electronics

MDPI

Review

Deep-Learning-Based Point Cloud Semantic Segmentation: A Survey

Rui Zhang †, Yichao Wu *,†, Wei Jin and Xiaoman Meng

School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China; zhangrui@ncwu.edu.cn (R.Z.); Z20211090841@stu.ncwu.edu.cn (W.J.); X20201090790@stu.ncwu.edu.cn (X.M.)

* Correspondence: wu546300070@gmail.com

† These authors contributed equally to this work.

Abstract: With the rapid development of sensor technologies and the widespread use of laser scanning equipment, point clouds, as the main data form and an important information carrier for 3D scene analysis and understanding, play an essential role in the realization of national strategic needs, such as traffic scene perception, natural resource management, and forest biomass carbon stock estimation. As an important research direction in 3D computer vision, point cloud semantic segmentation has attracted more and more researchers' attention. In this paper, we systematically outline the main research problems and related research methods in point cloud semantic segmentation and summarize the mainstream public datasets and common performance evaluation metrics. Point cloud semantic segmentation methods are classified into rule-based methods and point-based methods according to the representation of the input data. On this basis, the core ideas of each type of segmentation method are introduced, the representative and innovative algorithms of each type of method are elaborated, and the experimental results on the datasets are compared and analyzed. Finally, some promising research directions and potential tendencies are proposed.

Keywords: deep learning; point cloud semantic segmentation; convolutional neural network; feature representation learning; computer vision

Citation: Zhang, R.; Wu, Y.; Jin, W.; Meng, X. Deep-Learning-Based Point Cloud Semantic Segmentation: A Survey. *Electronics* **2023**, *12*, 3642. https://doi.org/10.3390/electronics12173642

Academic Editor: Luca Mesin

Received: 12 July 2023
Revised: 6 August 2023
Accepted: 14 August 2023
Published: 29 August 2023

1. Introduction

In recent years, with the booming development of a large group of emerging industries, such as smart cities, automotive navigation systems, augmented reality, and environmental assessment, a large amount of research related to 3D scene perception has been motivated. This research invariably requires the processing and analysis of huge amounts of 3D data. How to enhance the understanding of 3D scenes and extract effective high-level features has become an important scientific problem in 3D computer vision.

As a key form and essential information carrier of 3D data, a point cloud is a collection of points representing the information of objects in 3D scenes, which can be used as a digital representation of the real world. Point clouds usually contain coordinates, color, intensity values, and other attributes so that the original geometric structure of the object in 3D scenes can be retained to the maximum extent. As a key step in understanding 3D scenes, point cloud semantic segmentation is a technique that divides the original point cloud into several subsets with different semantic information and classifies each point into specific groups according to the degree of attribute similarity. At present, point cloud semantic segmentation has been widely applied to national strategic needs, such as autonomous driving [1], augmented reality [2], and transmission line inspection [3]. It has important research significance and broad development prospects.

In recent years, deep learning techniques have made breakthroughs in computer vision, and more and more computer vision tasks rely on convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), and

other derived neural network architectures. Due to their excellent feature learning capacity, the deep neural network has achieved remarkable results and occupied a dominant position in point cloud semantic segmentation. Deep-learning-based point cloud semantic segmentation methods can be subdivided into point-based methods and rule-based methods. The latter transforms the original point cloud into regular structures, such as 2D images and voxels, and automatically extracts features through neural networks to achieve the segmentation of different categories of objects in 3D scenes at the semantic level. However, due to the sparse and unstructured characteristics of point clouds, such operations not only increase the computational overhead but also lead to a large loss of key information, seriously affecting the accuracy of the segmentation methods. Therefore, it is crucial and urgent to explore how to further improve the performance of point cloud segmentation methods while ensuring that the original information is as complete as possible.

There have been some review papers on point cloud semantic segmentation [4–8], but a systematic summary analysis of the latest proposed segmentation methods and datasets is still needed. This paper aims to provide researchers with a comprehensive and systematic understanding of the current state of research in the field of point cloud segmentation by summarizing and analyzing the representative methods proposed from 2015 to 2023. As shown in Figure 1, this paper focuses on point cloud semantic segmentation, introducing and discussing the latest research progress in detail through the following seven sections. First, we analyze the characteristics of point clouds, and to address the challenges they pose, we classify point cloud semantic segmentation into rule-based segmentation and point-based segmentation according to the processing of methods. The representative and innovative implementations of each type of method are elaborated in detail. Furthermore, we introduce mainstream evaluation metrics in the field of point cloud semantic segmentation, summarize more than 20 datasets, and compare the performance results of different methods on the datasets. Finally, the future development trends and research focus of point cloud semantic segmentation are predicted and foreseen.

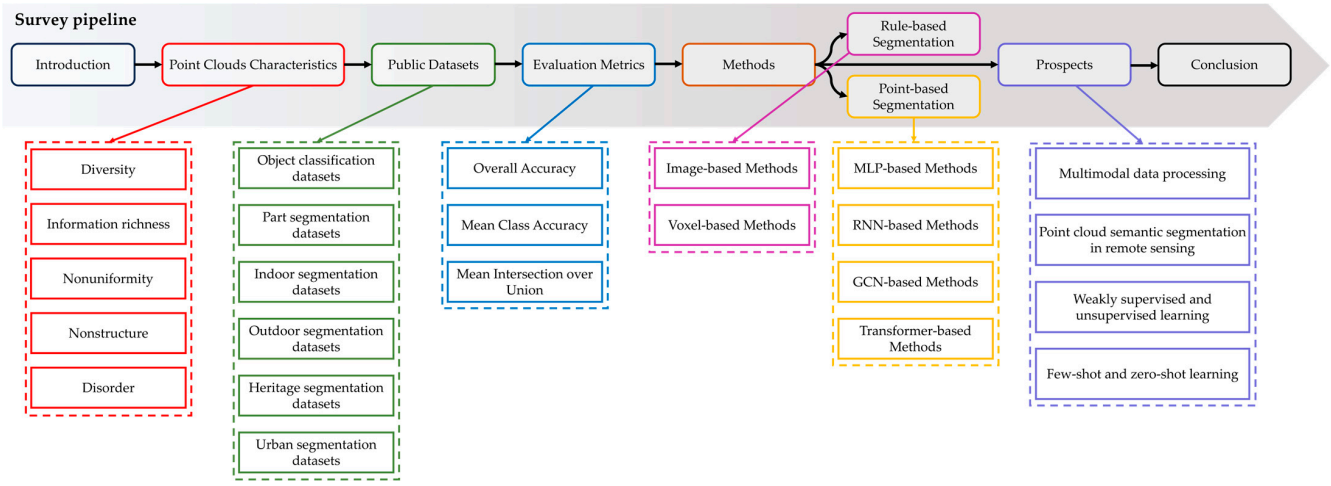


Figure 1. Illustration of survey pipeline. Different colors represent specific sections. Best viewed in color.

2. Point Cloud Characteristics

Compared with 2D images, 3D point clouds not only avoid the impact of the image acquisition process due to the complex structure of the objects, random lighting conditions, partial object occlusion or adhesion, and other limitations but also have the advantages of diversity and rich information contained. Accordingly, the processing and analysis of point clouds have become the focus of research in the field of 3D computer vision. However, since point clouds are characterized by nonuniformity, nonstructure, and disorder, it is necessary to process them effectively according to their characteristics. In this section, a comprehensive introduction to point cloud characteristics is presented to provide some reference for future research on point cloud semantic segmentation.

2.1. Diversity of Point Clouds

Depending on the different data acquisition principles and methods, point clouds can be roughly categorized into three types: image-derived point clouds, light detection and ranging (LiDAR) point clouds, and other point clouds. The image-derived point cloud is mainly obtained by stereo matching methods of RGB-D images acquired from depth sensors using time of flight (ToF), structured light, and other technologies. The LiDAR point cloud is obtained using the time delay between the emission of pulses from the laser and its reflection back to the receiver to measure the distance of the object's surface and combine it with the position and attitude information. According to the different carriers of the LiDAR system, it can be classified into fixed, handheld, vehicle-borne, airborne, and so on. Driven by the rapid development of sensor technology and the demand for applications, novel point clouds have been proposed, such as multisource fusion point clouds [9] and interferometric synthetic aperture radar (InSAR) point clouds [10]. Compared with regular point clouds, novel point clouds have fully demonstrated their value and unique advantages in related research over the past few years [11–13], which provides the possibility for innovation of application scenarios. Figure 2 shows the various types of point clouds acquired by different acquisition methods and devices.

(a) (b) (c) (d) (e)

Figure 2. Illustration of the diversity of point clouds: (**a**) point clouds scanned by Microsoft Kinect devices (SUN RGB-D [14]), (**b**) points generated from CAD models (ModelNet [15]), (**c**) point clouds scanned by Matterport (S3DIS [16]), (**d**) point clouds scanned by a vehicle-mounted laser scanning system (A2D2 [17]), (**e**) point clouds scanned by a mobile laser scanning system (Paris-Lille-3D [18]).

2.2. Information Richness of Point Clouds

The point cloud is the most direct and significant data carrier for describing the real world in the digital era and plays a vital role in national requirements and science research, which contains rich information. This information is derived from multiple dimensions, which is essential in processing and analyzing point clouds. Specifically, geometric information provides the spatial position and structure of the objects. Texture information describes the fine-grained features of the object surface. Color information contains RGB values or reflective intensities obtained from other sensors, which makes point clouds more visually realistic and enhances visualization. Normal information describes the direction of each point normal to the object's surface, which is necessary for tasks such as 3D reconstruction and geometry analysis. Semantic information indicates the object category to which a point belongs, which is necessary to achieve a deep understanding of 3D scenes. The high-dimensional and varied information carried by the point cloud provides a wealthy data resource for further research in 3D computer vision.

2.3. Nonuniformity of Point Clouds

Point cloud acquisition methods based on laser scanning equipment and depth sensors provide a direct and effective means for the 3D digital representation of the real world. For most of the 3D scenes at various scales, the object categories and point density distributions are different, and the penetration capacity of the point cloud acquisition equipment is limited, which can only reflect the surface of the object and almost completely ignore the internal structure, which leads to large differences in point cloud density in different regions, as shown in Figure 3. Therefore, extracting and understanding high-level features in 3D scenes is challenging.

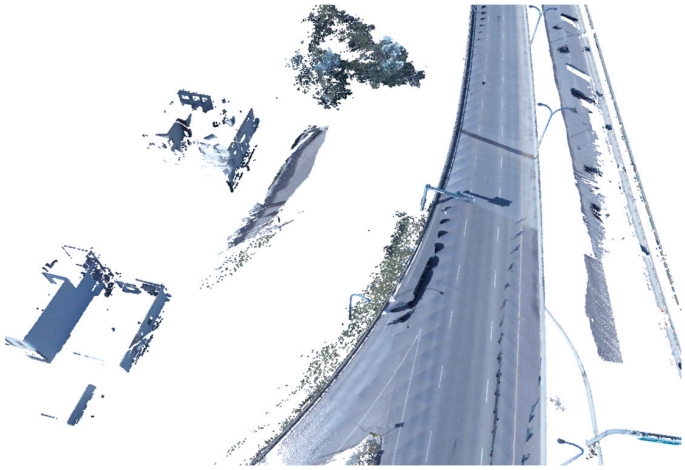


Figure 3. Illustration of the nonuniformity of point clouds.

