Review

Deep Learning Methods Applied to 3D Point Clouds Based Instance Segmentation: A Review

Desire Burume [1] and Shengzhi Du [2]

[1] Tshwane University of Technology; deziburume@gmail.com
[2] Tshwane University of Technology; sdu@gmail.com
* Correspondence: deziburume@gmail.com

Abstract: Beyond semantic segmentation,3D instance segmentation(a process to delineate objects of interest and also classifying the objects into a set of categories) is gaining more and more interest among researchers since numerous computer vision applications need accurate segmentation processes(autonomous driving, indoor navigation, and even virtual or augmented reality systems…) This paper gives an overview and a technical comparison of the existing deep learning architectures in handling unstructured Euclidean data for the rapidly developing 3D instance segmentation. First, the authors divide the 3D point clouds based instance segmentation techniques into two major categories which are proposal based methods and proposal free methods. Then, they also introduce and compare the most used datasets with regard to 3D instance segmentation. Furthermore, they compare and analyze these techniques performance (speed, accuracy, response to noise…). Finally, this paper provides a review of the possible future directions of deep learning for 3D sensor-based information and provides insight into the most promising areas for prospective research.

Keywords: Deep Learning; 3D Instance Segmentation; Datasets

1. Introduction

The instance segmentation is not an isolated domain of object detection but rather a natural evolution from semantic segmentation to fine refinement. The idea of instance segmentation emanates from classification which is essentially the process of predicting a whole input while classifying each object or alternatively naming a ranked list of the items in the input [1]. Localization (detection) is the following process of fine-grained regression yielding the classes and supplementary data such as those classes spatial position.

The rapid expansion of affordable 3D sensors of diverse types including 3D scanners, RGB-D cameras (Kinect, Apple depth cameras…) as well as LIDARS [2] have contributed immensely to the development of computer vision where semantic and instance segmentation are key components. Data acquired by these 3D sensors have the advantage of providing size information and rich geometric profile ([3], [4]). 3D data have the ability to better describe the surroundings for robots, for instance. Autonomous driving, medical treatment, remote sensing and robotics among others are good applications areas for 3D data [5].

3D instance segmentation is also closely connected to the tasks of 3D semantic segmentation and 3D object detection. 3D semantic segmentation serves to predict 3D points semantic labels, but it does not split dissimilar instances. Conversely, 3D object detection predicts the 3D bounding box of each particular object, but is unable to present a detailed mask of the 3D target object. Hence, 3D instance segmentation can safely be seen as an integrated task of 3D object detection and semantic segmentation.

In general, deep learning of 3D point clouds does encounter numerous considerable difficulties [5] such as the unstructured disposition of 3D point clouds and undersized datasets…In essence, this paper focuses on the study of the 3D point clouds based deep learning methods for instance segmentation. This review elaborates on the most used datasets for 3D segmentation such as NYUv2 [6], ShapeNet [7], ScanNet [8] to name a few. In fact, these datasets have immensely contributed to improve the research on 3D point clouds deep learning because they ushered in several new techniques dedicated to solving point cloud processing challenges including 3D segmentation and 3D object detection.

The simplest representation of 3D data is as a point cloud [6] which is a selection of 3D points each with three coordinates in a coordinate system (Cartesian or otherwise) [7]. Generally, point clouds only possess implicit neighborhood relations (unstructured data) contrary to 3D images which are structurally stored with clear neighborhood relationships [8]. From the few available review papers on deep learning on 3D data ([1], [8], [9], [10]), this paper is the first to specially center on the cross-analysis of deep learning methods for point clouds covering only 3D instance segmentation.

Compared to the existing literature, the major contributions of this paper are as follows:

1) As far as the author's knowledge, this paper is the first to cover exclusively and expansively the very important 3D point clouds instance segmentation task.
2) As opposed to existing reviews ([1], [8], [9], [10]), this paper exclusively centers on deep learning methods for 3D point clouds and not other types of 3D data.
3) This paper covers the most recent and advanced progress of deep learning for 3D instance segmentation on point clouds and the impact on their performance from data noise, scene variations, loss functions choices, hardware used…Therefore, it provides the reader with an all-round performances comparison of the 3D instance segmentation state-of-the-art methods.

The remainder of this paper is organized as follows. Firstly, section 2 introduces the 3D point clouds segmentation methods where the common deep learning networks architectures are explained. Next, section 3 describes existing benchmarks datasets, their focus and contents. Section 4 reviews and compares existing methods. Section 5 presents possible future research directions. Finally, section 6 summarizes the paper and draws conclusions.

2. 3D point clouds instance segmentation methods

2.1. Common deep learning networks architectures for raw point clouds processing

2.1.1. PointNet

Only Convolutional Neural Networks make it possible to learn the local regions using a hierarchical approach while the network drops deeper. Yet, convolution demands a structured network that point cloud data normally does not have. The inaugural framework that has used deep learning directly on unstructured point cloud is PointNet[11]. PointNet relies on two fundamental symmetric functions (a maxpooling function and a multilayer perceptron (MLP)) (Figure 1). Following the PointNet method, scores of techniques were introduced that consider the point clouds local structure.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

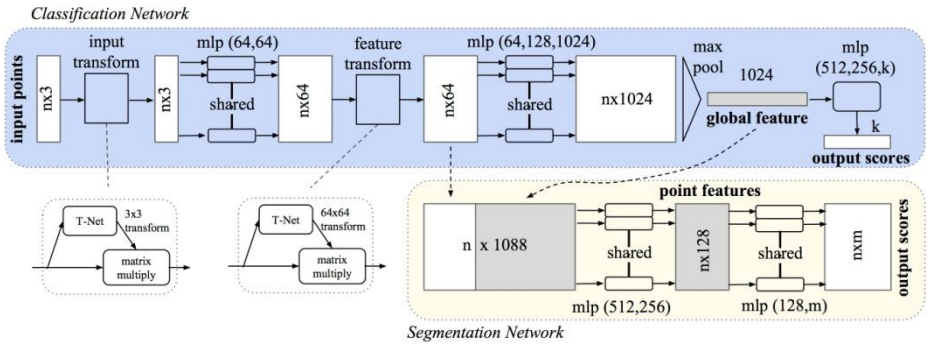


Figure 1. PointNet architecture

2.1.2. Other methods for raw point clouds processing

Many state-of-the-arts approaches are able to consider the point clouds local structure in a hierarchical fashion [12]. Since 3D instance segmentation is a natural evolution from semantic segmentation, this paper will endeavor to give a quick overview of the latter. In essence, 3D semantic segmentation methods can be divided in two broad categories such as point based and projection based (table 1 (a and b)). As stated before, 3D instance segmentation is a natural evolution from semantic segmentation as it aims to desalinate points in the 3D scene.

