan Li. 2020. deep learning on 3D point clouds. *Remote Sensing* 12, 11 (2020), 1729.
- [5] Alexandre Boulch, Joris Guerry, Bertrand Le Saux, and Nicolas Audebert. 2018. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Computers & Graphics* 71 (2018), 189–198.
- [6] Alexandre Boulch, Bertrand Le Saux, and Nicolas Audebert. 2017. Unstructured Point Cloud Semantic Labeling Using Deep Segmentation Networks. *3DOR* 2 (2017), 7.
- [7] Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, Luca Celotti, Florian Strub, Jean Rouat, Hugo Larochelle, and Aaron Courville. 2017. Home: A household multimodal environment. *arXiv preprint arXiv:1711.11017* (2017).
- [8] Yuanzhouhan Cao, Chunhua Shen, and Heng Tao Shen. 2016. Exploiting depth from single monocular images for object detection and semantic segmentation. *IEEE Transactions on Image Processing* 26, 2 (2016), 836–846.
- [9] Adria Casamitjana, Santi Puch, Asier Aduriz, and Verónica Vilaplana. 2016. 3D Convolutional Neural Networks for Brain Tumor Segmentation: a comparison of multi-resolution architectures. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer, 150–161.
- [10] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2018. Matterport3d: Learning from rgb-d data in indoor environments. (2018).
- [11] Shuqing Chen, Holger Roth, Sabrina Dorn, Matthias May, Alexander Cavallaro, Michael M Lell, Marc Kachelrieß, Hirohisa Oda, Kensaku Mori, and Andreas Maier. 2017. Towards automatic abdominal multi-organ segmentation in dual energy CT using cascaded 3D fully convolutional network. *arXiv preprint arXiv:1710.05379* (2017).
- [12] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. 2009. A benchmark for 3D mesh segmentation. *Acm transactions on graphics (tog)* 28, 3 (2009), 1–12.

- [13] Jiyu Cheng, Yuxiang Sun, and Max Q-H Meng. 2020. Robust Semantic Mapping in Challenging Environments. *Robotica* 38, 2 (2020), 256–270.
- [14] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. 2017. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3029–3037.
- [15] Hung-Yueh Chiang, Yen-Liang Lin, Yueh-Cheng Liu, and Winston H Hsu. 2019. A unified point-based framework for 3d segmentation. In *2019 International Conference on 3D Vision (3DV)*. IEEE, 155–163.
- [16] François Fleuret. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.
- [17] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3075–3084.
- [18] Özgür Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*. Springer, 424–432.
- [19] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. 2013. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572* (2013).
- [20] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5828–5839.
- [21] Angela Dai and Matthias Nießner. 2018. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 452–468.
- [22] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. 2018. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4578–4587.
- [23] Cathrin Elich, Francis Engelmann, Theodora Kontogianni, and Bastian Leibe. 2019. 3D Bird’s-Eye-View Instance Segmentation. In *German Conference on Pattern Recognition*. Springer, 48–61.
- [24] Ekrem Emre Yurdakul and Yucel Yemez. 2017. Semantic segmentation of rgbd videos with recurrent fully convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 367–374.
- [25] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 2020. 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9031–9040.
- [26] Francis Engelmann, Theodora Kontogianni, Alexander Hermans, and Bastian Leibe. 2017. Exploring spatial context for 3D semantic segmentation of point clouds. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 716–724.
- [27] Francis Engelmann, Theodora Kontogianni, and Bastian Leibe. 2020. Dilated Point Convolutions: On the Receptive Field Size of Point Convolutions on 3D Point Clouds. In *International Conference on Robotics and Automation (ICRA)*, Vol. 1.
- [28] Francis Engelmann, Theodora Kontogianni, Jonas Schult, and Bastian Leibe. 2018. Know what your neighbors do: 3D semantic segmentation of point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 0–0.
- [29] Heng Fan, Xue Mei, Danil Prokhorov, and Haibin Ling. 2017. RGB-D scene labeling with multimodal recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 9–17.
- [30] Mingtao Feng, Liang Zhang, Xuefei Lin, Syed Zulqarnain Gilani, and Ajmal Mian. 2020. Point Attention Network for Semantic Segmentation of 3D Point Clouds. *Pattern Recognition* (2020), 107446.
- [31] Fahimeh Fooladgar and Shohreh Kasaei. 2020. A survey on indoor RGB-D semantic segmentation: from hand-crafted features to deep convolutional neural networks. *Multimedia Tools and Applications* 79, 7 (2020), 4499–4524.