2.4. Nonstructure of Point Clouds

While 2D images are represented in the computer as the matrix, point clouds are more flexible. As shown in Figure 4, the spatial distribution of points is not limited to certain structured representations, and each local region contains different numbers of points, and the relative positions of pairs of points differ. This unstructured characteristic makes it difficult to manipulate the original point cloud using conventional convolution. For this reason, some researchers have attempted to transform point clouds into regular data that retain the original geometric structure by constructing voxels. Due to the limitation of the resolution size of the voxel and the computing power requirements, voxelization methods inevitably lead to the loss of a large amount of key information, and the complexity of the algorithms grows cubically with the increase in voxel refinement. Therefore, such methods do not apply to the processing of large-scale 3D scenes.

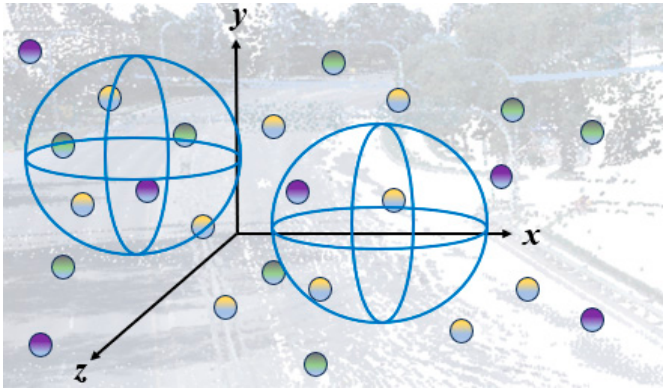
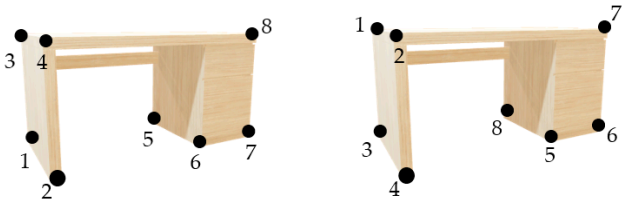


Figure 4. Illustration of the nonstructure of point clouds. Different colors represent different categories.

2.5. Disorder of Point Clouds

The disorder of point clouds means that the point cloud is essentially a collection of disordered points in 3D space. The order of the collected points varies greatly due to the variation of the object's posture, sensor type, and observation platform. The coordinates of each point can independently characterize the spatial position, while for a cluster of point clouds, the initial input order is not necessary, and each point is not associated with other points in the neighborhood. If the $n \times 3$ scale point cloud is input into the neural network, there are $n!$ kinds of arrangement and combination sequences. As shown in Figure 5, changing the order of key points describing the same desk in the figure generates different point cloud matrices and is not affected by the physical storage in the computer.

How to effectively solve the problem of disorder has become the key to the tasks of point cloud registration, point cloud classification, and point cloud semantic segmentation.



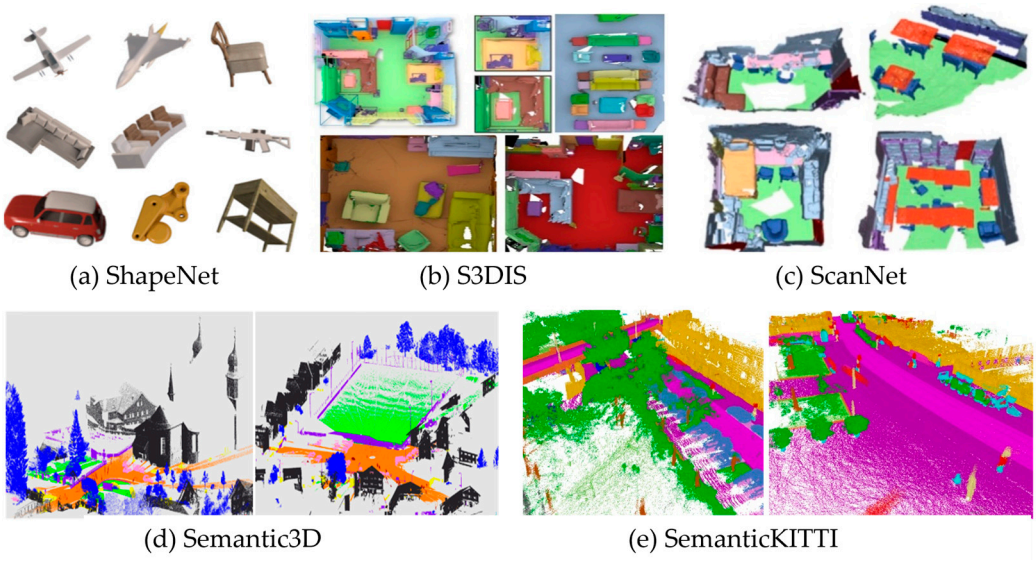
	x	y	z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
3	x_3	y_3	z_3
4	x_4	y_4	z_4
5	x_5	y_5	z_5
6	x_6	y_6	z_6
7	x_7	y_7	z_7
8	x_8	y_8	z_8

	x	y	z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
3	x_3	y_3	z_3
4	x_4	y_4	z_4
5	x_5	y_5	z_5
6	x_6	y_6	z_6
7	x_7	y_7	z_7
8	x_8	y_8	z_8

Figure 5. Illustration of the disorder of point clouds.

3. Public Datasets

With the rise of artificial intelligence, computer vision tasks need to utilize deep neural networks with larger parameter sizes and more complex architectures for high-level feature extraction. High-quality point cloud datasets are an important guarantee for the effective training of networks and the verification of the performance of the proposed segmentation algorithms. However, the collection and labeling of massive data require not only a lot of labor, material, and financial resources but also the guidance of domain experts and professional skills in related industrial software. To promote the development of point cloud semantic segmentation–related research, some research institutions provide semantically informative and reliable public datasets, and the use of these mainstream public datasets for network training and validation not only guarantees the fairness and validity of comparison with other networks but also provides a low-cost and feasible solution for building deep networks with excellent performance. This section highlights five datasets commonly used for point cloud semantic segmentation: ShapeNet [19], S3DIS [16], ScanNet [20], Semantic3D [21], and SemanticKITTI [22]. Figure 6 shows these datasets' annotation examples.



(a) ShapeNet (b) S3DIS (c) ScanNet

(d) Semantic3D (e) SemanticKITTI

Figure 6. (**a**–**e**) show the annotation samples of ShapeNet, S3DIS, ScanNet, Semantic3D, and SemanticKITTI.

ShapeNet: ShapeNet is a large dataset of 3D CAD models with rich annotations, which consists of two parts, ShapeNetCore and ShapeNetSem, where ShapeNetCore contains 55 common categories of about 51,300 3D models, and each model annotation consists of 2–5 parts. ShapeNetSem is a smaller, more densely annotated subset that validates and annotates more than 12,000 3D models in 270 categories with size, volume, shape, and other attributes.

S3DIS: The Stanford 3D Indoor Scene Dataset (S3DIS) is a large indoor scene dataset generated using a Matterport 3D laser scanner. The dataset covers 6 indoor regions consisting of more than 215 million points, 70,496 regular RGB images and 1413 equirectangular RGB images, and 272 indoor scenes with instance-level semantic annotations, covering a total area of more than 6000 m^2, with 13 categories, each point with surface normals, coordinates, semantic annotations, and other attributes. This dataset plays a key role in the learning of indoor scene features in 3D vision.

ScanNet: ScanNet is a dataset of indoor scenes composed of RGB-D video sequences. The dataset consists of 1513 scans of 707 indoor environments, generating 2.5 million RGB-D views with 21 categories. The attributes include not only the precalibration parameters, textures, and coordinates but also instance-level semantic annotations. This dataset is an important contribution to the realization of 3D scene perception.

Semantic3D: Semantic3D is a representative large-scale outdoor scene point cloud dataset, providing more than 30 different scenes, such as churches, stations, squares, soccer fields, and villages. Among them, 15 scenes are used for network training, and the remaining scenes are used for network testing. There are over 4 billion points in the scenes, including attributes such as coordinates, colors, intensity values, and other attributes, covering 8 categories, including artificial terrain, natural terrain, high vegetation, low vegetation, buildings, landscape, cars, and scanning artifacts. Considering the actual hardware situation of the researcher's development environment, two types of subdatasets are provided, semantic-8 and reduced-8. Semantic-8 has the complete test data, while reduced-8 contains only 4 subsets as test cases.

SemanticKITTI: SemanticKITTI is a large point cloud dataset of outdoor scenes around Karlsruhe, Germany, generated by automotive LiDAR, which plays a vital role in the study of the semantic segmentation of road traffic scenes in the field of autonomous driving. The dataset contains about 4.5 billion points in 28 categories, covering 22 sets of scene sequences, including city traffic, residential areas, highways, and rural roads. The sequences 0–10 are used for network training and the sequences 11–21 are used for network testing. This dataset provides a reliable benchmark for evaluating the performance of the models in the task of 3D outdoor scene target recognition and semantic segmentation.

The mainstream datasets of point cloud semantic segmentation are summarized according to name, year, type, application scenario, category, size, and sensor, as shown in Table 1.

Table 1. Summary of mainstream datasets for point cloud semantic segmentation (where R ← real-world environment, S ← synthetic environment in the Type column, Oc ← object classification, Ps ← part segmentation, Is ← indoor segmentation, Os ← outdoor segmentation, Hs ← heritage segmentation, Us ← urban segmentation in the Application Scenario column, Tm ← thousand models, Tf ← thousand frames, To ← thousand objects, Mp ← million points in the Size column, ALS ← airborne laser scanning, MLS ← mobile laser scanning, TLS ← terrestrial laser scanning, - ← information not available in the Sensor column).

Name	Year	Type	Application Scenario	Category	Size	Sensor
ModelNet10 [15]	2015	S	Oc	10	4.9 Tm	-
ModelNet40 [15]	2015	S	Oc	10	12.3 Tm	-
ScanObjectNN [23]	2019	R	Oc	15	15 To	-
ShapeNet [19]	2015	S	Ps	55	51.3 Tm	-

Table 1. *Cont.*

Name	Year	Type	Application Scenario	Category	Size	Sensor
ShapeNet Part [24]	2016	S	Ps	16	16.9 Tm	-
SUN RGB-D [14]	2015	R	Is	47	103.5 Tf	Kinect
S3DIS [16]	2016	R	Is	13	273.0 Mp	Matterport
ScanNet [20]	2017	R	Is	22	242.0 Mp	RGB-D
MIMAP [25]	2020	R	Is	-	22.5 Mp	XBeibao
ArCH [26]	2020	R	Hs	10	102.74 Mp	TLS
KITTI [27]	2012	R	Os	3	179.0 Mp	MLS
Semantic3D [21]	2017	R	Os	8	4000.0 Mp	MLS
Paris-rue-Madame [28]	2018	R	Os	17	20.0 Mp	MLS
Paris-Lille-3D [18]	2018	R	Os	9	143.0 Mp	MLS
ApolloScape [29]	2018	R	Os	24	140.7 Tf	RGB-D
SemanticKITTI [22]	2019	R	Os	25	4549.0 Mp	MLS
Toronto-3D [30]	2020	R	Os	8	78.3 Mp	MLS
A2D2 [17]	2020	R	Os	38	41.3 Tf	TLS
SemanticPOSS [31]	2020	R	Os	14	216 Mp	MLS
WHU-TLS [32]	2020	R	Os	-	1740.0 Mp	TLS
nuScenes [33]	2020	R	Os	31	34.1 Tf	Velodyne HDL-32E
PandaSet [34]	2021	R	Os	37	16.0 Tf	MLS
Panoptic nuScenes [35]	2022	R	Os	32	1100.0 Mp	MLS
TJ4DRadSet [36]	2022	R	Os	8	7.75 Tf	4D Radar
DALES [37]	2020	R	Us	8	505.0 Mp	ALS
LASDU [38]	2020	R	Us	5	3.12 Mp	ALS
SensatUrban [39]	2022	R	Us	13	2847.0 Mp	UAV Photogrammetry

4. Evaluation Metrics

For the quantitative evaluation of the model's performance, mainstream evaluation metrics are needed to sufficiently guarantee the fairness and validity of the experimental results. At present, researchers mostly use execution time, complexity, and accuracy as the benchmark for evaluating models. However, the time overhead of the segmentation algorithms is closely related to the hardware systems used by researchers, and few researchers provide data about the time and space complexity of the proposed methods. Therefore, this paper focuses on the accuracy evaluation metrics of the methods.