Table 1 a. Nomenclature of 3D semantic segmentation methods for projection based

Methods	Categories	Representative techniques
Projection based methods	Multi-view	DeePr3SS[13], SnapNet[14], TangentConv[15]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

	Spherical	SqueezeSeg[16], SqueezeSegV2[17], RangeNet++[18]
	Volumetric	SegCloud[19], SparseConvNet[20], MinkowskiNet[21], VV-Net[22]
	Permutohedral lattice	SPLATNet[23], LatticeNet[24]
	Hybrid	3DMV[158], UPB[25], MVPNet[26]

Table 1 b. Nomenclature of 3D semantic segmentation methods for point based

Methods	Categories	Representative techniques
Point based methods	Point-wise MLP	PointNet[11], Pointnet++[27], PointSIFT[28], Engelmann[29],

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

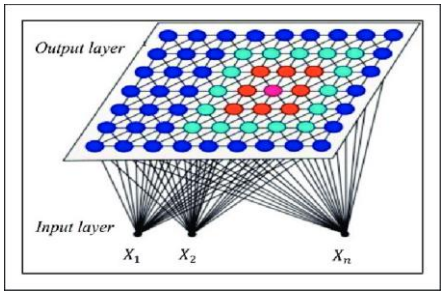
		3DContextNet[30], A-SCN[31], PointWeb[32], PAT[33], LSANet[34], Shell-Net[35], Rand-LA-Net[36]
	Point convolution	PointCNN[37], PCCN[38] A-CNN[39], ConvPoint[40], KPConv[41], DPC[42], InterpCNN[43]
	RNN based	RSNet[44], G+RCU[45], 3P-RNN[46]
	Graph based	DGCNN[47], SPG[48], SSP+SPG[49], GAC-Net[50], PAG[51], HDGCN[52], HPEIN[53], SPH3D-GCN[54], DPAM[55]

Because of the point cloud's lack of order, local structure definition relies on three fundamental actions: sampling, grouping and mapping of the nearest neighbor point features [56]. Sampling does decrease the amount of points in layers. Two main techniques for subsampling are used such as random point sampling (where every points is evenly prone to sampling) and farthest point sampling (where all the points are sampled according to the farthest point from the other points). There exist other sampling approaches like the Gumbel Subset Sampling and the uniform sampling [33].

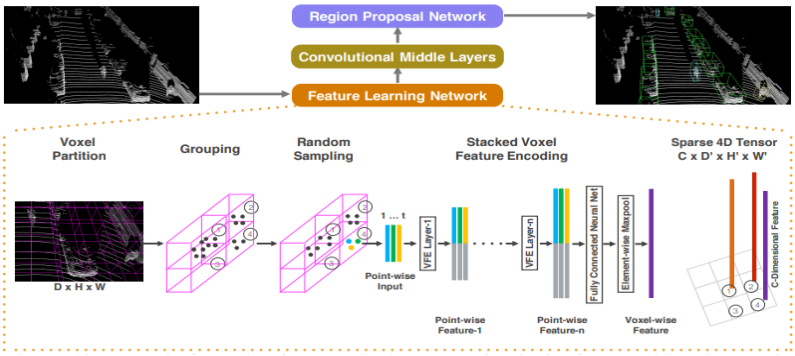
As for grouping, once the representative points are sampled, k-nearest neighbor algorithm chooses the nearest neighbor points to the representative points in order to form a plot. After sampling and grouping, the last action consists of mapping into a feature vector that will be the sought-for structure.

Unlike PointNet, there are numerous methods that consider learning features for each point in literature. Such is for instance PointNte++ [27] which completed PointNet by implementing it in local regions. Other methods are: VoxelNet, Self Organizing Map, Pointwise Convolution (Figure 2).

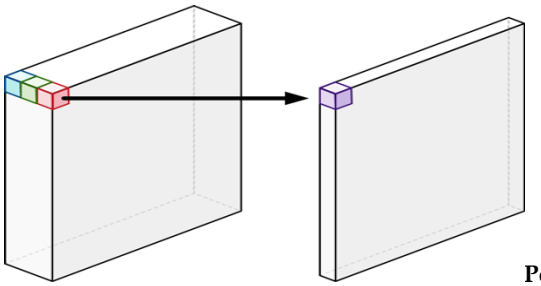
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1



SOMap



VoxelNet



Pointwise Convolution

Figure 2. Illustration of other methods for 3D point clouds features learning among which SOMap, VoxelNet and Pointwise Convolution.

VoxelNet [58] uses a parameter called Voxel Feature Encoding (VFE). A point cloud is initially shed into 3D voxels and then points are stored into the voxel they find themselves in. Self Organizing map (SOM) [59] does generate a self organizing network in SONet [60] where either techniques of point sampling (uniform, random and farthest) does choose centroids before further processing. In the Pointwise convolution [61] technique, no points are subsampled since the whole set of points are used in the convolution operation. Instead, in each point, nearest neighbor points are sampled according to the radius and size of a kernel.

Moreover, graph based approaches (Figure 3) were proposed that represent the point cloud using a graph structure where all points are considered to be a node. In a graph structure, the relationships between points are represented using edges and nodes.

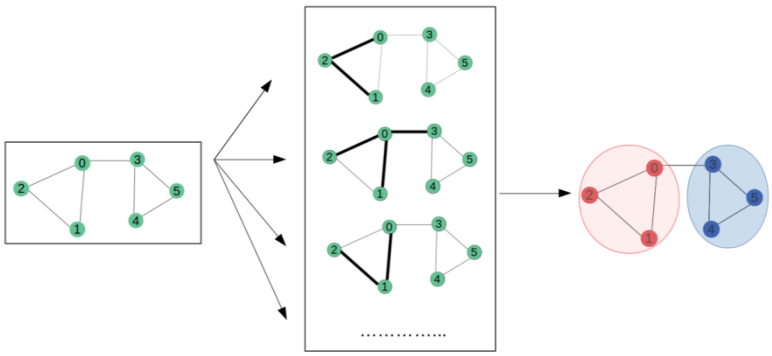


Figure 3. Graph based convolutional networks

Actually, Esteves et al. [62] use a different type of graph called kd-tree. The kd-tree is designed in a top-down manner to generate a feed-forward with learnable parameters in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

every layer. Thomas et al. [63] propose a Dynamic Graph CNN (DGCNN) method with a clear disparity from the standard graph network where the computed features from the prior layer are the basis of an edge updating mechanism which takes place after edgeConv layer.

2.2. 3D point clouds methods for instance segmentation

Generally, 3D instance segmentation techniques will be divided into 2 main categories (Table 2) such as region proposal based methods centered on the generation of a bounding box and region proposal free methods focusing on deep learning frameworks. These categories are the focus of this paper which will analyze, compare and study them.

Table 2. Nomenclature of deep learning methods for 3D instance segmentation

Categories	Representative methods
Region proposal methods	3D-BoNet [64], GSPN (R-PointNet) [65], PanopticFusion [66], LIDARSeg [67], 3DSIS [68], GICN [69]
Region proposal free methods	SGPN [70], MASC [71], Discriminative embeddings [72], MTML [73], Dynamic Region Growing [74], ClusterNet [75], PointGroup [76], 3D-BEVIS [77], 3D MPA [78], MT-PNet [79], ASIS [80], JSNet [81]

2.2.1. Region proposal based methods

This approach is generally divided into two main steps which are the objects proposal (bounding box proposal, object proposal) and the refinement task (mask prediction…). Yang et al. [64] followed this two-step solution and proposed a structure for 3D instance segmentation known as 3D-BoNet where the system simultaneously regress 3D bounding boxes and predicts point-level mask for all objects in the scene without the need of post-processing tasks such as clustering, feature sampling, non-maximum suppression or voting. However, 3D-BoNet also has some restrictions such as an inability to learn the weights of the diverse input point clouds, the absence of advanced feature fusion components to improve both semantic and instance segmentation at the same time.