- [32] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3354–3361.
- [33] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9224–9232.
- [34] Fabian Groh, Patrick Wieschollek, and Hendrik PA Lensch. 2018. Flex-convolution. In *Asian Conference on Computer Vision*. Springer, 105–122.
- [35] Joris Guerry, Alexandre Boulch, Bertrand Le Saux, Julien Moras, Aurélien Plyer, and David Filliat. 2017. Snapnet-r: Consistent 3d multi-view semantic labeling for robotics. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 669–678.
- [36] Yanrong Guo and Tao Chen. 2018. Semantic segmentation of RGBD images based on deep depth regression. *Pattern Recognition Letters* 109 (2018), 55–64.
- [37] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence* (2020).
- [38] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. [n.d.]. Learning Rich Features from RGB-D Images for Object Detection and Segmentation: Supplementary Material. ([n. d.]).
- [39] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan D Wegner, Konrad Schindler, and Marc Pollefeys. 2017. Semantic3d.net: A new large-scale point cloud classification benchmark. *arXiv preprint arXiv:1704.03847* (2017).
- [40] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. 2020. OccupSeg: Occupancy-aware 3D Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2940–2949.
- [41] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. 2019. MeshCNN: a network with an edge. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

- [42] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*. Springer, 213–228.
- [43] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.
- [44] Yang He, Wei-Chen Chiu, Margret Keuper, and Mario Fritz. 2017. Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4837–4846.
- [45] Pedro Hermosilla, Tobias Ritschel, Pere-Pau Vázquez, Álvar Vinacua, and Timo Ropinski. 2018. Monte carlo convolution for learning on non-uniformly sampled point clouds. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–12.
- [46] Nico Höft, Hannes Schulz, and Sven Behnke. 2014. Fast semantic segmentation of RGB-D scenes with GPU-accelerated deep neural networks. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, 80–85.
- [47] Ji Hou, Angela Dai, and Matthias Nießner. 2019. 3d-sis: 3d semantic instance segmentation of rgbd scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4421–4430.
- [48] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. 2020. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11108–11117.
- [49] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. 2016. Scenenn: A scene meshes dataset with annotations. In *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 92–101.
- [50] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. 2018. Pointwise convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 984–993.
- [51] Jing Huang and Suya You. 2016. Point cloud labeling using 3d convolutional neural network. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2670–2675.
- [52] Qiangui Huang, Weiyue Wang, and Ulrich Neumann. 2018. Recurrent slice networks for 3d segmentation of point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2626–2635.
- [53] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. *arXiv preprint arXiv:1602.07360* (2016).
- [54] Anastasia Ioannidou, Elisavet Chatzilari, Spiros Nikolopoulos, and Ioannis Kompatsiaris. 2017. Deep learning advances in computer vision with 3d data: A survey. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 1–38.
- [55] Bc Ján Ivanecký. 2016. *Depth estimation by convolutional neural networks*. Ph.D. Dissertation. Master thesis, Brno University of Technology.
- [56] Varun Jampani, Martin Kiefel, and Peter V Gehler. 2016. Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4452–4461.
- [57] Rens Janssens, Guodong Zeng, and Guoyan Zheng. 2018. Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 893–897.
- [58] Maximilian Jaritz, Jiayuan Gu, and Hao Su. 2019. Multi-view pointnet for 3d scene understanding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.
- [59] Jongmin Jeong, Tae Sung Yoon, and Jin Bae Park. 2018. Multimodal sensor-based semantic 3d mapping for a large-scale environment. *Expert Systems with Applications* 105 (2018), 1–10.
- [60] Haiyong Jiang, Feilong Yan, Jianfei Cai, Jianmin Zheng, and Jun Xiao. 2020. End-to-End 3D Point Cloud Instance Segmentation Without Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12796–12805.
- [61] Jindong Jiang, Zhijun Zhang, Yongqian Huang, and Lunan Zheng. 2017. Incorporating depth into both cnn and crf for indoor semantic segmentation. In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 525–530.
- [62] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. 2020. PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4867–4876.
- [63] Mingyang Jiang, Yiran Wu, Tianqi Zhao, Zelin Zhao, and Cewu Lu. 2018. Pointsift: A sift-like network module for 3d point cloud semantic segmentation. *arXiv preprint arXiv:1807.00652* (2018).
