

REVIEW

Open Access



A review of point cloud segmentation for understanding 3D indoor scenes

Yuliang Sun¹ , Xudong Zhang¹  and Yongwei Miao^{2*} 

Abstract

Point cloud segmentation is an essential task in three-dimensional (3D) vision and intelligence. It is a critical step in understanding 3D scenes with a variety of applications. With the rapid development of 3D scanning devices, point cloud data have become increasingly available to researchers. Recent advances in deep learning are driving advances in point cloud segmentation research and applications. This paper presents a comprehensive review of recent progress in point cloud segmentation for understanding 3D indoor scenes. First, we present public point cloud datasets, which are the foundation for research in this area. Second, we briefly review previous segmentation methods based on geometry. Then, learning-based segmentation methods with multi-views and voxels are presented. Next, we provide an overview of learning-based point cloud segmentation, ranging from semantic segmentation to instance segmentation. Based on the annotation level, these methods are categorized into fully supervised and weakly supervised methods. Finally, we discuss open challenges and research directions in the future.

Keywords: Point clouds, Scene understanding, Deep learning, Semantic segmentation, Instance segmentation

1 Introduction

Understanding indoor scenes is one of the essential tasks for computer vision and intelligence. The rapid development of depth sensors and three-dimensional (3D) scanners, such as RGB-D cameras and LiDAR, has increased interest in 3D indoor scene comprehension for a variety of applications, such as robotics [1], navigation [2] and augmented/virtual reality [3, 4]. The objective of 3D indoor scene understanding is to discern the geometric and semantic context of each interior scene component. There are a variety of 3D data formats, including depth images, meshes, voxels, and point clouds. Among them, point clouds are the most common non-discretized data representation in 3D applications and can be acquired directly by 3D scanners or reconstructed from stereo or multi-view images.

Point cloud segmentation, which attempts to decompose indoor scenes into meaningful parts and label each point, is a fundamental and indispensable step in understanding 3D indoor scenes. Point clouds provide the original spatial information, making them the preferred data format for segmenting indoor scenes. Segmentation of indoor scene point clouds can be divided into semantic segmentation and instance segmentation. Semantic segmentation assigns each point with a scene-level object category label. Instance segmentation is more difficult and requires individual object identification and localization. Unlike outdoor point cloud segmentation, which addresses dynamic objects, indoor point cloud segmentation commonly handles cluttered man-made objects with regularly designed shapes. Indoor point cloud data are usually captured by consumer-level sensors with short ranges, while outdoor point clouds are commonly collected by LiDAR. Indoor point cloud segmentation faces several challenges. First, point cloud data are typically large and voluminous, with varying qualities from different sensors. This makes it difficult to efficiently process and accurately annotate point cloud data. Second, indoor scenes are typically cluttered

*Correspondence: ywmiao@hznu.edu.cn

²School of Information Science and Technology, Hangzhou Normal University, Hangzhou, 311121, China

Full list of author information is available at the end of the article

with severe occlusions. It is challenging to accurately segment objects when they are hidden or close together. Third, unlike regular data structures in two-dimensional (2D) images, point cloud data are sparse and unorganized, making it difficult to apply sophisticated 2D segmentation methods directly to 3D point clouds. Moreover, annotating 3D data is time-consuming and labor-intensive, limiting the ability of fully supervised learning. Existing indoor point cloud datasets are limited and suffer from long-tailed distributions.

Much effort has been devoted to the task of point cloud segmentation. Traditional geometry-based solutions for point cloud segmentation mainly include clustering-based, model-based, and graph-based methods [5]. The majority of these methods rely on hand-crafted features with heuristic geometric constraints. Deep learning has made significant progress in 2D vision [6–8], leading to advances in point cloud segmentation. In recent years, point cloud based deep neural networks [9] have demonstrated the ability to extract more powerful features and provide more reliable geometric cues for better understanding 3D scenes. Learning from 3D data has become a reality with the availability of public datasets such as ShapeNet, ModelNet, PartNet, ScanNet, Semantic3D, and KITTI. Recently, weakly supervised learning for point cloud segmentation has become a popular research topic, because it attempts to learn features from limited annotated data.

This paper provides a comprehensive review of point cloud segmentation for indoor 3D scene understanding, especially methods based on deep learning. We will introduce the primary datasets and methods used for indoor scene point cloud segmentation, analyze the current research trends in this area, and discuss future directions for development. The structure of this paper is as follows. Section 2 begins by introducing 3D indoor datasets that are used for understanding 3D scenes. Section 3 presents a brief review of geometry-based point cloud segmentation methods. Section 4 reviews indirect learning approaches with structured data. Section 5 provides a comprehensive survey of existing point cloud based learning frameworks employed for 3D scene segmentation. Section 6 introduces recent learning-based segmentation methods with multi-modal data. Section 7 summarizes the performance of indoor point cloud segmentation using different methods. Section 8 discusses open questions and future research directions. Section 9 concludes the paper.

2 3D indoor scene point cloud datasets

The emergence of 3D datasets has led to the development of deep learning-based segmentation methods, which play a crucial role in advancing the field and promoting progress in research and applications. Public benchmarks have proven to be highly effective in facilitating framework evaluation and comparison. By providing real-world data

with ground truth annotations, these benchmarks offer a foundation for researchers to test their algorithms and enable fair comparisons between different approaches. The two most commonly used 3D indoor scene point cloud datasets are ScanNet [10] and S3DIS [11].