Yi et al. set p the Generative Shape Proposal Network (GSPN) [65] as an algorithm for point cloud instance segmentation that leverages an innovative structure named Region-based PointNet (R-PoinNet) to allow both adjustable proposal enhancement and instance segmentation generation. A Point RoIAlign layer is introduced to accumulate features and to allow the network process the proposals. For the most part, the triumph of GSPN emerges from its reliance on geometric shape in object proposal with the advantage of removing proposals with low objectness. With PanopticFusion [66], Narita et al. used a voxel based semantic technique at the level of stuff and things that tightly predicts object classes from a background area and independently section random forefront objects to recreate a major scene using a volumetric map depiction. However, the network lacks global stability against lasting pose drift to incur major throughput mapping w.r.t. to dynamic environments. Zhang et al. [67] introduced LIDARSeg which is an outdoor LiDAR point clouds based network for 3D instance segmentation for the segmentation and localization of little objects while it also remains an adequate answer to single-shot instance prediction and to severe class imbalances. Basically, this method learns a feature representation by using self-attention blocks in point clouds.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

Hou et al. [68] established a neural network structure called 3D-SIS for instance semantic segmentation of RGB-D scans. Here, the major contribution is the ability to simultaneously learn from both geometric and RGB input, hence allowing precise instance estimations. The network builds a bounding box regression followed by instance (mask) segmentation for all objects. The network is fully convolutional meaning it can run well in a single shot for sizeable 3D environments. Liu et al. [69] presents a novel method called Gaussian Instance Center Network (GICN), which approximates the distributions of instance centers spread in the entire scene as Gaussian center heatmaps. Centers instance size prediction, bounding boxes generation and final instance masks are obtained as a result of the predicted heatmaps. Possible improvements would consist of enhancing accuracy in finding centers for the challenging semantic classes and using metric learning to learn feature embeddings.

2.2.1. Region proposal free methods

Proposal-free semantic methods group points into instances using a system of similarity metrics while methods relying on segmentation predict the semantic labels to gather points into object instances. These techniques learn 3D instance segmentation using deep learning framework. Wang et al. [70] brought forth a Similarity Group Proposal Network (SGPN) which learns a similarity matrix to articulate group proposals of object instances. The structure first uses PointNet and PointNet++ in order to pull out significant characteristics in every point of the specific point cloud. Here points from he identical objects are considered to carry similar features to create a similarity matrix. The problem is that the size of the similarity matrix increases quadratically making it impossible to process very large scenes due to memory constraints. Meanwhile, Liu and Furukawa [71] presented MASC, a novel based straightforward and efficient process to learn the ressemblance between points to group them into instances. The network adopts sparse convolution and propose a clustering algorithm based on learned multi-scale affinities to tackle the 3D instance segmentation problem. But the network is slow since the clustering algorithm is implemented sequentially even if it is parallel in theory. Liang et al. [72] suggested a 3D CNN called sub-manifold sparse convolutional network which simultaneously produces semantic predictions and instance embeddings. Here a loss function takes into account both the embedding and structure data where a graph convolutional neural network uses an attention-based k-nearest neighbor (KNN) to allocate diverse weights to various neighbors. Lahoud et al. [73] suggested a multi-task learning problem (MTML) which is a 3D instance segmentation process that centers on volumetric scenes from either multi-view stereo or depth sensors. In MTML, voxel-based scenes use metric learning to approximate the scene object centers directional information to achieve the segmentation results.

Hu et al. [74] introduced a method called 3D-SIS where segmentation is coupled with labeling. In fact, the method first segment the point clouds into surface patches and clusters them into group patches using unsupervised learning. The method thus avoids parameter sensitivity by producing both semantic labeling results and robust patch segmentation. Though the method manages occlusions very well, its performance is affected in extremely occluded scenes which could be resolved by generalizing the patch segmentation.

Shao et al. [75] proposed ClusterNet model for 3D instance segmentation of objects in RGB-D images. The task is presented as a regression problem followed by clustering. Specifically, the model makes pixel-wise predictions followed by sequential clustering in the feature space to deduce the object instances. To overcome shortcomings, an improved version would be to assimilate that approach with semantic segmentation. Jiang et al. [76] have introduced a innovative end-to-end bottom-up structure called Point Group with three key components (clustering, backbone network and ScoreNet) that specially centers on an improved grouping by investigating the empty space among objects. The network has a two-branch architecture (point features extraction and semantic labels prediction)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

and actually pushes every point to its own object centroid. However, this model requires a refinement section to alleviate the semantic inaccuracy predicament that affects the instance assemblage and should investigate the prospect of supervision techniques to enhance its performance.

To implement instance segmentation over complete 3D scans, Elich et al. [77] brought up 3D-BEVIS which is a 2D-3D bird's eye view framework which learns global consistent instances features from a u-shaped fully convolution network. The point clouds features are mutually predicted by a graph neural network. However the technique is unable to show very well numerous occluded objects in its 2D representations. A possible future solution could be to consider different 2D representations to rise above the bird's-eye view restrictions. Engelmann et al. [78] propose 3D-MPA an object-centric 3D instance segmentation technique to create instance proposals based on a graph convolutional network to enable better interactions among neighboring proposals. The absolute object instances will be reached when putting together several proposals rather than pruning them with non-maximum-suppression. Even though multi-proposal aggregation yields good promise for object detection, there still remain challenges such as the combination possibilities between tracking and detection in sequences that are semi-dynamic for instance.

On the other side, Pham et al. [79] present the Multi-Task Point-wise Network (MT-PNet) that concurrently addresses the challenges of instance and semantic segmentation in3D point clouds. It develops a multi-task point wise network (semantic classes prediction and same object instance points embedding similar) followed by a conditional random field model to integrate the semantic and objects instance labels. Wang et al. [80] introduced a method called Associatively Segmenting Instances and Semantics (ASIS) which makes semantic segmentation benefit its instance level counterpart through learning an instance embedding from a semantically aware point cloud. Semantic features belonging to the one instance are combined to correctly improve semantic predictions. This process has a possibility to be extended to panoptic segmentation and beyond.

Moreover, Zhao and Tao [81] proposed JSNet which deals with both the 3D point clouds semantic and instance segmentation at the same time which contains an efficient backbone network (vigorous features extraction) and also a point cloud feature fusion module (discriminative features detection).