- [64] Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, and Siddhartha Chaudhuri. 2017. 3D shape segmentation with projective convolutional networks. In *proceedings of the IEEE conference on computer vision and pattern recognition*. 3779–3788.
- [65] Roman Klokov and Victor Lempitsky. 2017. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *Proceedings of the IEEE International Conference on Computer Vision*. 863–872.
- [66] Deyvid Kochanov, Aljoša Ošep, Jörg Stückler, and Bastian Leibe. 2016. Scene flow propagation for semantic mapping and object discovery in dynamic street scenes. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1785–1792.
- [67] Artem Komarichev, Zichun Zhong, and Jing Hua. 2019. A-CNN: Annularly convolutional neural networks on point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7421–7430.
- [68] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 2019. 3d instance segmentation via multi-task metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 9256–9266.
- [69] Loic Landrieu and Martin Simonovsky. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4558–4567.

- [70] Felix Järemo Lawin, Martin Danelljan, Patrik Tosteberg, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. 2017. Deep projective 3D semantic segmentation. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 95–107.
- [71] Truc Le and Ye Duan. 2018. Pointgrid: A deep network for 3d shape understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9204–9214.
- [72] Huan Lei, Naveed Akhtar, and Ajmal Mian. 2019. Octree guided CNN with spherical kernels for 3D point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9631–9640.
- [73] Huan Lei, Naveed Akhtar, and Ajmal Mian. 2020. Spherical kernel for efficient graph convolution on 3d point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
- [74] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. 2019. Deepgcns: Can gcns go as deep as cnns?. In *Proceedings of the IEEE International Conference on Computer Vision*. 9267–9276.
- [75] Jiaxin Li, Ben M Chen, and Gim Hee Lee. 2018. So-net: Self-organizing network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9397–9406.
- [76] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. 2018. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE transactions on medical imaging* 37, 12 (2018), 2663–2674.
- [77] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhua Di, and Baoquan Chen. 2018. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems* 31 (2018), 820–830.
- [78] Ying Li, Lingfei Ma, Zilong Zhong, Dongpu Cao, and Jonathan Li. 2019. Tgnet: Geometric graph cnn on 3-d point cloud segmentation. *IEEE Transactions on Geoscience and Remote Sensing* 58, 5 (2019), 3588–3600.
- [79] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. 2016. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In *European conference on computer vision*. Springer, 541–557.
- [80] Zhidong Liang, Ming Yang, Liuyuan Deng, Chunxiang Wang, and Bing Wang. 2019. Hierarchical depthwise graph convolutional neural network for 3d semantic segmentation of point clouds. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 8152–8158.
- [81] Zhidong Liang, Ming Yang, and Chunxiang Wang. 2019. 3D graph embedding learning with a structure-aware loss function for point cloud semantic instance segmentation. *arXiv preprint arXiv:1902.05247* (2019).
- [82] Di Lin, Guangyong Chen, Daniel Cohen-Or, Pheng-Ann Heng, and Hui Huang. 2017. Cascaded feature network for semantic segmentation of rgb-d images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1311–1319.
- [83] Chen Liu and Yasutaka Furukawa. 2019. MASC: multi-scale affinity with sparse convolution for 3d instance segmentation. *arXiv preprint arXiv:1902.04478* (2019).
- [84] Fangyu Liu, Shuaipeng Li, Liqiang Zhang, Chenghu Zhou, Rongtian Ye, Yuebin Wang, and Jiwen Lu. 2017. 3DCNN-DQN-RNN: A deep reinforcement learning framework for semantic parsing of large-scale 3D point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*. 5678–5687.
- [85] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. 2015. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence* 38, 10 (2015), 2024–2039.
- [86] Hong Liu, Wenshan Wu, Xiangdong Wang, and Yueliang Qian. 2018. RGB-D joint modelling with scene geometric information for indoor semantic segmentation. *Multimedia Tools and Applications* 77, 17 (2018), 22475–22488.
- [87] Jing Liu, Yuhang Wang, Yong Li, Jun Fu, Jiangyun Li, and Hanqing Lu. 2018. Collaborative deconvolutional neural networks for joint depth estimation and semantic segmentation. *IEEE transactions on neural networks and learning systems* 29, 11 (2018), 5655–5666.
- [88] Weiping Liu, Jia Sun, Wanyi Li, Ting Hu, and Peng Wang. 2019. Deep learning on point clouds and its application: A survey. *Sensors* 19, 19 (2019), 4188.