Presently, overall accuracy (OA), mean class accuracy (mAcc), and mean intersection over union (mIoU) are used as the metrics to evaluate the performance of point cloud semantic segmentation methods. For the convenience of description, the notations appearing later are indicated here: Assuming that there are $N + 1$ semantic classes (including empty class), M_{ij} denotes the number of units with actual semantic type i but predicted type j and vice versa for M_{ji}. M_{ii} denotes the number of units with actual semantic type i and predicted type i.

OA: OA is the ratio of the number of samples correctly predicted by the segmentation algorithms to the total number of samples, as shown in Equation (1):

$$\mathrm{OA} = \frac{\sum_{i=0}^{N} M_{ii}}{\sum_{i=0}^{N} \sum_{j=0}^{N} M_{ij}} \tag{1}$$

mAcc: mAcc is an improvement of OA, which calculates the precision for each category separately, and then averages the summed results according to the number of categories, as shown in Equation (2):

$$\mathrm{mAcc} = \frac{1}{N+1} \sum_{i=0}^{N} \frac{M_{ii}}{\sum_{j=0}^{N} M_{ij}} \tag{2}$$

mIoU: mIoU is the most important index to evaluate the performance of the segmentation methods, which first calculates the ratio between the intersection of the predicted and true regions of the models for each category, and then calculates the average value of the summed results according to the number of categories, as shown in Equation (3):

$$\mathrm{mIoU} = \frac{1}{N+1} \sum_{i=0}^{N} \frac{M_{ii}}{\sum_{j=0}^{N} M_{ij} + \sum_{i=0}^{N} M_{ji} - M_{ii}} \tag{3}$$

Considering the simplicity and representativeness, three evaluation metrics, OA, mAcc, and mIoU, are selected in this paper to compare and analyze different point cloud semantic segmentation methods for researchers' reference.

5. Point Cloud Semantic Segmentation Methods

In recent years, with the increasing development of deep learning technologies, point cloud semantic segmentation based on deep learning has attracted a great deal of attention from researchers. These methods achieve automatic extraction of point cloud features and better performance than conventional methods [40–42]. Deep-learning-based point cloud semantic segmentation methods can be divided into rule-based and point-based methods. The key to rule-based methods is to solve the problem of disordered and unstructured input data. This type of method transforms the original point cloud into structured data that can be easily represented, and then inputs into the networks to extract features, thus realizing semantic segmentation. Point-based methods directly use point clouds and extract the features. Appendix A shows the timeline of the development of deep-learning-based point cloud semantic segmentation methods since 2015.

5.1. Rule-Based Segmentation

Existing research classifies rule-based point cloud semantic segmentation methods into two main strategies: image-based methods and voxel-based methods. This type of method transforms point clouds into structured data that can be processed by conventional CNNs, as shown in Figure 7.

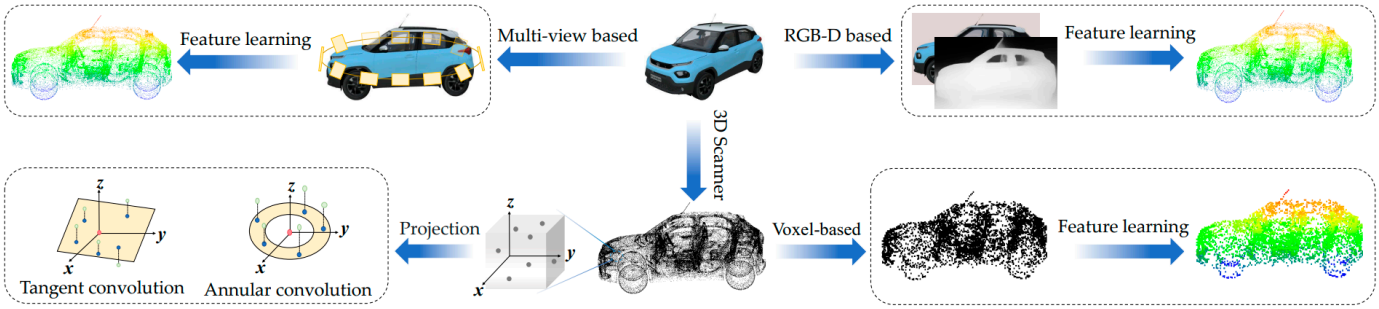


Figure 7. Illustration of point cloud regularization. Points of different colors represent different categories.

5.1.1. Image-Based Methods

(1) Multiview Image-Based

In the early stage, deep-learning-based methods could not deal with 3D data effectively and required dimensionality reduction. Su et al. [43] projected the original point cloud in multiple viewpoints to obtain 2D images from different viewpoints, then used the proposed network MVCNN to extract features and aggregate them in the pooling layer, and finally remapped the aggregated features back to the point cloud to achieve segmentation. This method achieves better accuracy and is a pioneer in solving the unstructured problem of point clouds. Feng et al. [44] improved MVCNN by increasing the number of projection views, obtaining feature vectors by CNN for images obtained from 12 views individually, and grouping the prediction scores from the fully connected layers. The group-level features are combined into the object features by weighting and then averaging between different groups. Aiming to improve the problem of point cloud loss in structured processing, You et al. [45] proposed a point cloud segmentation network, PVRNet, that fully considers the relationship between points and views, which fully integrates view features and points features through the correlation prediction module and proposes two correlation feature fusion methods, i.e., point cloud correlation feature fusion methods with a single viewpoint and point cloud correlation feature fusion method with multiple viewpoints. Finally, the features of both are aggregated to further improve the network's capacity to understand the deep-level features of objects in 3D scenes. Milioto et al. [46] proposed an efficient GPU-based k-NN postprocessing method that can be used to address discretization and inferential ambiguity. Robert et al. [47] computed occlusion-aware mappings between 3D points and 2D pixels, and then aggregated relevant image features for each point through observation conditions based on the attention scheme. This method achieved 74.4 mIoU on S3DIS with sixfold cross-validation, which set a new state-of-the-art for large-scale indoor semantic segmentation. However, since multiview is only an approximate abstraction of the object, there might be partial occlusions and defects in the objects themselves, and it is difficult to cover all objects for large-scale scenes with multiview image-based methods. Therefore, few such methods have been used for point cloud semantic segmentation in recent studies.

(2) RGB-D Image-Based

The depth image takes the distance from the laser scanning device to each object in space as the key information and reflects the geometry of the object's surface. Depth images are usually generated from point clouds in spherical coordinates based on azimuth and zenith angles. Boulch et al. [48] proposed a semantic segmentation network, SnapNet, for fusing depth image features and achieved impressive results on semantic-8. The method first preprocesses the point cloud and generates viewpoints, selects different viewpoints to generate RGB images and depth images, and then uses a fully convolutional neural network to annotate the RGB images and depth images, and finally back-projects the labels into the point cloud to obtain the semantic segmentation results. Guerry et al. [49] improved SnapNet by proposing SnapNet-R, which can process multiple views simultaneously compared with SnapNet, thus obtaining more dense labels and further improving the performance. Since the maximum pooling operation during feature aggregation leads to a partial loss of local information, Wu et al. [50] proposed SqueezeSeg, a point cloud semantic segmentation network based on a conditional random field (CRF) and depth images, which uses spherical projection to transform sparse point clouds into 2D images to feed into SqueezeNet [51] for 3D classification and semantic segmentation and uses CRF as the recursive layers to further optimize the results. However, the accuracy of this method is sensitive to the noise generated in the point cloud acquisition process. SqueezeSegV2 [52] improves SqueezeSeg by adding the context aggregation module (CAM) to increase the perceptual field of the network and improve the efficiency of using contextual information, which makes the network more robust to the noises and outliers generated during point cloud acquisition. Considering the nonuniform distribution of spatial features in point clouds, Xu et al. [53]

proposed SqueezeSegV3 with spatially adaptive convolution (SAC), which uses different filters for different neighborhood locations in the point cloud projection–generated images, thus making full use of the capacity of the network. In a recent study, Yang et al. [54] proposed a novel framework, SAM3D, by leveraging the Segment Anything Model (SAM) for 3D vision, which first utilizes SAM to predict the segmentation results of RGB images and then adopts the bidirectional merging approach to project the 2D masks of adjacent frames into 3D point clouds. Finally, the 3D masks predicted from different frames are gradually merged into the 3D mask of the whole 3D scene. Table 2 compares the performance of image-based point cloud semantic segmentation methods on the datasets.

Table 2. Comparison of image-based point cloud semantic segmentation methods.