3. Benchmarks datasets for 3D instance segmentation

3.1.Outdoors 3D datasets

These datasets are those that are built from outdoor scenes. They are manufactured using various tools such as Lidar, lasers scanners, cameras…

The most used outdoors datasets used for 3D segmentation and classifications are:

- ASL [82] (ETH Zurich) made from structured and unstructured environments using Hokuyo UTM-30LX. It contains 8 point cloud sequences.
- Apollo [83][84] (Baidu Research) gathered from 3D car, object detection and localization and contains Several thousands of images of car instances.
- BLVD [85] (Xian Jiaotong University-China) stemming from 5D event recognition, 4 D object tracking and 5D intention prediction and contains 654 video clips with 120k frames (frame rate is 10fps/sec) and 3D annotations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

- DBNet [86] (Xiamen University) from LiDAR, video record, GPS and drivers conduct monitoring using Velodyne laser and boasts 1000 kms of driving.
- iQmulus [87] (Mines ParisTech) made using Riegl LMS-Q120i and contains 50 classes of 300 million points.
- KITTI [88] (Karlsruhe Institute of Technology) for optical flow estimation, stereo image, 3D detection, 3d tracking… using GPS/IMU, cameras, Velodyne HDL-64E 3D laser scanner and has 93000 depth maps of 5 groups.
- NCLT [89] (University of Michigan) for LiDAR point cloud, 2D images, INS and GPS ground truth using Velodyne-32 LiDAR scanner and contains several trajectories inside the University of Michigan.
- NPM3D [90] (PSL Research University) for classification and segmentation using a Velodyne HDL-32E and GPS/INS to collect 1,431Mpoints data.
- NuScenes [91] (nuTonomy) for classification, velocity, size, localization using 6 cameras, 1 lidar and 5 radars and contains 8 attributes and 23 classes labeled with 3D bounding boxes.
- Oxford Robotcar [92] (University of Oxford) focusing on LiDAR point cloud, 2D images, vehicles INS/GPS ground truth collected by 1 SICK LD-MRS 3D LIDAR and 2 SICK LMS-151 2D LiDAR and contains various difficult environment such as season, traffic, weather …
- Semantic3D [93] (ETH Zurich) for semantic segmentation collected through static terrestrial laser scanners and contains 4 billion points of terrain, difficult vegetation…

3.2. Indoor 3D datasets

Contrary to the outdoors datasets, these ones are those built from indoor scenes. They are also manufactures using tools such as Kinect, Sturcture from motion, laser scanners, cameras…

The most used indoor datasets used for 3D segmentation and classifications are as follows:

- 3DMatch [94] (Princeton University) with 62 scenes.
- Matterport3D [95] (Princeton University) focusing on camera poses and semantic segmentation and has over 10 000 panoramic scenes of hundred large- buildings.
- Multisensor Indoor Mapping and Positioning Dataset [96] (Xiamen University) for dense laser scanning of indoor scenes from multiples sensors.
- NYUDv2 [97] (New York University) for segmentation labels collected using the Kinectv1 and has a thousand and a half RGB-D images from 464 various indoor scenes.
- S3DIS [98] (Stanford University) for semantic and instance segmentation labels and contains 13 groups of 3 different buildings with 271 rooms.
- ScanNet [99] (Stanford University) for Object classification and semantic segmentation collected an occipital structure sensor and has thousands of scanned scenes of indoor environments.
- SceneNN [100] (Singapore University of Technology and Design) for semantic labels collected with a Kinectv2 and contains 101 indoor scenes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

- SUN3D [101] (Princeton University) for camera pose and object labels collected using structure from motion (SfM) and is made of videos from 254 different spaces in 41 buildings.

4. Comparative study and analysis

4.1. Runtime analysis

Computation speed is a very important metric for most systems which meet rigid requirements regarding the amount of time spent on execution, training… However, it is helpful to provide a system speed with a comprehensive description of the hardware used to execute it (Table 3) that could help other researchers, Despite the various hardware and the different training methods used, the table clearly shows that the 3D-SIS network outperforms all the other methods regardless of categories with only about 10 epochs needed for training.

Table 3. Various methods showing according to their categories, the optimization used, the graphical processing unit (GPU) and their training time.

Categories	Methods	GPU used	Training time
Region Proposal based	3D SIS[68]	Single Nvidia GTX 1080 Ti	10 epochs
	PanopticFusion[66]	Intel Core i7-7800X	30 epochs
	3D-BoNet[64]	TitanX	20 epochs
	GICN[69]	Two Nvidia GTX1080Ti	50 epochs
Region Proposal free	3D-MPA[78]	Nvidia Titan Xp (12 Gb)	-
	PoitnGroup[76]	Titan Xp Nvidia Titan Xp	-
	SGPN (R-PointNet)[70]	Titan 1080 Ti	-
	MASC[71]		20 epochs
	MT-PNet[79]	Single Nvidia Titan X	50 epochs

4.2. Accuracy

To evaluate the accuracy in computer vision, the overall mean average precision is an important factor. The mean Average Precision (mAP) is the most used because it is a measure that combines reacll and precision. Table 4 shows the comparison of the methods performance on the ScanNet (v2) dataset while table 5 is didplaying the accuracy comparison on the 3DSIS Dataset for area 5 and 6.

Using ScanNet (v2) as dataset, Table 4 illustrates that no method is simultaneously the complete solution for both categories of methods and metrics. A complete solution should outperform all other methods both in 25 and 50% mAP at th same time. So @25% 3D MPA performs better while @50% GICN (superseding Pointgroup slightly) performs well.

Using the 3DSIS 6-fold CV dataset, Table 5 on the contrary shows 3D MPA strength on both metrics by performing better than all other methods in making it the best solution on all methods for the 3D-SIS dataset. In conclusion, it appears that the region proposal free methods achieve better accuracy in general on both ScanNet (v2) and 3DSIS datasets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

Table 4. ScanNet (v2) Validation set mean Average Precision (mAP) of the various methods according to their categories

Categories	Methods	mAP@50%	mAP@25%
Region Proposal based	GSPN[65]	19.3	53.4
	3D-BoNet[64]	48.8	-
	PanopticFusion[66]	47.8	-
	3DSIS[68]	38.2	35.7
	GICN[69]	63.8	-
Region Proposal free	3D MPA[78]	30.6	72.4
	3D BEVIS[77]	24.8	-
	SGPN (R-PointNet)[70]	35.3	22.2
	MASC[71]	44.7	-
	MTML[73]	55.0	-
	PointGroup[76]	63.6	-

Table 5. 3DSIS 6-fold CV mean Average Precision (mAP) of the various methods according to their categories

Categories	Methods	mAP@50%	mAP@25%
Region Proposal based	3D-BoNet[64]	65.6	47.6
Region Proposal free	3D MPA[78]	66.7	64.1
	PointGroup[76]	64.0	-
	SGPN(R-PointNet)[70]	54.4	-
	ASIS[80]	63.6	47.5
	JSNet[81]	54.4	53.9
	MT-PNet[79]	24.9	-

4.3. Robustness to noise

The presence of noise in the 3D LIDAR point clouds in the proposal based methods will result in the loss of points in the expected results. But in terms of voting for classification or segmentation a strategy of "one point one vote" would help to estimate the categories of each object as well as minimize the impact of the extreme values of unanticipated noises. By showing where to sample within the object distribution, the scene observations help in directing the proposal generation process. For instance, GSPN [65] encodes noisy observations in the object space to produce an instance based feature extraction process.