- [89] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. 2018. Affinity derivation and graph merge for instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 686–703.
- [90] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. 2019. Point-Voxel CNN for efficient 3D deep learning. In *Advances in Neural Information Processing Systems*. 965–975.
- [91] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.
- [92] Lingni Ma, Jörg Stückler, Christian Kerl, and Daniel Cremers. 2017. Multi-view deep learning for consistent semantic mapping with rgb-d cameras. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 598–605.
- [93] Yanni Ma, Yulan Guo, Hao Liu, Yinjie Lei, and Gongjian Wen. 2020. Global Context Reasoning for Semantic Segmentation of 3D Point Clouds. In *The IEEE Winter Conference on Applications of Computer Vision*. 2931–2940.
- [94] Daniel Maturana and Sebastian Scherer. 2015. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 922–928.
- [95] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. 2017. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)*. IEEE, 4628–4635.
- [96] Hsien-Yu Meng, Lin Gao, Yu-Kun Lai, and Dinesh Manocha. 2019. Vv-net: Voxel vae net with group convolutions for point cloud segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 8500–8508.
- [97] Gregory P Meyer, Jake Charland, Darshan Hegde, Ankit Laddha, and Carlos Valdes-Gonzalez. 2019. Sensor fusion for joint 3d object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

- [98] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. 2019. RangeNet++: Fast and accurate LiDAR semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4213–4220.
- [99] Miguel Monteiro, Mário AT Figueiredo, and Arlindo L Oliveira. 2018. Conditional random fields as recurrent neural networks for 3d medical imaging segmentation. *arXiv preprint arXiv:1807.07464* (2018).
- [100] Guy M Morton. 1966. A computer oriented geodetic data base and a new technique in file sequencing. (1966).
- [101] Arsalan Mousavian, Hamed Pirsiavash, and Jana Košecká. 2016. Joint semantic segmentation and depth estimation with deep convolutional networks. In *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 611–619.
- [102] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. 2019. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. *arXiv preprint arXiv:1903.01177* (2019).
- [103] Muzammal Naseer, Salman Khan, and Fatih Porikli. 2018. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE Access* 7 (2018), 1859–1887.
- [104] I Oguz, H Bogunović, S Kashyap, MD Abràmoff, X Wu, and M Sonka. 2016. LOGISMOS: A Family of Graph-Based Optimal Image segmentation methods. In *Medical Image Recognition, Segmentation and Parsing*. Elsevier, 179–208.
- [105] Jay Patravali, Shubham Jain, and Sasank Chilamkurthy. 2017. 2D-3D fully convolutional neural networks for cardiac MR segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 130–139.
- [106] Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 2019. Real-time progressive 3D semantic segmentation for indoor scenes. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1089–1098.
- [107] Quang-Hieu Pham, Thanh Nguyen, Binh-Son Hua, Gemma Roig, and Sai-Kit Yeung. 2019. JSIS3D: joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8827–8836.
- [108] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 652–660.
- [109] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* 30 (2017), 5099–5108.
- [110] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 2017. 3d graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 5199–5208.
- [111] Aman Raj, Daniel Maturana, and Sebastian Scherer. 2015. Multi-scale convolutional architecture for semantic segmentation. *Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RITR-15-21* (2015).
- [112] Dario Rethage, Johanna Wald, Jürgen Sturm, Nassir Navab, and Federico Tombari. 2018. Fully-convolutional point networks for large-scale point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 596–611.
- [113] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. 2017. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3577–3586.
- [114] Hayko Riemenschneider, András Bódis-Szomorú, Julien Weissenberg, and Luc Van Gool. 2014. Learning where to classify in multi-view semantic segmentation. In *European Conference on Computer Vision*. Springer, 516–532.
- [115] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
- [116] Radu Alexandru Rosu, Peer Schütt, Jan Quenzel, and Sven Behnke. 2019. Laticenet: Fast point cloud segmentation using permutohedral lattices. *arXiv preprint arXiv:1912.05905* (2019).
- [117] Holger Roth, Masahiro Oda, Natsuki Shimizu, Hirohisa Oda, Yuichiro Hayashi, Takayuki Kitasaka, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. 2018. Towards dense volumetric pancreas segmentation in CT using 3D fully convolutional networks. In *Medical Imaging 2018: Image Processing*, Vol. 10574. International Society for Optics and Photonics, 105740B.