Method	Year	Dataset	Performance			Contribution
			OA	mAcc	mIoU	
MVCNN [43]	2015	ModelNet40	90.1%	-	-	The first multiview CNN
SnapNet [48]	2017	Sun RGB-D	-	67.4%	-	Generate RGB and depth views by 2D image views
		Semantic3D	88.6%	70.8%	59.1%	
SnapNet-R [49]	2017	Sun RGB-D	78.1%	-	38.3%	Improvements to SnapNet
GVCNN [44]	2018	ModelNet40	93.1%	-	-	Grouping module to learn the connections and differences between views
SqueezeSeg [50]	2018	KITTI	-	-	29.5%	Data conversion from 3D to 2D using spherical projection
SqueezeSegV2 [52]	2018	KITTI	-	-	39.7%	Introducing a context aggregation module to SqueezeSeg
PVRNet [45]	2019	ModelNet40	93.6%	-	-	Consider relationships between points and views, and fuse features
RangeNet++ [46]	2019	KITTI	-	-	52.2%	GPU-accelerated postprocessing +RangNet++
SqueezeSegV3 [53]	2020	SemanticKITTI	-	-	55.9%	Proposing the spatially adaptive and context-aware convolution
Robert et al. [47]	2022	S3DIS	-	-	74.4%	Introducing an attention scheme for multiview image-based methods
		ScanNet	-	-	71.0%	

5.1.2. Voxel-Based Methods

The use of voxelization methods to handle point clouds is another idea for transforming unstructured data into structured data. The process of voxelization is to represent an object as voxels that are closest to the object. VoxNet [55] was the first to use the voxelization method to transform unstructured point clouds into regular voxels and then use 3D CNN to predict the semantic labels of the occupied voxels by standard convolution operations. Although this method solved the problem of unstructured point clouds, it also had the limitations of low efficiency of voxel arrangement due to the sparsity and high computational complexity of 3D CNN. Su et al. [56] designed SPLATNet for sparse voxels, which first interpolates the original point cloud to the sparse voxel by splat operation, then convolves the occupied voxels by convolve operation, and finally, interpolates the output features to the original point cloud by slice operation. This method significantly improves the efficiency by using the index structure to convolve only the occupied voxels. To alleviate the impact of a point cloud scale on performance, Rosu et al. [57] proposed LatticeNet with PointNet as the backbone, which can convolve sparse voxels quickly while keeping the computational overhead low and then project the features back to the point cloud through the DeformSlice module. This method has shown effectiveness in handling large-scale point clouds. Tchapmi et al. [58] proposed an end-to-end semantic segmenta-

tion network, SEGCloud, combined with a 3D fully convolutional network, which first voxelizes the point cloud, then applies 3D CNN to generate downsampled voxel labels, and then transforms the voxel labels back to point labels by a trilinear interpolation layer, finally, combining the point features with the interpolated scores using a 3D fully connected conditional random field and postprocessing to obtain fine-grained semantic information. However, due to the sparsity of the point cloud itself, the voxelized units are still sparse and discrete, and these cause unnecessary computational overhead. In response, researchers have tried to transform sparse point clouds into nonuniform voxels, for example, using the octree instead of fixed-size voxels. OctNet, proposed by Riegler et al. [59], uses the octree to divide 3D scenes into nonuniform voxels of varying sizes according to the distribution density of points and allows computational resources to be concentrated in voxel-dense regions, which saves computational overhead to some extent. O-CNN, proposed by Wang et al. [60], uses the octree to divide the point cloud into several nodes, takes the average normal vector of nodes as input of the network, and utilizes 3D CNN for feature learning. The complexity of the method grows squarely with the depth of the octree, which saves computational resource consumption to some extent and is suitable for 3D classification and semantic segmentation tasks of high-resolution voxels. For more effective handling of sparsely distributed points, Meng et al. [61] proposed a kernel-based interpolated variational autoencoder architecture to encode the local geometry within each voxel and utilized the radial basis function to compute a local, continuous representation within each voxel. This method provides richer fine-grained features without increasing parameters, improving the expressive capacity and leading to more robust results. Recently, some meaningful work was presented where PCSCNet [62] avoids the discretization error from larger-sized voxels through cross-entropy loss and position-aware loss, keeping the efficiency in the case of lower voxel resolutions. SIEV-Net [63] utilizes a hierarchical voxel feature encoding module and a height information complement module to minimize the height information lost during the point feature aggregation process. Table 3 compares the performance of voxel-based point cloud semantic segmentation methods on the datasets.

Table 3. Comparison of voxel-based point cloud semantic segmentation methods.

Method	Year	Dataset	Performance			Contribution
			OA	mAcc	mIoU	
VoxNet [55]	2015	ModelNet10	-	92.0%	-	The first method to process raw point clouds using voxelization
		ModelNet40	85.9%	83.0%	-	
SEGCloud [58]	2015	ShapeNet Part	-	-	79.4%	Combining 3DFCNN with fine representation using trilinear interpolation and conditional random field
		ScanNet	73.0%	-	-	
		S3DIS	-	57.4%	48.9%	
		Semantic3D	88.1%	73.1%	61.3%	
		KITTI	-	49.5%	36.8%	
OctNet [59]	2017	ModelNet10	90.0%	-	-	Divide the space into nonuniform voxels using unbalanced octrees
		ModelNet40	83.8%	-	-	
O-CNN [60]	2017	ModelNet40	90.2%	-	-	Making 3D-CNN feasible for high-resolution voxels
		ShapeNet Part	-	-	85.9%	
SPLATNet [56]	2018	ShapeNet Part	-	83.7%	-	Hierarchical and spatially aware feature learning
VV-Net [61]	2019	ShapeNet Part	-	-	87.4%	Using the radial basis function to compute the localized continuous representation within each voxel
		S3DIS	87.8%	-	78.2%	
LatticeNet [57]	2020	ShapeNet Part	-	83.9%	-	Proposing a novel slicing operator for computational efficiency
		ScanNet	-	-	64.0%	
		SemanticKITTI	-	-	52.9%	

Table 3. *Cont.*

Method	Year	Dataset	Performance			Contribution
			OA	mAcc	mIoU	
PCSCNet [62]	2022	nuScenes	-	-	72.0%	Reducing the voxel discretization error
		SemanticKITTI	-	-	62.7%	
SIEV-Net [63]	2022	KITTI	-	-	62.6%	Effectively reduces loss of height information

5.2. Point-Based Segmentation

Rule-based segmentation methods solve the limitation that 2D CNNs cannot be directly applied to point clouds, but there are challenges, such as loss of key information and high complexity. To solve the mentioned challenges, researchers have started to focus on the research of point-based segmentation, which can be divided into multilayer perceptron–based method (MLP-based method), recurrent neural network–based method (RNN-based method), graph convolution network–based method (GCN-based method), and transformer-based method (transformer-based method). Figure 8 shows the basic framework of the point-based segmentation network. It should be noted that the internal structures of the encoder and decoder are different for each network.

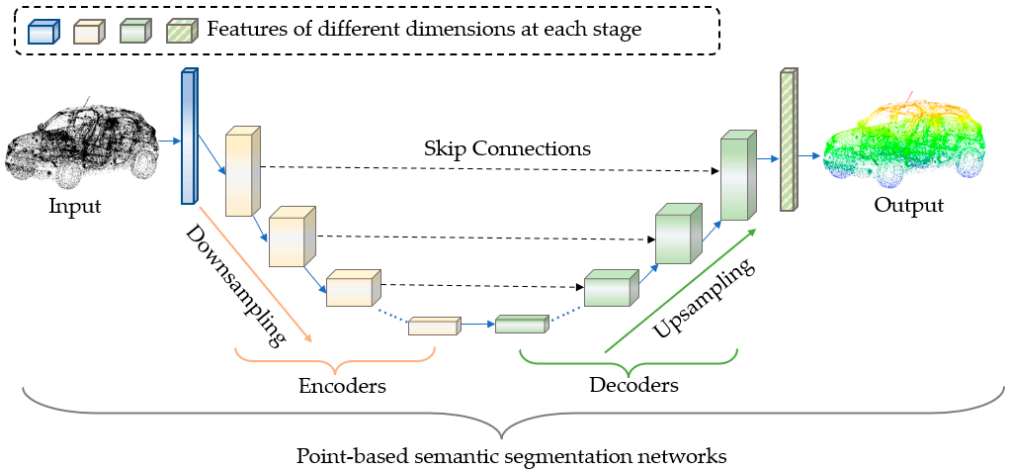


Figure 8. Basic frameworks of point-based CNNs. Points of different colors represent the learned features of different car parts.

5.2.1. MLP-Based Methods

MLP is a commonly used architecture in point cloud processing, which is a feed-forward neural network consisting of multiple fully connected layers. MLP-based methods usually use shared MLPs as the basic structure of the network, which means that all points in the point cloud share the same parameters. Qi et al. [64] proposed a pioneering network, PointNet, which takes the original point cloud as input, sums the feature of each point by the symmetric function and extracts the feature vector with the maximum value in each dimension, extracts the features of each point independently using MLP, and finally aggregates the features of all points using the maximum pooling layers to obtain the global representation. PointNet effectively solves the problems of permutation invariance and rotation invariance of point clouds. However, the local and interaction information with other points in the neighborhood learned by PointNet is insufficient because deeper layer features cannot cover a larger spatial extent. To address this limitation, Qi et al. [65] improved PointNet by proposing a deep hierarchical network, PointNet++, which consists of the sampling layer, grouping layer, and PointNet backbone network. First, the farthest point sampling (FPS) algorithm is used to select the point with the largest spatial separation in the high-dimensional space as the center of the local region to ensure that the data dimensionality is reduced while preserving the main geometrical structure, and

then the local regions are constructed by using the grouping module. Finally, the backbone network is used to recursively learn the features of the local region. Although this network solves the problem of extraction of local features, the capacity to capture information, such as direction and distance between points, is still insufficient. Jiang et al. [66] developed a PointSIFT module that can efficiently explore the neighborhoods in multiple directions. The module uses orientation-encoding units to describe eight crucial orientations and achieves the learning of multiscale features by stacking several orientation-encoding units. To capture the correlation between neighboring points, Zhao et al. [67] designed PointWeb, a network based on the adaptive feature adjustment (AFA) module. The network densely connects each point with others in a local region, exploring in depth the interactions between point pairs. For each local region, an impact map carrying the impact of the elements between point pairs is applied to the feature difference map. Then, the features are adaptively pushed and pulled according to the adaptively learned impact indicators, which in turn achieves the dynamic adjustment and assignment of features beneficial in the point cloud classification and segmentation tasks. SO-Net, proposed by Li et al. [68], models the spatial distribution of point clouds by constructing a self-organizing map (SOM) and performs hierarchical feature extraction on each point and SOM node, and then aggregates the obtained set of feature vectors into global features by averaging pooling. Finally, the semantic features representing the input point cloud are recovered from the global features. Zhang et al. [69] proposed a novel yet effective ShellConv convolution operator that uses the statistics of concentric spherical shells to define representative features to resolve the ambiguity of point order, enabling conventional convolution to be performed on these features. Based on ShellConv, an efficient neural network named ShellNet is further built, which recursively computes each spatial neighborhood and aggregates the statistics of different regions by maximum pooling, while maintaining fewer layers, achieving the balance of efficiency and accuracy.

To further improve the networks' capacity to understand 3D scenes, researchers have tried to introduce the attention mechanism in MLP-based methods. Yan et al. [70] designed PointASNL with strong robustness to noisy point clouds through an adaptive sampling (AS) module based on the attention mechanism. This module adaptively adjusts the features of the sampled points by augmenting the neighborhood points obtained from the FPS algorithm and reweighting the features according to the learned attention weights, thus effectively mitigating the bias caused by the outliers in original point clouds. Hu et al. [71] proposed a lightweight neural network, RandLA-Net, for large-scale point cloud processing, which introduces local spatial encoding (LocSE) units to preserve geometric features and uses the attention-based pooling unit to achieve feature aggregation. By stacking LocSE units and pooling units to increase the perceptual field, the network effectively enhances the understanding of local regions and achieves significant improvement in computational efficiency. Ma et al. [72] provided a new perspective by designing the pure residual MLP network PointMLP, a model equipped with a proposed lightweight geometric affine module that achieves state-of-the-art performance on the ScanObjectNN dataset. Table 4 compares the performances of MLP-based point cloud semantic segmentation methods on the datasets.

Table 4. Comparison of MLP-based point cloud semantic segmentation methods.