In the proposal free methods, the idea to generate object proposals from directional information alone is tiresome because of the noise while the clustering predicament is harder and less efficient. Therefore, mean shift clustering for objects proposals and directional information are the prevalent methods. Since real data contains missing data, noise and outliers, the projection of each points in the feature space could be less discriminative and clusters could overlap.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

4.4. Performance on scenes variations

Proposal based methods such as GICN [69] are capable to approximate the center distributions spread in the entire scene as Gaussian center heatmaps correctly even for unknown 3D scenes. Furthermore, PanopticFusion [66] implements total scene reconstruction and solid semantic labelling with the capacity to distinguish each object and easily realizes context-aware interactions while naturally visualizing the occlusion effects. SGPN [70] is weak for road scenes with very scarece points but LIDARSeg [67] is able to recognize smaller objects and not have several false positives.

For the proposal free methods, there exist fundamental limitations such as for scenes where various occluded objects are not well visible. As an example, 3D-BEVIS [77] is able to learn global instance features which are consistent over a full scene.

In rare situations, the results of MTML [73] turn out to be worse due to unproperly scanned scenes parts while ClusterNet [38] model does force RGB and depth data to obtain robust 3D instance segmentation in cluttered scenes with much occlusions. SGPN [70] is unable to process really large scenes in the order 10 plus ponts.

4.5. Loss functions study for performance improvement

In 3D-BEVIS [77] approach for sizeable distances between similar pixels the network uses a discriminative loss as the cross-entropy loss where the semantic losses are gauged with the negative logarithm of the class frequency.

In GICN [69], the proposed network is trained and optimized by a joint loss L_{total} made of numerous loss expressions such that $L_{total} = L_{center} + L_{bound} + L_{IoU} + L_{mask} + L_{size}$. The GICN [69] bounding box loss L_{bound} is simple when compared with the multi-criteria loss used for box prediction by 3D-BoNet [64] because it requires a box association layer to choose the mapping between the ground-truth bounding boxes and the predicted ones.

In a case of a lack of focal loss, GICN helps the network resolve the imbalance predicament of predicting diverse instances. The difference in using a vanilla cross entropy loss and the focal loss is wide. An improvement of 14% on mAP can be observed. This result shows that without the focal loss strategy th GICN network tends to learn merely the simpler instances.

In 3D-SIS [68],the detection algorithm brings out features that serve as input to the 3D Region Proposal Network (3D-RPN) and the 3D Region of Interest to predict both the locations of the bounding box and the labels of the object class. The 3D-RPN training is by merging ground-truth object annotations with previously designed anchors. In addition, the 3D-SIS [68] uses a two-set cross entropy loss for the objectiveness measurement. Finally, a Huber loss is used for the bounding box regression.

In order to constrain the direction of the predicted offset, the PointGroup [76] network formulates a direction loss because it finds it difficult to regress precise offsets most importantly for boundary points of sizeable objects. The binary cross-entropy loss is used as the 3D-SIS [68] score loss to allow the whole framework to train in an end-to-end approach where the total loss is an addition of the all the losses (semantic, direction, regression, score).

In order to learn the feature embeddings (where one maps each voxel to a feature space, the other allocates a 3D vector to each voxel), MTML [73] introduces a multi-assignment loss function that is reduced in training. The first component of the loss promotes discrimination among various instances in the feature space, while the second component castigates vector angular deviations from the selected direction. For the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2021 doi:10.20944/preprints202111.0228.v1

directional loss, MTML[73] aims to generate a vector feature that would locally describe the intra-cluster relationship without being affected by other clusters. We choose the vector to be the one pointing towards the ground truth center of the object. To learn this vector feature, we attend to the following directional loss.

Generally, the MTML [73] network trained with a multitask loss always performs better than the single-assignment one. This is confirmer by the results on the synthetic dataset and further supports the premise that the directional loss appends more discriminative features.

In Clusternet [75], the loss function stimulates the network to project a point in feature space to allow that points of equivalent instances would be close to each other while points of different instances would be separated by a wide margin. It is defined as an object-centric training loss consisting of the sum of the semantic mask loss (cross-entropy between the ground truth and estimated semantic segmentation), the cluster center loss, the pixel-wise loss, the variance loss and finally the violation loss.

In ASIS [80], the classical cross-entropy loss supervises the semantic segmentation step at training time. The loss function is formulated as an addition of the which L_{var} aims to drag embeddings to the average embedding of the instance, i.e the instance center, the L_{dist} allows instances to fend each other off and the L_{reg} is a regilarization term to bind the embedding values. In 3D-MPA [78], the mask loss is introduced as a focal loss instead of a cross-entropy losss in order to manage the class imbalance in the foreground and background while the model is trained with the multi-task loss in an end-to-end manner.

Basically, in 3D-BoNet [64], the loss function has two goals which are to minimize the Euclidean distance between the associated predicted box and each ground truth box, but also to maximize the coverage of suitable points inside of every predicted box. For the loss function, 3D-BoNet uses the focal loss with with default hyper-parameters in the place of the standard cross-entropy loss for optimization.

GSPN [65] is trained to reduce a multi-task loss function L_{GSPN} defined for every prospective object proposal. In fact, L_{GSPN} is an addition of five expressions such as the shape generation loss $L_{gen,}$ the shape geneation per-point confidence loss $\mathbf{L_e}$, the KL loss LKL, the center prediction loss L_{center} and the objectness loss L_{obj}. The loss of the MTPNet [79] is the summation of the branches of the losses where $L = L_{prediction} + L_{embedding.}$ The prediction loss $L_{prediction}$ is then by the cross-entropy. MTPNet emplys a discriminative loss to introduce the embedding loss $L_{embedding}$. In JSNet [81], the loss function L at trainig time consists of a summation of the semantic segmentation loss L_{sem} and the instance embedding loss L_{ins} where L_{sem} is defined with the classical cross-entropy loss.

4.6. Metric as a learning tool

2D instance segmentation methods use metric learning as one of the very important tool. A frequent metric-learning approach is to identify a pair-wise loss for learning appropriate pixel embeddings so that points from the same instance are closer together. Comparable approaches can be implemented in the learning of embeddings for 3D instance segmentation. For example, MTML [73] learns intra-instance and inter-instance relationships, and develops a post-processing task for semantic segmentation with a mean-shift clustering to assemble the 3D points. The labels are defined by learning a metric which assembles parts of one object instance and predicts the direction of the object's center of mass. ASIS [80] and JSIS3D [79] also choose mean-shift clustering to deduce instance segmentation clusters. One more strategy for metric learning consists of training a network to estimate the instance affinity sore if it predicts whether two points are from one object instance such as in MASC [71].

After outputting initial embeddings for all points, the discriminative embeddings [72] method argues that ponts belonging to one instance will present comparable embeddings while ponts of different instances would be far in the embedding space. This is a normal predicament in the metric learning. Nonetheless, for the 3D instance segmentation task, points of the same instance will have both embedding features and geometric associations in the 3D space. The network combines such embedding features with structure information to attain more discriminative final results. So the network needs a metric to measure the likeness among embeddings and argues that the KNN uses the spatial distance of points rather than the embedding distance as metric.