- [118] Holger R Roth, Hirohisa Oda, Yuichiro Hayashi, Masahiro Oda, Natsuki Shimizu, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. 2017. Hierarchical 3D fully convolutional networks for multi-organ segmentation. *arXiv preprint arXiv:1704.06382* (2017).
- [119] Holger R Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. 2018. An application of cascaded 3D fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics* 66 (2018), 90–99.
- [120] Xavier Roynard, Jean-Emmanuel Deschaud, and François Fleuret. 2018. Paris-Lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *The International Journal of Robotics Research* 37, 6 (2018), 545–557.
- [121] Yiru Shen, Chen Feng, Yaoqing Yang, and Dong Tian. 2018. Mining point cloud local structures by kernel correlation and graph pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4548–4557.
- [122] Hanyu Shi, Guosheng Lin, Hao Wang, Tzu-Yi Hung, and Zhenhua Wang. 2020. SpSequenceNet: Semantic Segmentation Network on 4D Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4574–4583.
- [123] Nathan Silberman and Rob Fergus. 2011. Indoor scene segmentation using a structured light sensor. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 601–608.
- [124] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*. Springer, 746–760.

- [125] Martin Simonovsky and Nikos Komodakis. 2017. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3693–3702.
- [126] Qi Song, Junjie Bai, Dongfeng Han, Sudershan Bhatia, Wenqing Sun, William Rockey, John E Bayouth, John M Buatti, and Xiaodong Wu. 2013. Optimal co-segmentation of tumor in PET-CT images with context information. *IEEE transactions on medical imaging* 32, 9 (2013), 1685–1697.
- [127] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 567–576.
- [128] Yafei Song, Xiaowu Chen, Jia Li, and Qipeng Zhao. 2017. Embedding 3d geometric features for rigid object part segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 580–588.
- [129] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. 2018. Splatnet: Sparse lattice networks for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2530–2539.
- [130] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*. 945–953.
- [131] Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, and Qian-Yi Zhou. 2018. Tangent convolutions for dense prediction in 3d. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3887–3896.
- [132] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. 2017. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*. IEEE, 537–547.
- [133] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Fleuret, and Leonidas J Guibas. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*. 6411–6420.
- [134] Sepehr Valipour, Mennatullah Siam, Martin Jagersand, and Nilanjan Ray. 2017. Recurrent fully convolutional networks for video segmentation. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 29–36.
- [135] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage* 155 (2017), 159–168.
- [136] Nitika Verma, Edmond Boyer, and Jakob Verbeek. 2018. Feastnet: Feature-steered graph convolutions for 3d shape analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2598–2606.
- [137] Chu Wang, Babak Samari, and Kaleem Siddiqi. 2018. Local spectral graph convolution for point set feature learning. In *Proceedings of the European conference on computer vision (ECCV)*. 52–66.
- [138] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. 2018. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging* 37, 7 (2018), 1562–1573.
- [139] Jinghua Wang, Zhenhua Wang, Dacheng Tao, Simon See, and Gang Wang. 2016. Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In *European Conference on Computer Vision*. Springer, 664–679.
- [140] Pengyu Wang, Yuan Gan, Panpan Shui, Fenggen Yu, Yan Zhang, Songle Chen, and Zhengxing Sun. 2018. 3D shape segmentation via shape fully convolutional networks. *Computers & Graphics* 70 (2018), 128–139.
- [141] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. 2015. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2800–2809.
- [142] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. 2017. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–11.
- [143] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. 2018. Deep parametric continuous convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2589–2597.
- [144] Weiyue Wang and Ulrich Neumann. 2018. Depth-aware cnn for rgb-d segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 135–150.
- [145] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. 2018. Sgn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2569–2578.
- [146] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. 2019. Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4096–4105.
- [147] Yunhai Wang, Shmulik Asafi, Oliver Van Kaick, Hao Zhang, Daniel Cohen-Or, and Baoquan Chen. 2012. Active co-analysis of a set of shapes. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 1–10.
- [148] Yuan Wang, Tianyue Shi, Peng Yun, Lei Tai, and Ming Liu. 2018. Pointseg: Real-time semantic segmentation based on 3d lidar point cloud. *arXiv preprint arXiv:1807.06288* (2018).
- [149] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)* 38, 5 (2019), 1–12.