Method	Year	Dataset	Performance			Contribution
			OA	mAcc	mIoU	
PointNet [64]	2017	ModelNet40	89.2%	86.2%	-	The first method for processing raw point clouds
		ShapeNet Part	-	-	83.7%	
		ScanNet	73.9%	-	14.7%	
		S3DIS	78.6%	-	47.7%	

Table 4. *Cont.*

Method	Year	Dataset	Performance			Contribution
			OA	mAcc	mIoU	
PointNet++ [65]	2017	ModelNet40	90.7%	-	-	Improvements to PointNet and design of hierarchical network architecture
		ShapeNet Part	-	-	85.1%	
		ScanNet	84.5%	-	34.3%	
		S3DIS	81.0%	-	54.5%	
SO-Net [68]	2018	ModelNet10	94.1%	-	-	SOM for modeling the spatial distribution of points
		ModelNet40	90.8%	-	-	
		ShapeNet	-	-	84.6%	
PointSIFT [66]	2018	ScanNet	86.2%	-	-	Integration of multidirectional features using orientation-encoding convolution
		S3DIS	88.7%	-	70.2%	
PointWeb [67]	2019	ModelNet40	92.3%	89.4%	-	Proposing an adaptive feature adjustment module for interactive feature exploitation
		S3DIS	86.9%	66.6%	60.3%	
ShellNet [69]	2019	ScanNet	85.2%	-	-	Proposing an efficient point cloud processing network using statistics from concentric spherical shells
		S3DIS	87.1%		66.8%	
		Semantic3D	93.2%	-	69.4%	
RandLA-Net [71]	2020	Semantic3D	94.8%	-	77.4%	Proposing a lightweight network that exploits large receptive fields and keeps geometric details through LFAM
		SemanticKITTI	-	-	53.9%	
PointASNL [70]	2020	ModelNet10	95.9%	-	-	Proposing a local–nonlocal module with strong noise robustness
		ModelNet40	93.2%	-	-	
		ScanNet	-	-	63.0%	
		S3DIS	-	-	68.7%	
PointMLP [72]	2022	ModelNet40	94.1%	91.5%	-	A pure residual MLP network
		ScanObjectNN	86.1%	84.4%	-	

5.2.2. RNN-Based Methods

In the field of 2D image processing, RNNs can better capture the contextual information between pixels to significantly improve the learning ability of deep neural networks. In the field of point cloud processing, RNNs can also be used to learn the contextual information between point pairs. Fan et al. [73] proposed a point recurrent neural network for moving point cloud processing, which achieves the fusion of pointwise features and state features by correlating the spatiotemporal information and better solves the limitation that the features from points in different periods cannot be operated directly due to the disorder of point clouds. To better capture the multiscale contextual interaction information and achieve the extraction of adjacent features, Ye et al. [74] proposed a novel end-to-end semantic segmentation network named 3P-RNN to solve the problem of extracting local geometric features under different point density distributions. This 3P-RNN consists of two main components, namely, pointwise pyramidal pooling module and bi-directional hierarchical RNN; the former is used to extract contextual interaction information at different scales to achieve multilevel semantic feature fusion, and the latter for capturing long-range spatial relations. Huang et al. [75] designed a lightweight segmentation network, RSNet, which can efficiently learn the local geometric structure. The network consists of a slice pooling layer, RNN layers, and a slice unpooling layer. Specifically, the slice pooling layer maps the features of unordered points into an ordered sequence of feature vectors, then inputs the sequence into RNN layers for processing and updating, thus achieving effec-

tive interaction of spatial contextual information. In the end, the slice unpooling layer reverses the projection and assigns updated features to each point to obtain the semantic segmentation results.

Zhao et al. [76] proposed DAR-Net, a point cloud segmentation network supporting dynamic feature aggregation, fully considering the differences between the sizes of objects in complex 3D scenes. The network uses RNN to recursively process disordered point clouds, forms a backbone consisting of key points by aggregating middle-level features, and adaptively adjusts the model perceptual field as well as key point weights, thus achieving an accurate grasp of local and global features. Experimental results show that the proposed approach outperforms static pooling methods significantly when dealing with large-scale point clouds. The 3DCNN-DQN-RNN proposed by Liu et al. [77] uses 3D CNN to learn and encode the location, color, and other attributes of points from multiscale; efficiently locates the position of points belonging to a particular category through a deep Q-network (DQN); and feeds the correlated feature vectors into the residual RNN to further extract richer high-level features. Table 5 compares the performances of RNN-based point cloud semantic segmentation methods on the datasets.

Table 5. Comparison of RNN-based point cloud semantic segmentation methods.

Method	Year	Dataset	Performance			Contribution
			OA	mAcc	mIoU	
3DCNN-DQN-RNN [77]	2017	S3DIS	70.8%	-	-	Combining 3DCNN, DQN, and RNN in a single framework
3P-RNN [74]	2018	S3DIS	86.9%	-	56.3%	Designing a novel pointwise pyramid pooling module
RSNet [75]	2018	ShapeNet Part	-	-	84.9%	Processing point clouds using bidirectional RNN
		ScanNet	-	48.4%	39.4%	
		S3DIS	-	59.4%	51.9%	
DAR-Net [76]	2019	ScanNet	-	61.6%	55.8%	The network supports dynamic feature aggregation

5.2.3. GCN-Based Methods

A graph convolutional network (GCN) models the real-world problem as the interaction and information transfer between neighboring nodes in a graph and has been widely used in knowledge graphs, recommendation systems, and other fields. To this end, researchers further extend the applicability of GCN by transforming the point cloud into a graph structure and formulating computational strategies for nodes and edges, fully exploiting the interaction between point pairs and effectively transferring the learned information, which provides a new paradigm and solution for a deeper semantic perception. To solve the problem of feature homogeneity in graphs, Simonovsky et al. [78] designed a point cloud segmentation network by setting fixed radiuses to divide spatial regions and then connecting the neighboring points in the same region with edges and assigning attributes, such as coordinates, color, and intensity values, to achieve the construction of the graph structure. By performing edge-conditioned convolution (ECC) in the neighborhood, the extraction of edge features between point pairs in the local area is achieved. Wang et al. [79] proposed the dynamic graph convolutional neural network (DGCN), which extracts the features of the centroid by constructing local neighborhood graphs and using dynamic edge convolution (EdgeConv) to obtain the edge feature vectors of the centroids and k-nearest neighboring points. Then, the global features and the local spatial features output by each EdgeConv are fused to further improve the network's capacity to recognize similar features in the feature space and the semantic segmentation performance. However, DGCN has a high computational complexity when performing EdgeConv and suffers from the problem of network gradient disappearance. Lei et al. [80] proposed

a discrete spherical convolution (SPH3D) operator, which divides the spatial region nonuniformly on the spherical coordinate system and specifies a set of trainable parameters to extract features. This metric-based kernel is applied in GCN without relying on edge convolution, which makes more benefits in computational efficiency. Lu et al. [81] designed the PointNGCNN with the feature matrix and Laplacian matrix of each neighborhood as inputs and used the neighborhood graph filter constructed based on Chebyshev polynomials to achieve the learning of neighborhood geometric features in Cartesian space and feature space. Finally, the pointwise semantic descriptors are obtained by fully connected layers. Experimental results show that PointNGCNN achieves good performance in the 3D recognition and segmentation tasks. Li et al. [82] proposed point convolution (P_{conv}) and point pooling (P_{pool}) for 3D points based on the graph structure and designed a novel point cloud feature learning network, PointVGG. Among them, P_{conv} learns the geometric information between the center point and its neighboring points. P_{pool} acquires a more detailed local geometric representation by aggregating points. Zhang et al. [83] proposed an architecture AF-GCN based on graph convolution and the self-attention mechanism. The network uses graph convolution to learn local features in the shallow coding stages, and in the deeper stages, long-range contexts are modeled more efficiently by the graph attentive filter (GAF).

For most GCNs, convolution operations are usually only suitable for the feature extraction of structurally fixed graphs. Considering the complexity of graph structures and the heterogeneity in connecting modes, Zhang et al. [84] efficiently organized the point cloud by constructing a hybrid index structure based on Kd-Octree and generated patch-based feature descriptors at leaf nodes as input for 3D pairwise point cloud matching. Li et al. [85] designed an adaptive graph convolutional neural network, AGCN, which can take arbitrary-sized graphs as input. The network uses spectral graph convolution (SGC) to achieve the adaptive transformation of graph topology based on the scale of inputs and the relevance of contextual information, which better solves the problem of inadequate learning of contextual information and geometric features. Landrieu et al. [86] designed a novel deep-learning-based network to address the challenge of large-scale point clouds in semantic segmentation, which, when unsupervised, partitions the original point cloud into geometrically homogeneous elements, represents them as superpoints and constructs a superpoint graph (SPG). SPGs provide rich edge features and accurate representations of contextual relationships between object parts in point clouds by embedding superpoints and using a gated recurrent unit (GRU), and experimental results show impressive results in Semantic3D and S3DIS datasets. Geng et al. [87] proposed a structural representation algorithm for local embedding superpoint graphs (LE-SPG) and then designed a gated integration graph convolutional network (GIGCN) for feature learning and semantic segmentation of the graphs. To prevent the model from gradient vanishing or exploding during training, the hidden states of gated recurrent units (GRUs) in each layer are integrated using a new layer called gated hidden state integration (GHSI), and backpropagation is strengthened by giving the loss function direct access to each layer, fully absorbing the features from different layers.

GCN-based methods extend convolution operations and graph representations to 3D space, which provides a new research idea for processing raw point clouds. At present, researchers have enhanced the learning capacity of networks for local and global information by introducing attention mechanisms and constructing dynamic graphs, which have led to significant achievements of GCNs in the field of point cloud processing. Table 6 compares the performances of GCN-based point cloud semantic segmentation methods on the datasets.

Table 6. Comparison of GCN-based point cloud semantic segmentation methods.

Method	Year	Dataset	Performance			Contribution
			OA	mAcc	mIoU	
SPG [86]	2018	S3DIS	85.5%	73.0%	62.1%	Introducing superpoint graph with rich edge features
		SensatUrban	85.3%	44.4%	37.3%	
		Semantic3D (reduced-8)	94.0%	-	73.2%	
		Semantic3D	92.9%	-	76.2%	
DGCN [79]	2019	ModelNet40	92.2%	90.2%	-	Proposing the EdgeConv operator
		ShapeNet Part	-	-	85.2%	
		S3DIS	84.1%	-	56.1%	
SPH3D-GCN [80]	2020	ModelNet40	85.5%	73.0%	62.1%	Proposing the SPH3D operator
		S3DIS	86.4%	66.5%	58.0%	
PointNGCNN [81]	2020	ModelNet40	92.8%	-	-	Using Chebyshev polynomials as the neighborhood graph filter to extract neighborhood geometric features
		ShapeNet Part	-	-	85.6%	
		ScanNet	84.9%	-	-	
		S3DIS	87.3%	-	-	
PointVGG [82]	2021	ModelNet40	93.6%	91.0%	-	Proposing point convolution and point pooling operations
		ShapeNet Part	-	-	86.1%	
AF-GCN [83]	2023	ShapeNet Part	-	-	85.3%	Combining graph convolution and self-attention mechanisms
		ScanNet	-	-	71.8%	
		S3DIS	-	-	78.4%	
3DGraphSeg [87]	2023	Semantic3D	94.7%	-	76.8%	Proposing a local embedding super-point graph to alleviate gradient vanishing or exploding

5.2.4. Transformer-Based Methods

Transformer is a new deep learning architecture based on self-attention mechanisms, which was originally applied to natural language processing (NLP) tasks, such as sentiment analysis and machine translation. In recent years, inspired by the fruitful results in NLP, researchers have tried to apply Transformer to the field of computer vision and achieved impressive results [88–90]. Point clouds are essentially a set of unordered, unstructured sparse points, and the core of the Transformer architecture is the self-attention mechanism and feed-forward neural network, which does not depend on the order of the points and is more suitable for point cloud processing than CNN architectures.