In 3D-MPA [78], bottom-up approaches use metric-learning techniques to learn a per-point feature embedding space that is then grouped into object instances. This approach is able to successffully deal with outliers, but it deeply relies on altering cluster parameters. The SGPN [70] method is also intimately connected to similarity metric learning, which has been extensively used in deep learning on a multitude of tasks such as person re-identification [102], matching [103], image retrieval [104], and face recogition [105]. Basically, the SGPN method uses metric learning in a different way in that it regress the probability of two points belonging to the one group and formulates the similarty matrix as group proposals to deal with variety of a number of instances.

5. Possible future research directions

Based on this review which studies and compares the most prominent methods in the 3D instace segmentation field, possible future research would address the increasing need for large datasets, the problem of an efficient segmentation of voluminous point clouds, the semantic inaccuracy problem by exploring the possiblity of incorporating auto-supervision techniques that could further enhance the system achievement as wellas the ability to learn the dynamic point clouds spatio-temporal data which could enhance the output of ensuing tasks such as 3D object recognition and segmentation.

Many others avenues exist such as new applications turned towards context-aware augmented reality and intelligent autonomous robotics or combination of detections using tracking in partially dynamic situations but also the possibility to separate object proposals in 4D space.

6. Conclusions

This paper has established a comparative analysis of the releveant techniques for 3D perception specifically 3D instance segmentation including 3D region proposal based methods and 3D region free proposal methods. All datasets normally used in 3D point clouds instance segmentation have been presented. A complete nomenclature, performance evaluation as well as the respective contributions of these 3D instance segmentation methods was discussed. Prospective research possibilities were discussed. This research is a unique review paper in the literature which compares and discusses the various deep learning based instance segmentation techniques.

Funding: This research review received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Garcia-Garcia, A.; Gracia-Rodriguez, J. ; Oprea, S.; Orts-Escolano, S. ; Villena-Martinez ,V.; A Review on Deep Learning Techniques Applied to Semantic Segmentation. arXiv: 1704.06857v1 [cs.CV], 22 Apr 2017.
2. Liang, Z.; Feng, Y.; Guo, Y.; Chen, W.; Liu, H.; Qiao, L.; Zhang, J.; Zhou, L. Stereo matching using multi-level cost volume and multi-scale feature constancy. In TPAMI, 2019.

3. Guo, Y.; Bennamoun, M.; Lu, M.; Sohel, F.; Wan, J. Rotational projection statistics for 3D local surface description and object recognition. In IJCV, 2013, vol. 105, no. 1, pp. 63–86.
4. Guo, Y.; Bennamoun, M.; Lu, M.; Sohel, F.; Wan, J. 3D object recognition in cluttered scenes with local surface features: a survey. In TPAMI, 2014, vol. 36, no. 11, pp. 2270–2287.
5. Chen, X.; Li, B.; Ma, H.; Wan, J.; Xia, T. Multi-view 3D object detection network for autonomous driving. In CVPR, 2017, pp. 1907–1915.
6. Nguyen, M. T.; Bao, P. T.; Dang, T. V.; Thi, M. K. Generating Point Cloud from measurement and shapes on Convolutional Neural network: An Application for building 3D human model. In CIN, 2019.
7. Mineo, C.; Pierce, S. G.; Summan, R. Novel algorithm for 3D surface point cloud boundary detection and edge reconstruction. doi.org: 10.016/j.jcde.2018.02.001.
8. Griffits, D.; Boehm, J. A Review on Deep Learning Techniques for 3D Sensed Data Classification. In Remote Sensing, June 2019, 11, 1499.
9. Liu, W.; Hu, T.; Li, W.; Sun, J.;Wang, P. Deep Learning on Point Clouds and Its Application: A Survey. In Sensors, September 2019, doi: 10.3390/s19194188.
10. Guo, Y.; Hu, Q.; Liu, H.; Liu, L.; Wang, H.; Bennamoun,V. Deep Learning for 3D Point Clouds: A Survey. In IEEE TPAMI, 2020.
11. Qi, C. R., Mo, K.; Su H.; Guibas L. J. PointNet: Deep learning on point sets for 3D classification and segmentation. In CVPR, 2017.
12. Catalucci, S.; Senin, N. State-of-the-art in point cloud analysis, Advances in Optical Form and Coordinate Metrology, Chap 2, State-of-the-art in point cloud analysis, December 2020, Pages 2-1 to 2-48.
13. Lawin, F. J. ; Danelljan, M.; Tosteberg, P.; Bhat, G.; Khan, F. S.; Felsberg, M. Deep projective 3D semantic segmentation. In Proceedings of International Conference on Computer Analysis of Images and Patterns, 2017, pp. 95-107.
14. Boulch, A. ; Le Saux, B. ; Audebert, N. Unstructured point cloud semantic labeling using deep segmentation networks. In Proceedings of the Eurographics Workshop on 3D Object Retrieval, 2017.
15. Tatarchenko, M.; Park, J.; Koltun, V.; Zhou, Q.-Y. Tangent convolutions for dense prediction in 3D. In CVPR, 2018, pp. 3887–3896.
16. Wu, B.; Wan, A.; Yue, X.; Keutzer, K. SqueezeSeg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3D lidar point cloud. In ICRA, 2018, pp. 1887–1893.
17. Wu, B. ; Zhou, X.; Zhao, S. ; Yue, X.; Keutzer, K. SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In ICRA, 2019, pp. 4376–4382.
18. Milioto, A.; Vizzo, I.; Behley, J.; Stachniss, C. Rangenet++: Fastand accurate lidar semantic segmentation. In IROS, 2019.
19. Tchapmi, L.; Choy, C.; Armeni, I.; Gwak, J.; Savarese, S. SEGCloud: Semantic segmentation of 3D point clouds. In 3DV, 2017, pp. 537–547.
20. Graham, B.; Engelcke, M.; van der Maaten, L. 3D semantic segmentation with submanifold sparse convolutional networks. In CVPR, 2018, pp. 9224–9232.
21. Choy, C.; Gwak, J.; Savarese, S. 4D spatio-temporal convnets: Minkowski convolutional neural networks. arXiv preprint arXiv:1904.08755, 2019.
22. Meng, H.-Y.; Gao, L.; Lai, Y.; Manocha, D. VV-Net: Voxel vae net with group convolutions for point cloud segmentation. arXiv preprint arXiv:1811.04337, 2018.
23. Su, H.; Jampani, V.; Sun, D.; Maji, S.; Kalogerakis, E.; Yang, M.-H. ; Kautz, J. Splatnet: Sparse lattice networks for point cloud processing. In CVPR, 2018, pp. 2530–2539.
24. Rosu, R. A.; Sch ¨ utt, P.; Quenzel, J.; Behnke, S. Latticenet: Fast point cloud segmentation using permutohedral lattices. arXiv preprint arXiv:1912.05905, 2019.
25. Chiang, H.; Lin, Y.; Liu, Y.; Hsu, W. H. A unified point-based framework for 3D segmentation. arXiv preprint arXiv:1908.00478, 2019.
26. Jaritz, M.; Gu, J.; Su, H. Multi-view pointnet for 3D scene understanding. In ICCVW, 2019.
27. Qi, C. R.; Yi, L.; Su, H.; Guibas, L. J. PointNet++: Deephierarchical feature learning on point sets in a metric space. In NeurIPS, 2017, pp. 5099–5108.
28. Jiang, M.; Wu, J.; Lu, C. PointSIFT : A sift-like network module for 3D point cloud semantic segmentation. arXiv preprint arXiv:1807.00652, 2018.
29. Engelmann F.; Kontogianni, T.; Schult, J.; Leibe, B. Know what your neighbors do: 3D semantic segmentation of point clouds. In ECCV, 2018, pp. 0-0.
30. Zeng, W.; Gevers, T. 3DContextNet: K-d tree guided hierarchical learning of point clouds using local and global contextual cues. In ECCV, 2018.
31. Xie, S.; Liu, S.; Chen, Z.; Tu, Z. Attentional shape-contextnet for point cloud recognition. In CVPR, 2018, pp. 4606–4615.
32. Zhao, H.; Jiang, L.; Fu, C.-W.; Jia, J. PointWeb: Enhancing local neighborhood features for point cloud processing. In CVPR, 2019, pp. 5565–5573.
33. Yang, J.; Zhang, Q.; Ni, B.; Li, L.; Liu, J.; Zhou, M.; Tian, Q. Modeling point clouds with self-attention and gumbel subset sampling. arXiv preprint arXiv:1904.03375, 2019.
34. Chen, L.-Z.; Li, X.-Y.; Fan, D.-P.; Cheng, M.-M.; Wang, K.; Lu, S.-P. LSANet: Feature learning on point sets by local spatial attention. arXiv preprint arXiv:1905.05442, 2019.
35. Zhang, Z.; Hua, B.-S.; Yeung, S.-K. ShellNet: Efficient point cloud convolutional neural networks using concentric shells statistics. arXiv preprint arXiv:1908.06295, 2019.

36. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. arXiv preprint arXiv:1911.11236, 2019.
37. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. PointCNN: Convolution on x-transformed points. In NeurIPS, 2018, pp. 820–830.
38. Wang, S.; Suo, S.; Ma, W.-C.; Pokrovsky, A.; Urtasun, R. Deep parametric continuous convolutional neural networks. In CVPR, 2018, pp. 2589–2597.
39. Komarichev, A.; Zhong, Z.; Hua, J. A-CNN: Annularly convolutional neural networks on point clouds. In CVPR, 2019, pp.7421–7430.
40. Boulch, A. ConvPoint: continuous convolutions for point cloudsprocessing. arXiv preprint:1904.02375, 2019.
41. Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; Guibas, L. J. KPConv: Flexible and deformable convolution for point clouds. arXiv preprint arXiv:1904.08889, 2019.
42. Engelmann, F.; Kontogianni, T.; Leibe, B. Dilated point convolutions: On the receptive field of point convolutions. arXiv preprint arXiv:1907.12046, 2019.
43. Mao, J.; Wang, X; Li, H. Interpolated convolutional networks for 3d point cloud understanding. In ICCV, 2019, pp. 1578–1587.
44. Huang, Q.; Wang, W.; Neumann, U. Recurrent slice networks for 3D segmentation of point clouds. In CVPR, 2018, pp. 2626–2635.
45. Engelmann, F.; Kontogianni, T.; Hermans, A.; Leibe, B. Exploring spatial context for 3D semantic segmentation of point clouds. In ICCV, 2017, pp. 716–724.
46. Ye, X.; Li, J.; Huang, H. ; Du, L.; Zhang, X. 3D recurrent neural networks with context fusion for point cloud semantic segmentation. In ECCV, 2018, pp. 403–417.
47. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M.; Solomon, J. M. Dynamic graph CNN for learning on point clouds. ACM TOG, 2019.
48. Landrieu, L.; Simonovsky, M. Large-scale point cloud semantic segmentation with superpoint graphs. In CVPR, 2018, pp. 4558–4567.
49. Landrieu, L.; Boussaha, M. Point cloud over-segmentation with graph-structured deep metric learning. arXiv preprint arXiv:1904.02113, 2019.
50. Wang, L.; Huang, Y.; Hou, Y.; Zhang, S.; Shan, J. Graph attention convolution for point cloud semantic segmentation. In CVPR, 2019, pp. 10 296–10 305.
51. Pan, L.; Chew, C.-M.; Lee, G. H. Pointatrousgraph: Deep hierarchical encoder-decoder with atrous convolution for point clouds. arXiv preprint arXiv:1907.09798, 2019.
52. Liang, Z.; Yang, M.; Deng, L.; Wang, C.; Wang, B. Hierarchical depth-wise graph convolutional neural network for 3d semantic segmentation of point clouds. In ICRA. IEEE, 2019, pp. 8152–8158.
53. Jiang, L.; Zhao, H.; Liu, S.; Shen, X.; Fu, C.-W.; Jia, J. Hierarchical point-edge interaction network for point cloud semantic segmentation. In ICCV, 2019, pp. 10 433–10 441.
54. Lei, H.; Akhtar, N.; Mian, A. Spherical convolutional neural network for 3d point clouds. 2018.
55. Liu, J.; Ni, B.; Li, C. ; Yang, J.; Tian, Q. Dynamic points agglomeration for hierarchical point sets learning. In ICCV, 2019, pp. 7546–7555.
56. Bello, S. A.; Yu, S.; Wang, C.; Adam, J.M.; Li, J. Review: Deep Learning on 3D Point Clouds, In Remote Sensing, Volume 12, Issue 11, 2020.
57. Yang, J.; Zhang, Q.; Li, L.; Liu, J.; Ni, B.; Tian, Q.; Zhou, M. Modeling point clouds with self-attention and gumbel subset. arXiv preprint arXiv:1904.03375, 2019.
58. 21Voxelnet.
59. 22 Self Organizing Map
60. 23 Self organizing network
61. 24Pointwise Convolution
62. Esteves, C.; Allen-Blanchette, C.; Makadia, A.; Daniilidi, K. Learning so(3) equivariant representations with spherical CNNs. In ECCV, 2017, pp. 52–68.
63. Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor field networks: Rotation and translation equivariant neural networks for 3D point clouds. arXiv preprint arXiv:1802.08219, 2018.
64. Yang, B.; Wang, J.; Clark, R.; Hu,. Q.; Wang, S.; Markham, A.; Trigoni, N. Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds. arXiv:1906.01140 [cs.CV], 2019.
65. Yi, L.; Sung, M.; Wang, H.; Zhao, W.; Guibas, L. J. GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud. arXiv: 1812.03320v1 [cs.CV], Dec 2018.
66. Kaji, Y.; Narita, G.; Seno, T.; Ishikawa, T. PanopticFusion: Online Volumetric Semantic Mapping at the Level of Stuff and Things. arXiv: 1903.01177v2 [cs.CV], Sep 2019.
67. Zhang, F.; Guan, C.; Fang, J.; Bai, S.; Yang, R.; Torr, P. H.S.; Prisacariu, V. Instance segmentation of Lidar point clouds. In ICRA, 2020.
68. Hou, J.; Dai, A.; Nießner, M. 3D Semantic Instance Segmentation of RGB-D Scans. In CVPR, 2019.
69. Liu, S.; Yu, S.; Wu, S. Learning Gaussian Instance Segmentation in Point Clouds. arXiv:2007.09860v1 [cs.CV] 20 Jul 2020.
70. Wang, W.; Yu, R.; Huang, Q. ; Neumann, U. SGPN: Similarity Group Proposal Network for 3D Point Cloud Instance Segmentation. In CVPR., 2018.