- [150] Zongji Wang and Feng Lu. 2019. VoxSegNet: Volumetric CNNs for semantic part segmentation of 3D shapes. *IEEE transactions on visualization and computer graphics* (2019).
- [151] Li-Yi Wei. 2008. Parallel Poisson disk sampling. *Acm Transactions On Graphics (tog)* 27, 3 (2008), 1–9.

- [152] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. 2018. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1887–1893.
- [153] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. 2019. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4376–4382.
- [154] Wenxuan Wu, Zhongang Qi, and Li Fuxin. 2019. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9621–9630.
- [155] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209* (2018).
- [156] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1912–1920.
- [157] Yu Xiang and Dieter Fox. 2017. Da-rnn: Semantic mapping with data associated recurrent neural networks. *arXiv preprint arXiv:1703.03098* (2017).
- [158] Yuxing Xie, TIAN JiaoJiao, and XiaoXiang Zhu. 2020. Linking Points With Labels in 3D: A Review of Point Cloud Semantic Segmentation. *IEEE Geoscience and Remote Sensing Magazine* (2020).
- [159] Zhuyang Xie, Junzhou Chen, and Bo Peng. 2020. Point clouds learning with attention-based graph convolution networks. *Neurocomputing* (2020).
- [160] Chengfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. 2020. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. *arXiv preprint arXiv:2004.01803* (2020).
- [161] Haotian Xu, Ming Dong, and Zichun Zhong. 2017. Directionally convolutional networks for 3d shape segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 2698–2707.
- [162] Xun Xu and Gim Hee Lee. 2020. Weakly Supervised Semantic Point Cloud Segmentation: Towards 10x Fewer Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13706–13715.
- [163] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. 2018. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 87–102.
- [164] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. 2019. Learning object bounding boxes for 3d instance segmentation on point clouds. *Advances in neural information processing systems* 32 (2019), 6740–6749.
- [165] Shichao Yang, Yulan Huang, and Sebastian Scherer. 2017. Semantic 3D occupancy mapping through efficient high order CRFs. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 590–597.
- [166] Xin Yang, Lequan Yu, Shengli Li, Xu Wang, Na Wang, Jing Qin, Dong Ni, and Pheng-Ann Heng. 2017. Towards automatic semantic segmentation in volumetric ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 711–719.
- [167] Xin Yang, Lequan Yu, Shengli Li, Huaxuan Wen, Dandan Luo, Cheng Bian, Jing Qin, Dong Ni, and Pheng-Ann Heng. 2018. Towards automated semantic segmentation in prenatal volumetric ultrasound. *IEEE transactions on medical imaging* 38, 1 (2018), 180–193.
- [168] Xiaoqing Ye, Jiamao Li, Hexiao Huang, Liang Du, and Xiaolin Zhang. 2018. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 403–417.
- [169] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. 2016. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)* 35, 6 (2016), 1–12.
- [170] Li Yi, Hao Su, Xingwen Guo, and Leonidas J Guibas. 2017. Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2282–2290.
- [171] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. 2019. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3947–3956.
- [172] Fenggen Yu, Kun Liu, Yan Zhang, Chenyang Zhu, and Kai Xu. 2019. Partnet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9491–9500.
- [173] Guodong Zeng, Xin Yang, Jing Li, Lequan Yu, Pheng-Ann Heng, and Guoyan Zheng. 2017. 3D U-net with multi-level deep supervision: fully automatic segmentation of proximal femur in 3D MR images. In *International workshop on machine learning in medical imaging*. Springer, 274–282.
- [174] Wei Zeng and Theo Gevers. 2018. 3DContextNet: Kd tree guided hierarchical learning of point clouds using local and global contextual cues. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 0–0.
- [175] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. 2020. PolarNet: An Improved Grid Representation for Online LiDAR Point Clouds Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9601–9610.
- [176] Cheng Zhao, Li Sun, Pulak Purkait, Tom Duckett, and Rustam Stolkin. 2018. Dense rgb-d semantic mapping with pixel-voxel neural network. *Sensors* 18, 9 (2018), 3099.
- [177] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. 2019. PointWeb: Enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5565–5573.
- [178] Yongheng Zhao, Tolga Birdal, Haowen Deng, and Federico Tombari. 2019. 3D point capsule networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1009–1018.
- [179] Zisha Zhong, Yusung Kim, Leixin Zhou, Kristin Plichta, Bryan Allen, John Buatti, and Xiaodong Wu. 2018. 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 228–231.