Guo et al. [91] innovatively introduced the Transformer architecture into point cloud processing and proposed a novel network PCT for point cloud classification and semantic segmentation. The network uses coordinate-based input embedding modules and offset-attention modules with strong robustness to ensure the inherent order invariance of transformers to avoid the ordering of the point cloud and conducts feature learning through the self-attention mechanism. Zhao et al. [92] designed self-attention layers for point clouds and applied these to construct self-attention Point Transformer networks for point cloud processing. This network is based on self-attention operators, using the subtraction relation and adding the trainable, parameterized position encoding to the attention vector and transformation features. In addition, residual point transformer blocks are constructed with the Point Transformer as the core to facilitate the exchange of information between local feature vectors. Engel et al. [93] designed a multiheaded attention network with strong robustness to point clouds, which constructs input sequences by top-k operations and extracts the latent features of local geometric and spatial relations from different subspaces

based on the learned scores through SortNet. Then, the local features are correlated with the global features through a multiheaded attention mechanism, which then better captures spatial relationships and geometric features and demonstrates competitive performance in point cloud classification and segmentation tasks.

To address the problem of the large computational overhead of multihead attention mechanisms, Yang et al. [94] designed a point cloud processing network named PAT with group shuffle attention (GSA) and Gumbel subset sampling (GSS) as the core operations, which largely improved the performance by deeply mining the relationships between the elements of point sets. Among them, GSA is a parameter effective for self-attention operation for learning relationships between points. GSS serves as an effective alternative to the widely used FPS with the advantages of permutation invariance, task agnostic, and differentiability, which enables effective learning on high-dimensional representations. Zhong et al. [95] designed a novel point-based network named multilevel multiscale transformer (MLMST), which consists of three modules: point pyramid transformer (PPT), multiscale transformer (MST), and multilevel transformer (MLT). Among them, PPT captures context information from different resolutions and scales, MST aims to model the context interaction across different scales and enhances the expressive capability of the network, and MLT learns the cross-level information interaction to further aggregate geometric and semantic features. Han et al. [96] designed a deep neural network, named DTNet, mainly consisting of dual point cloud transformer (DPCT) modules, which enhances the information transfer and interaction by aggregating the pointwise and channelwise multihead self-attention models to efficiently learn contextual features at different resolutions and scales from the perspective of spatial position and channel and connecting the outputs of different modules element by element. In turn, the expression capability of the network is improved. Lai et al. [97] proposed Stratified Transformer, which can be used to capture long-range contexts and demonstrates high performance in point cloud segmentation. For each query point, it densely samples nearby points and sparse distant points in a stratified way. In addition, to cope with the challenges posed by irregular point arrangements, the network's representation and generalization capabilities are further enhanced by designing adaptive contextual relative position encoding and point embedding to achieve an effective fusion of local and long-range features. Most existing Transformer-based methods provide the same feature-learning paradigm for all 3D points, ignoring the huge differences in object sizes in 3D scenes. In this regard, Zhou et al. [98] designed a novel size-aware Transformer framework that introduces multiscale features to each attention layer and allows each point to adaptively choose its attentive fields through the multigranular attention (MGA) scheme and the reattention module. Experimental results show that SAT achieves balanced performance on different categories of S3DIS and ScanNet datasets, which demonstrates the superiority of modeling categories of different sizes. Table 7 compares the performance of Transformer-based point cloud semantic segmentation methods on the datasets.

Table 7. Comparison of Transformer-based point cloud semantic segmentation methods.

Method	Year	Dataset	Performance			Contribution
			OA	mAcc	mIoU	
PAT [94]	2019	ModelNet40	91.7%	-	-	Pioneering Transformer-based processing of point clouds
		S3DIS	-	-	64.28%	
PCT [91]	2021	ModelNet40	93.2%	-	-	Proposing a coordinate-based embedding module and an offset attention module
		S3DIS	-	67.7%	61.33%	
Point Transformer [92] (Zhao et al.)	2021	ModelNet40	93.7%	90.6%	-	Facilitating interactions between local feature vectors through residual transformer blocks
		S3DIS	90.2%	81.9%	73.5%	
		ShapeNet Part	-	-	86.6%	

Table 7. *Cont.*

Method	Year	Dataset	Performance			Contribution
			OA	mAcc	mIoU	
Point Transformer [93] (Engel et al.)	2021	ModelNet40	92.8%	-	-	Proposing a multihead attention network
		ShapeNet	-	-	85.9%	
MLMST [95]	2021	ModelNet10	95.5%	-	-	Proposing a multilevel multiscale Transformer
		ModelNet40	92.9%	-	-	
		ShapeNet Part	-	-	86.4%	
		S3DIS	-	-	62.9%	
DTNet [96]	2021	ModelNet40	92.9%	90.4%	-	Proposing a novel dual-point cloud Transformer architecture
		ShapeNet Part	-	-	85.6%	
Stratified Transformer [97]	2022	ShapeNet Part	-	-	86.6%	Adaptive contextual relative position encoding and point embedding effective learning of long-range contexts
		ScanNet	-	-	73.7%	
SAT [98]	2023	ScanNet	-	-	74.2%	Proposing a multigranular attention scheme and a reattention module
		S3DIS	-	78.8%	72.6%	

6. Prospects

As the focus of research in 3D computer vision, point cloud semantic segmentation is playing an increasingly prominent role in a large number of emerging industries, including smart cities, automatic navigation systems, and virtual reality. Based on the existing research, this paper summarizes the key issues and development trends and provides the following outlook on future research directions.

(1) Multimodal data processing. Point cloud semantic segmentation methods from different research perspectives are based on different data forms (e.g., 2D images, voxels, point clouds). However, the data of a single form can hardly satisfy the all-around understanding and representation of 3D scenes. To this end, Xu et al. [99] proposed a point cloud semantic perception network based on voxels and graph-structured data. The network transforms the raw point cloud into voxels, constructs an adjacency graph for spatial contexts, and encodes the representation to realize the association of local geometric features between voxels. Liu et al. [100] proposed a dual-branch network named PVCNN for parallel processing of points and voxels, in which the voxel-based feature extraction branch aggregates coarse-grained features in the neighborhood, and the point-based branch uses MLP to achieve the extraction of fine-grained features. Therefore, designing lightweight and efficient multimodal data processing networks is an innovative idea to improve the performance of point cloud semantic segmentation methods.

(2) Point cloud semantic segmentation in remote sensing. The point cloud is one of the common data carriers in the field of remote sensing. At present, although there are some point cloud datasets with large data volumes, such as SemanticKITTI, Semantic3D, and DALES for outdoor scenes, the existing data are still insufficient to satisfy the demand for semantic segmentation of super-large-scale urban scenes. For this reason, it is significant to construct high-quality and reliable spatiotemporal remote sensing datasets to support scientific research on remote sensing point cloud semantic segmentation. In recent studies, Unal et al. [101] innovatively proposed a novel strategy named Scribbles that can effectively simplify data annotation and published the first LiDAR point cloud dataset based on this strategy, ScribbleKITTI. This weak annotation approach does not need to finely annotate the boundaries of the object, but simply determines the start and end points of a line annotation, thus saving human, material, and financial resources to a great extent. Therefore, using this strategy to simplify the annotation of datasets may be the research direction and development trend of re-

mote sensing point cloud semantic segmentation in the future. In addition, due to the different focus of remote sensing and computer vision, the performance evaluation system in computer vision is not fully applicable in remote sensing. How to build a standardized and unified performance evaluation system is the focus of future research on remote sensing point cloud semantic segmentation.

(3) Weakly supervised and unsupervised learning. The performance of deep-learning-based methods relies on a large amount of data, but the existing datasets are far from satisfying the development needs. By using the weakly supervised learning strategy with only a small amount of weakly labeled data or the unsupervised learning strategy to train networks, the data hunger problem due to insufficient datasets can be largely alleviated. In this regard, Yang et al. [102] proposed an unsupervised point cloud semantic segmentation network by combining co-contrastive learning and a mutual attention sampling strategy, which deeply explores the contextual interactions between point pairs and accurately identifies points with strong cross-domain correlations through the object sampler and the background sampler, showing impressive performance on ScanObjectNN and S3DIS datasets. Xie et al. [103] designed an unsupervised pretraining strategy, PointContrast, to dynamically adjust the distance between features by comparing the matching of points before and after point cloud transformation in different views of the same scene. The method demonstrates its effectiveness in point cloud semantic segmentation and 3D target detection tasks across six different benchmarks for indoor, outdoor, and synthetic datasets, while also proving the feasibility that the learned representation can generalize across domains.

(4) Few-shot and zero-shot learning. Deep learning is a data-driven technique that relies heavily on labeled samples. Due to the limitations of small size, uneven quality, and unbalanced data volume of different categories, few-shot and zero-shot learning strategies have been developed to solve the problem of overdependence on sample data. Specifically, the few-shot learning [104,105] strategy extracts key information from sample data with only a small amount of labeled samples so that the pretrained model can generalize to categories that did not occur during training. The zero-shot learning strategy [106,107] uses a limited number of samples that have no intersection with the categories in test sets to train models and achieve the construction of cross-domain representations by learning cross-domain features. The few-shot and zero-shot learning strategies provide a new research concept for achieving point cloud classification and semantic segmentation in the absence of sample data, which is instructive.

7. Conclusions

Point cloud semantic segmentation is a popular research topic in 3D computer vision. To segment large-scale point clouds more efficiently and robustly, researchers have developed different types of methods in the past few years and achieved some significant progress. In this paper, we discuss the diversity, information richness, nonuniformity, nonstructure, and disorder of point clouds and summarize representative public datasets and mainstream evaluation metrics. Based on a broad review, we believe that the size, quality, and diversity of datasets are the key factors for training deep models. The main challenges of existing 3D datasets can be summarized as follows:

(1) Difference in sensor types and data acquisition platforms leads to certain obstacles in processing different datasets by the models.

(2) The density of point clouds in 3D space is extremely nonuniform and the datasets are commonly long-tailed, which leads to the uneven focus of the models on different object categories in scene understanding.

(3) The diversity of 3D dataset types leads to large differences in the categories and numbers of objects in each scene, which poses challenges to the cross-domain learning capability of the models.

In addition, an organized survey of point cloud semantic segmentation methods is conducted, and based on the different methods of processing the original input data, such methods are divided into rule-based methods and point-based methods. The first type of method utilizes different data processing strategies to transform the original point cloud into structured data that can be easily represented by multiviews, voxels, etc. and indirectly extracts the key features from the point cloud to achieve the purpose of 3D semantic segmentation. This type of method focuses on how to effectively address the disordered and unstructured problem of point clouds. For complex and large-scale point clouds, the second type of method outperforms the rule-based method in terms of both data processing efficiency and performance. The reason is that the methods can directly extract features from the original point cloud, which can retain the geometric structure and intrinsic features of the objects maximally. On this basis, this paper systematically reviews the research of point cloud semantic segmentation and comprehensively classifies, elaborates, and summarizes the methods proposed in recent years from a variety of perspectives. Finally, we provide an insightful discussion of the outstanding issues and identify potential research directions.