71. Liu, C.; Furukawa, Y. MASC: Multi-scale Affinity with Sparse Convolution for 3D Instance Segmentation. Technical Report. arXiv:1902.04478v1 [cs.CV], 12 Feb 2019.
72. Liang, Z.; Yang, M.; Wang, C. 3D graph embedding learning with a structure-aware loss function for point cloud semantic instance segmentation. arXiv preprint arXiv: 1902.05247, 2019.
73. Lahoud, J.; Ghanem, B.; Pollefeys, M.; Oswald, M.R. 3D Instance Segmentation via Multi-Task Metric Learning. arXiv: 1906.08650v2, June 2019.
74. Dai, A.; Hou, J.; Nießner, M. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In CVPR, pp. 4421–4430, 2018.
75. Shao, L.; Tian, Y.; Bohg, J. ClusterNet: 3D Instance Segmentation in RGB-D Images. arXiv: 1807.08894v2 [cs.RO], Sep 2018.
76. Jiang, L.; Zhao, H.; Shi, S.; Liu, S.; Fu, C.; Jia, J. PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. arXiv:2004.01658v1 [cs.CV], Apr 2020.
77. Elich, C.; Engelmann, F.; Kontogianni, T.; Leibe, B. 3D-BEVIS: Birds-eye-view instance segmentation. arXiv preprint arXiv: 1904.02199, 2019.
78. Engelmann, F.; Bokeloh, M.; Fathi, A.; Leibe, B.; Nießner, M. 3D-MPA: Multi Proposal Aggregation for 3D Semantic Instance Segmentation. arXiv:2003.13867v1 [cs.CV], March 2020.
79. Pham, Q.; Nguyen, D. T.; Hua, B.; Roig, G.; Yeung, S. JSIS3D: Joint semantic-instance segmentation of 3D point clouds with multi-task pointwise networks and multi-value conditional random fields. In CVPR, pp. 8827–8836, 2019.
80. Wang, X.; Liu, S.; Shen, X.; Jia, J. Associatively Segmenting Instances and Semantics in Point Clouds. arXiv: 1902.09852v2 [cs.CV], Feb 2019.
81. Zhao, L.; Tao, W. JSNet: Joint Instance and Semantic Segmentation of 3D Point Clouds. arXiv:1912.09654v1 [cs.CV] , December 2019.
82. Geiger , A.; Lenz, P.; Stiller, C.; Urtasun, R. 3D Semantic Vision meets robotics: The KITTI dataset, IJRR, (2013).
83. Song, X.; Wang, P.; Zhou, D.; Zhu, R.; Guan, C.; Dai, Y.; Su, H.; Li, H.; Yang, R. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In IEEE CCVPR, 2019, pp. 5452– 5462.
84. Lu, W.; Zhou, Y.; Wan, G.; Hou, S.; Song, S. L3-net: Towards learning based lidar localization for autonomous driving. In: Proceedings of the IEEE CCVPR, 2019, pp. 6389–6398.
85. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1 Year, 1000km: The Oxford RobotCar Dataset, IJRR, 36 (2017) 3–15.
86. Carlevaris-Bianco, N.; Ushani, A. K.; Eustice, R. M. University of Michigan north campus long-term vision and lidar dataset, The IJRR 35 (2016) 1023–1035.
87. Br´edif, M.; Vallet, B.; Serna, A.; Marcotegui, B.; Paparoditis, N. Terramobilita/Iqmulus urban point cloud classification benchmark, 2014.
88. Pomerleau, F.; Liu, M.; Colas, F.; Siegwart, R. Challenging datasets for point cloud registration algorithms, The IJRR 31 (2012) 1705–1711.
89. Chen, Y.; Wang, J.; Li, J.; Lu, C.; Luo, Z.; Xue, H.; Wang, C. Lidar-Video driving dataset: Learning driving policies effectively. In CCVPR, 2018, pp. 5870–5878.
90. Roynard, X.; Deschaud, J.-E.; Goulette, F. Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. In the IJRR 37 (2018) 545–557.
91. Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019).
92. Xue, J.; Fang, J.; Li, T.; Zhang, B.; Zhang, P.; Ye, Z.; Dou, J. BLVD: Building a large-scale 5d semantics benchmark for autonomous driving. In ICRA, 2019.
93. Hackel, T.; Savinov, N.; Ladicky, L.; Wegner, J. D.; Schindler, K.; Pollefeys, M. Semantic3d-Net: A new large-scale point cloud classification benchmark. arXiv preprint arXiv:1704.03847 (2017).
94. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In ECCV, Springer, 2012, pp. 746–760.
95. Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; Zhang, Y. Matterport3d: Learning from rgb-d data in indoor environments. In the International Conference on 3D Vision (3DV) (2017).
96. Wang, C.; Hou, S.; Wen, C.; Gong, Z.; Li, Q.; Sun, X.; Li, J. Semantic line framework-based indoor building modeling using backpacked laser scanning point cloud. In ISPRS Journal of photogrammetry and remote sensing 143. 2018. pp:150–166.
97. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor Segmentation and Support Inference from RGBD Images. In ECCV, 2012.
98. Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I. K.; Fischer, M.; Savarese, S. 3D semantic parsing of large scale indoor spaces. In CVPR 2016, USA,, 2016, pp. 1534–1543.
99. Chang, X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; others. Shapenet: An information rich 3d model repository. arXiv preprint arXiv:1512.03012, 2015.
100. Hua, B.-S.; Pham, Q.-H.; Nguyen, D. T.; Tran, M.-K.; Yu, L.-F.; Yeung, S.-K. Scenenn: A scene meshes dataset with annotations. In 3D Vision (3DV), 2016.
101. Xiao, J.; Owens, A.; Torralba, A. Sun3d: A database of big spaces reconstructed using sfm and object labels. In IEEE ICCV, 2013, pp. 1625–1632.
102. Yi, D.; Lei, Z.; Liao, S.; Li, S. Z. Deep metric learning for person re-identification. In ICPR, 2014.
103. Han, X.; Leung, T.; Jia, Y.; Sukthankar, R.; Berg, A. C. Matchnet: Unifying feature and metric learning for patch-based matching. In CVPR, 2015.
104. Frome, A.; Singer, Y.; Sha, F.; Malik, J. Learning globally consistent local distance functions for shape-based image retrieval and classification. In ICCV, 2007.

 doi:10.20944/preprints202111.0228.v1

Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In CVPR, 2005.