Author Contributions: Conceptualization, R.Z. and Y.W.; methodology, R.Z. and Y.W.; formal analysis, Y.W., W.J., and X.M.; investigation, Y.W. and W.J.; data curation, Y.W., W.J., and X.M.; writing—original draft preparation, Y.W.; writing—review and editing, R.Z. and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This study is undertaken with the support of the National Natural Science Foundation of China (NSFC) (Grant No. 42371466).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

Appendix A

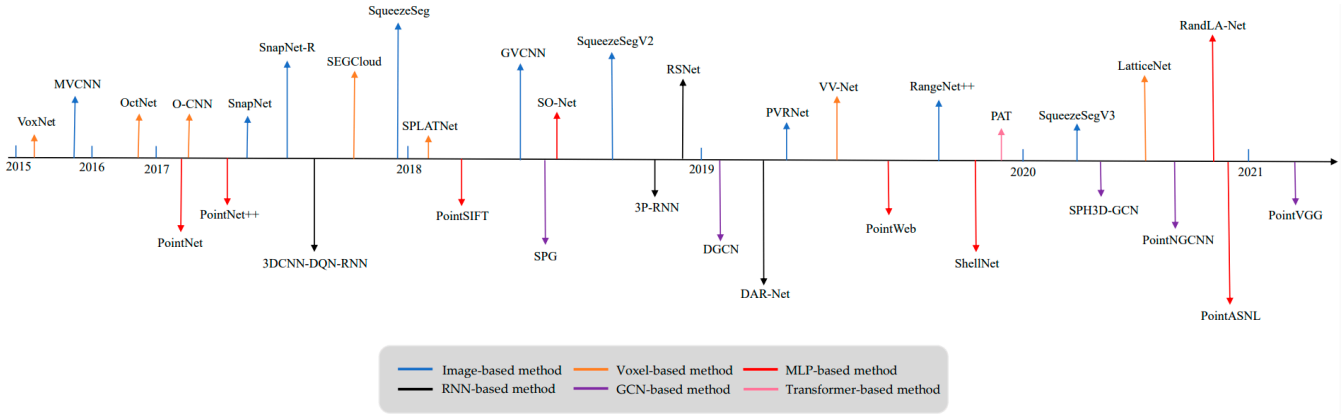


Figure A1. Timeline of deep Learning-based point cloud semantic segmentation methods. Different colors represent different types of methods. Image-based method [43–50] and [52,53], Voxel-based method [55–63], MLP-based method [64–72], RNN-based method [74–77], GCN-based method [79–83] and [86,87] Transformer-based method [91–98]. Best viewed in color.

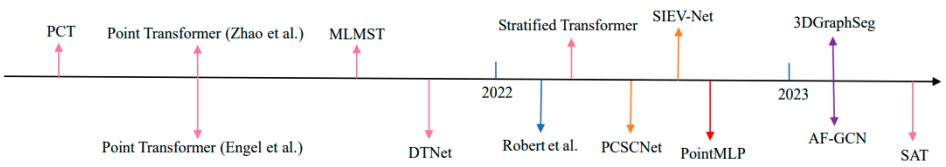


Figure A2. Timeline of deep-learning-based point cloud semantic segmentation methods (continued) [47,92,93]. Best viewed in color.

References

1. Kang, D.; Wong, A.; Lee, B.; Kim, J. Real-time semantic segmentation of 3D point cloud for autonomous driving. *Electronics* **2021**, *10*, 1960. [CrossRef]
2. Jin, Y.-H.; Hwang, I.-T.; Lee, W.-H. A mobile augmented reality system for the real-time visualization of pipes in point cloud data with a depth sensor. *Electronics* **2020**, *9*, 836. [CrossRef]
3. Wang, G.; Wang, L.; Wu, S.; Zu, S.; Song, B. Semantic Segmentation of Transmission Corridor 3D Point Clouds Based on CA-PointNet++. *Electronics* **2023**, *12*, 2829. [CrossRef]
4. Zhang, R.; Li, G.; Wunderlich, T.; Wang, L. A survey on deep learning-based precise boundary recovery of semantic segmentation for images and point clouds. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102411. [CrossRef]
5. Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* **2022**, *493*, 626–646. [CrossRef]
6. Diab, A.; Kashef, R.; Shaker, A. Deep Learning for LiDAR Point Cloud Classification in Remote Sensing. *Sensors* **2022**, *22*, 7868. [CrossRef] [PubMed]
7. Yang, S.; Hou, M.; Li, S. Three-Dimensional Point Cloud Semantic Segmentation for Cultural Heritage: A Comprehensive Review. *Remote Sens.* **2023**, *15*, 548. [CrossRef]
8. Jhaldiyal, A.; Chaudhary, N. Semantic segmentation of 3D LiDAR data using deep learning: A review of projection-based methods. *Appl. Intell.* **2023**, *53*, 6844–6855. [CrossRef]
9. Pan, Y.; Xia, Y.; Li, Y.; Yang, M.; Zhu, Q. Research on stability analysis of large karst cave structure based on multi-source point clouds modeling. *Earth Sci. Inform.* **2023**, *16*, 1637–1656. [CrossRef]
10. Tong, X.; Zhang, X.; Liu, S.; Ye, Z.; Feng, Y.; Xie, H.; Chen, L.; Zhang, F.; Han, J.; Jin, Y. Automatic Registration of Very Low Overlapping Array InSAR Point Clouds in Urban Scenes. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–25. [CrossRef]
11. Masciulli, C.; Gaeta, M.; Berardo, G.; Pantozzi, G.; Stefanini, C.A.; Mazzanti, P. ML-based characterization of PS-InSAR multi-mission point clouds for ground deformation classification. In Proceedings of the EGU General Assembly 2023, Vienna, Austria, 24–28 April 2023; p. EGU23-14546. [CrossRef]
12. Hu, L.; Tomás, R.; Tang, X.; López Vinielles, J.; Herrera, G.; Li, T.; Liu, Z. Updating Active Deformation Inventory Maps in Mining Areas by Integrating InSAR and LiDAR Datasets. *Remote Sens.* **2023**, *15*, 996. [CrossRef]
13. da Silva Ruiz, P.R.; Almeida CM de Schimalski, M.B.; Liesenberg, V.; Mitishita, E.A. Multi-approach integration of ALS and TLS point clouds for a 3-D building modeling at LoD3. *Int. J. Archit. Comput.* **2023**. [CrossRef]
14. Song, S.; Lichtenberg, S.P.; Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 567–576.
15. Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3d shapenets: A deep representation for volumetric shapes. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1912–1920.
16. Armeni, I.; Sener, O.; Zamir, A.R.; Jiang, H.; Brilakis, I.; Fischer, M.; Savarese, S. 3d semantic parsing of large-scale indoor spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1534–1543.
17. Geyer, J.; Kassahun, Y.; Mahmudi, M.; Ricou, X.; Durgesh, R.; Chung, A.S.; Hauswald, L.; Pham, V.H.; Mühlegg, M.; Dorn, S. A2d2: Audi autonomous driving dataset. *arXiv* **2020**, arXiv:2004.06320.
18. 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. *Int. J. Robot. Res.* **2018**, *37*, 545–557. [CrossRef]
19. Chang, A.X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H. Shapenet: An information-rich 3d model repository. *arXiv* **2015**, arXiv:1512.03012.
20. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5828–5839.
21. Hackel, T.; Savinov, N.; Ladicky, L.; Wegner, J.D.; Schindler, K.; Pollefeys, M. Semantic3d. net: A new large-scale point cloud classification benchmark. *arXiv* **2017**, arXiv:1704.03847.
22. Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; Gall, J. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of the Korea, 27 October–2 November 2019; pp. 9297–9307.
23. Uy, M.A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, T.; Yeung, S.-K. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of the Korea, 27 October–2 November 2019; pp. 1588–1597.
24. Yi, L.; Kim, V.G.; Ceylan, D.; Shen, I.-C.; Yan, M.; Su, H.; Lu, C.; Huang, Q.; Sheffer, A.; Guibas, L. A scalable active framework for region annotation in 3d shape collections. *ACM Trans. Graph.* **2016**, *35*, 1–12. [CrossRef]
25. Wang, C.; Dai, Y.; Elsheimy, N.; Wen, C.; Retscher, G.; Kang, Z.; Lingua, A. ISPRS Benchmark on Multisensory Indoor Mapping and Positioning. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2020**, *5*, 117–123. [CrossRef]
26. Matrone, F.; Lingua, A.; Pierdicca, R.; Malinverni, E.; Paolanti, M.; Grilli, E.; Remondino, F.; Murtiyoso, A.; Landes, T. A benchmark for large-scale heritage point cloud semantic segmentation. In Proceedings of the XXIV ISPRS Congress, Nice, France, 31 August–2 September 2020; pp. 1419–1426.
27. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.

28. Serna, A.; Marcotegui, B.; Goulette, F.; Deschaud, J.-E. Paris-rue-Madame database: A 3D mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods. In Proceedings of the 4th International Conference on Pattern Recognition, Applications and Methods ICPRAM 2014, Lisbon, Portugal, 29 July 2014.
29. Huang, X.; Cheng, X.; Geng, Q.; Cao, B.; Zhou, D.; Wang, P.; Lin, Y.; Yang, R. The apolloscape dataset for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 954–960.
30. Tan, W.; Qin, N.; Ma, L.; Li, Y.; Du, J.; Cai, G.; Yang, K.; Li, J. Toronto-3D: A large-scale mobile lidar dataset for semantic segmentation of urban roadways. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 202–203.
31. Pan, Y.; Gao, B.; Mei, J.; Geng, S.; Li, C.; Zhao, H. Semanticposs: A point cloud dataset with large quantity of dynamic instances. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 687–693.
32. Dong, Z.; Liang, F.; Yang, B.; Xu, Y.; Zang, Y.; Li, J.; Wang, Y.; Dai, W.; Fan, H.; Hyyppä, J. Registration of large-scale terrestrial laser scanner point clouds: A review and benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *163*, 327–342. [CrossRef]
33. 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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
34. Xiao, P.; Shao, Z.; Hao, S.; Zhang, Z.; Chai, X.; Jiao, J.; Li, Z.; Wu, J.; Sun, K.; Jiang, K. PandaSet: Advanced Sensor Suite Dataset for Autonomous Driving. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 3095–3101.
35. Fong, W.K.; Mohan, R.; Hurtado, J.V.; Zhou, L.; Caesar, H.; Beijbom, O.; Valada, A. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robot. Autom. Lett.* **2022**, *7*, 3795–3802. [CrossRef]
36. Zheng, L.; Ma, Z.; Zhu, X.; Tan, B.; Li, S.; Long, K.; Sun, W.; Chen, S.; Zhang, L.; Wan, M. TJ4DRadSet: A 4D Radar Dataset for Autonomous Driving. *arXiv* **2022**, arXiv:2204.13483.
37. Varney, N.; Asari, V.K.; Graehling, Q. Dales: A large-scale aerial lidar data set for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 186–187.
38. Ye, Z.; Xu, Y.; Huang, R.; Tong, X.; Li, X.; Liu, X.; Luan, K.; Hoegner, L.; Stilla, U. Lasdu: A large-scale aerial lidar dataset for semantic labeling in dense urban areas. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 450. [CrossRef]
39. Hu, Q.; Yang, B.; Khalid, S.; Xiao, W.; Trigoni, N.; Markham, A. SensatUrban: Learning Semantics from Urban-Scale Photogrammetric Point Clouds. *Int. J. Comput. Vis.* **2022**, *130*, 316–343. [CrossRef]
40. Jiang, X.Y.; Meier, U.; Bunke, H. Fast range image segmentation using high-level segmentation primitives. In Proceedings of the Third IEEE Workshop on Applications of Computer Vision: WACV'96, Sarasota, FL, USA, 2–4 December 1996; pp. 83–88.
41. Besl, P.J.; Jain, R.C. Segmentation through variable-order surface fitting. *IEEE Trans. Pattern Anal. Mach. Intell.* **1988**, *10*, 167–192. [CrossRef]
42. Rabbani, T.; Van Den Heuvel, F.; Vosselmann, G. Segmentation of point clouds using smoothness constraint. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2006**, *36*, 248–253.
43. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 945–953.
44. Feng, Y.; Zhang, Z.; Zhao, X.; Ji, R.; Gao, Y. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 264–272.
45. You, H.; Feng, Y.; Zhao, X.; Zou, C.; Ji, R.; Gao, Y. PVRNet: Point-view relation neural network for 3D shape recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 29–31 January 2019; pp. 9119–9126.
46. Milioto, A.; Vizzo, I.; Behley, J.; Stachniss, C. Rangenet++: Fast and accurate lidar semantic segmentation. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 4213–4220.
47. Robert, D.; Vallet, B.; Landrieu, L. Learning Multi-View Aggregation in the Wild for Large-Scale 3D Semantic Segmentation. *arXiv* **2022**, arXiv:2204.07548.
48. Boulch, A.; Guerry, J.; Le Saux, B.; Audebert, N. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Comput. Graph.* **2018**, *71*, 189–198. [CrossRef]
49. Guerry, J.; Boulch, A.; Le Saux, B.; Moras, J.; Plyer, A.; Filliat, D. Snapnet-r: Consistent 3d multi-view semantic labeling for robotics. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 669–678.
50. 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 Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 1887–1893.
51. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
52. 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 Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 4376–4382.
53. Xu, C.; Wu, B.; Wang, Z.; Zhan, W.; Vajda, P.; Keutzer, K.; Tomizuka, M. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 1–19.
54. Yang, Y.; Wu, X.; He, T.; Zhao, H.; Liu, X. SAM3D: Segment Anything in 3D Scenes. *arXiv* **2023**, arXiv:2306.03908.

55. Maturana, D.; Scherer, S. Voxnet: A 3d convolutional neural network for real-time object recognition. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–3 October 2015; pp. 922–928.
56. Su, H.; Jampani, V.; Sun, D.; Maji, S.; Kalogerakis, E.; Yang, M.-H.; Kautz, J. Splatnet: Sparse lattice networks for point cloud processing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2530–2539.
57. Rosu, R.A.; Schütt, P.; Quenzel, J.; Behnke, S. Latticenet: Fast point cloud segmentation using permutohedral lattices. *arXiv* **2019**, arXiv:1912.05905.
58. Tchapmi, L.; Choy, C.; Armeni, I.; Gwak, J.; Savarese, S. Segcloud: Semantic segmentation of 3d point clouds. In Proceedings of the 2017 international conference on 3D vision (3DV), Qingdao, China, 10–12 October 2017; pp. 537–547.
59. Riegler, G.; Osman Ulusoy, A.; Geiger, A. Octnet: Learning deep 3d representations at high resolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3577–3586.
60. Wang, P.-S.; Liu, Y.; Guo, Y.-X.; Sun, C.-Y.; Tong, X. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Trans. Graph.* **2017**, *36*, 1–11. [CrossRef]
61. Meng, H.-Y.; Gao, L.; Lai, Y.-K.; Manocha, D. Vv-net: Voxel vae net with group convolutions for point cloud segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of the Korea, 27 October–2 November 2019; pp. 8500–8508.
62. Park, J.; Kim, C.; Kim, S.; Jo, K. PCSCNet: Fast 3D semantic segmentation of LiDAR point cloud for autonomous car using point convolution and sparse convolution network. *Expert Syst. Appl.* **2023**, *212*, 118815. [CrossRef]
63. Yu, C.; Lei, J.; Peng, B.; Shen, H.; Huang, Q. SIEV-Net: A structure-information enhanced voxel network for 3D object detection from LiDAR point clouds. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
64. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
65. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5105–5114.
66. Jiang, M.; Wu, Y.; Zhao, T.; Zhao, Z.; Lu, C. Pointsift: A sift-like network module for 3d point cloud semantic segmentation. *arXiv* **2018**, arXiv:1807.00652.
67. Zhao, H.; Jiang, L.; Fu, C.-W.; Jia, J. Pointweb: Enhancing local neighborhood features for point cloud processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5565–5573.
68. Li, J.; Chen, B.M.; Lee, G.H. So-net: Self-organizing network for point cloud analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9397–9406.
69. Zhang, Z.; Hua, B.-S.; Yeung, S.-K. Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of the Korea, 27 October–2 November 2019; pp. 1607–1616.
70. Yan, X.; Zheng, C.; Li, Z.; Wang, S.; Cui, S. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5589–5598.
71. 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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11108–11117.
72. Ma, X.; Qin, C.; You, H.; Ran, H.; Fu, Y. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv* **2022**, arXiv:2202.07123.
73. Fan, H.; Yang, Y. PointRNN: Point recurrent neural network for moving point cloud processing. *arXiv* **2019**, arXiv:1910.08287.
74. Ye, X.; Li, J.; Huang, H.; Du, L.; Zhang, X. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 403–417.
75. Huang, Q.; Wang, W.; Neumann, U. Recurrent slice networks for 3d segmentation of point clouds. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2626–2635.
76. Zhao, Z.; Liu, M.; Ramani, K. DAR-Net: Dynamic aggregation network for semantic scene segmentation. *arXiv* **2019**, arXiv:1907.12022.
77. Liu, F.; Li, S.; Zhang, L.; Zhou, C.; Ye, R.; Wang, Y.; Lu, J. 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, Venice, Italy, 22–29 October 2017; pp. 5678–5687.
78. Simonovsky, M.; Komodakis, N. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3693–3702.
79. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic graph cnn for learning on point clouds. *Acm Trans. Graph.* **2019**, *38*, 1–12. [CrossRef]
80. Lei, H.; Akhtar, N.; Mian, A. Spherical kernel for efficient graph convolution on 3d point clouds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3664–3680. [CrossRef] [PubMed]
81. Lu, Q.; Chen, C.; Xie, W.; Luo, Y. PointNGCNN: Deep convolutional networks on 3D point clouds with neighborhood graph filters. *Comput. Graph.* **2020**, *86*, 42–51. [CrossRef]

82. Li, R.; Zhang, Y.; Niu, D.; Yang, G.; Zafar, N.; Zhang, C.; Zhao, X. PointVGG: Graph convolutional network with progressive aggregating features on point clouds. *Neurocomputing* **2021**, *429*, 187–198. [CrossRef]
83. Zhang, N.; Pan, Z.; Li, T.H.; Gao, W.; Li, G. Improving Graph Representation for Point Cloud Segmentation via Attentive Filtering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1244–1254.
84. Zhang, R.; Li, G.; Wiedemann, W.; Holst, C. KdO-Net: Towards Improving the Efficiency of Deep Convolutional Neural Networks Applied in the 3D Pairwise Point Feature Matching. *Remote Sens.* **2022**, *14*, 2883. [CrossRef]
85. Li, R.; Wang, S.; Zhu, F.; Huang, J. Adaptive graph convolutional neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–3 February 2018.
86. Landrieu, L.; Simonovsky, M. Large-scale point cloud semantic segmentation with superpoint graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4558–4567.
87. Geng, Y.; Wang, Z.; Jia, L.; Qin, Y.; Chai, Y.; Liu, K.; Tong, L. 3DGraphSeg: A Unified Graph Representation-Based Point Cloud Segmentation Framework for Full-Range Highspeed Railway Environments. *IEEE Trans. Ind. Inform.* **2023**. [CrossRef]
88. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
89. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.
90. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
91. Guo, M.-H.; Cai, J.-X.; Liu, Z.-N.; Mu, T.-J.; Martin, R.R.; Hu, S.-M. Pct: Point cloud transformer. *Comput. Vis. Media* **2021**, *7*, 187–199. [CrossRef]
92. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.; Koltun, V. Point transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 16259–16268.
93. Engel, N.; Belagiannis, V.; Dietmayer, K. Point transformer. *IEEE Access* **2021**, *9*, 134826–134840. [CrossRef]
94. Yang, J.; Zhang, Q.; Ni, B.; Li, L.; Liu, J.; Zhou, M.; Tian, Q. Modeling point clouds with self-attention and gumbel subset sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3323–3332.
95. Zhong, Q.; Han, X.-F. Point cloud learning with transformer. *arXiv* **2021**, arXiv:2104.13636.
96. Han, X.-F.; Jin, Y.-F.; Cheng, H.-X.; Xiao, G.-Q. Dual Transformer for Point Cloud Analysis. *arXiv* **2021**, arXiv:2104.13044. [CrossRef]
97. Lai, X.; Liu, J.; Jiang, L.; Wang, L.; Zhao, H.; Liu, S.; Qi, X.; Jia, J. Stratified Transformer for 3D Point Cloud Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8500–8509.
98. Zhou, J.; Xiong, Y.; Chiu, C.; Liu, F.; Gong, X. SAT: Size-Aware Transformer for 3D Point Cloud Semantic Segmentation. *arXiv* **2023**, arXiv:2301.06869.
99. Xu, Y.; Hoegner, L.; Tuttas, S.; Stilla, U. Voxel-and graph-based point cloud segmentation of 3d scenes using perceptual grouping laws. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *4*, 43–50. [CrossRef]
100. Liu, Z.; Tang, H.; Lin, Y.; Han, S. Point-Voxel CNN for Efficient 3D Deep Learning. *arXiv* **2019**, arXiv:1907.03739. [CrossRef]
101. Unal, O.; Dai, D.; Van Gool, L. Scribble-Supervised LiDAR Semantic Segmentation. *arXiv* **2022**, arXiv:2203.08537.
102. Yang, C.-K.; Chuang, Y.-Y.; Lin, Y.-Y. Unsupervised Point Cloud Object Co-segmentation by Co-contrastive Learning and Mutual Attention Sampling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7335–7344.
103. Xie, S.; Gu, J.; Guo, D.; Qi, C.R.; Guibas, L.; Litany, O. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 574–591.
104. Liu, M.; Zhu, Y.; Cai, H.; Han, S.; Ling, Z.; Porikli, F.; Su, H. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 21736–21746.
105. Sharma, C.; Kaul, M. Self-supervised few-shot learning on point clouds. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 7212–7221.
106. He, S.; Jiang, X.; Jiang, W.; Ding, H. Prototype Adaption and Projection for Few- and Zero-Shot 3D Point Cloud Semantic Segmentation. *IEEE Trans. Image Process.* **2023**, *32*, 3199–3211. [CrossRef]
107. Abdelreheem, A.; Skorokhodov, I.; Ovsjanikov, M.; Wonka, P. SATR: Zero-Shot Semantic Segmentation of 3D Shapes. *arXiv* **2023**, arXiv:2304.04909.