ScanNet. ScanNet [10] is an RGB-D video dataset encompassing more than 2.5 million views across more than 1500 scans. This dataset is captured by RGB-D cameras and extensively annotated with essential information such as 3D camera poses, surface reconstructions, and instance-level semantic segmentations. This dataset has led to remarkable advancements in state-of-the-art performance across various 3D scene understanding tasks, including object detection, semantic segmentation, instance segmentation, and computer-aided design (CAD) model retrieval. ScanNet v2, the modified released version, has meticulously gathered 1,513 scans that have been annotated with impressive surface coverage. In the semantic segmentation task, the V2 version is labeled with annotations for 20 classes of 3D voxelized objects. Each of these classes corresponds to a specific furniture category or room layout, allowing for a more granular understanding and analysis of the captured indoor scenes. This makes ScanNetV2 one of the most active online evaluation datasets tailored for indoor scene semantic segmentation. Apart from semantic segmentation benchmarks, ScanNetV2 also provides benchmarks for instance segmentation and scene type classification.

ScanNet200. ScanNet200 [12] was developed on the basis of ScanNetV2 to overcome the limited set of 20 class labels. It significantly expands the number of classes to 200, representing an order of magnitude increase compared to the previous version. This annotation enables a better capture and understanding of real-world indoor scenes with a more diverse range of objects. This new benchmark allows for a more comprehensive analysis of performance across different object class distributions by splitting the 200 classes into three sets. Specifically, the “head” set comprises 66 categories with the highest frequency, the “common” set consists of 68 categories with less frequency, and the “tail” set contains the remaining categories.

S3DIS. The Stanford large-scale 3D indoor spaces dataset [11], known as S3DIS, acquired through the Matterport scanner, is another highly popular dataset that has been extensively employed in point cloud segmentation. This dataset comprises 272 room scenes divided into 6 distinct areas. Each point within the scene is assigned a semantic label corresponding to one of the 13 pre-defined categories, such as walls, tables, chairs, cabinets, and others. This dataset is specifically curated for large-scale indoor semantic segmentation.

Cornell RGBD. This dataset [13] provides 52 labeled indoor scenes comprising point clouds with RGB values. It consists of 24 labeled office scenes and 28 labeled home

scenes. The point cloud data are generated from the original RGB-D images via the RGBD-SLAM method. This dataset contains approximately 550 views with 2495 labeled segments across 27 object classes, providing valuable resources for previous research and development in indoor scene understanding.

Washington RGB-D dataset. This dataset [14] includes 14 indoor scene point clouds, which are obtained via RGB-D image registration and stitching. It provides annotations of 9 semantic category labels, such as sofas, teacups, and hats.

3 Geometry-based segmentation

Geometry-based solutions for understanding indoor scenes can be classified as clustering or region growing, or model fitting based methods. By incorporating heuristic geometric constraints, most of these methods use hand-crafted features. The intuition behind geometry-based methods is that man-made environments normally consist of many geometric structures.

Clustering or region growing. These approaches assume that points in close proximity to each other are more likely to belong to the same object or surface. By considering the geometric properties of these neighboring points, such as spatial coordinates and surface normals, these methods can identify regions that share similarities in these properties. Mattausch et al. [15] proposed a method for segmenting indoor scenes by identifying repeated objects from multi-room indoor scanning data. To represent the indoor scenes, they employed a collection of nearly planar patches. These patches were clustered based on a patch similarity matrix, which was constructed using shape geometrical descriptors. Using this approach, the researchers aimed to effectively segment indoor scenes by exploiting the inherent repeated object structures. Hu et al. [16] partitioned point clouds into surface patches using the dynamic region growing method to generate initial segmentation. By leveraging this intermediate data representation, the model can better account for shape variations and enhance its ability to classify objects.

Model fitting. Model fitting is proposed as a more efficient and robust strategy, particularly in scenarios where noise and outliers are present. Nan et al. [17] introduced a search-classify pipeline for scene modeling, utilizing pre-trained object categories to aid in the process. Similarly, Li et al. [18] proposed an object-retrieved approach that replaces scanned data with objects sourced from 3D shape databases. In another approach, Shi et al. [19] trained classifiers for both objects and object groups, which allows the decomposition of indoor sub-scenes. These methods rely primarily on the availability and diversity of current CAD datasets, which limits their effectiveness. Alternatively, another strategy involves employing primitive-based approaches, where indoor scenes are decomposed into a

collection of geometric primitives. By utilizing this strategy, researchers aim to capture the essential geometric information of a scene without being overly reliant on extensive CAD datasets. These primitive-based approaches offer an alternative means of scene decomposition. The most widely used primitive fitting method is random sample consensus (RANSAC) [20]. Monszpart et al. [21] processed large-scale indoor point clouds via RANSAC-based plane fitting. Sun et al. [22] developed a graph-cut segmentation method to group primitives found in indoor sub-scenes. The primitives, such as plane and cylinder, were extracted by RANSAC and oriented by PCA. Yu et al. [23] further used a patch relation classifier to group planar patches and achieve instance segmentation.

These geometry-based methods cannot directly assign object classes or instance labels to each point. Postprocessing is normally required to produce the final segmentation. Recently, several approaches have used geometry-based segmentation as the pre-processing step and generated mid-level scene representation as the inputs of deep learning frameworks. For instance, Huang et al. [24] clustered on-surface voxels to provide a compact representation of 3D scenes. Landrieu and Simonovsky [25] partitioned the scan data into superpoints, which are geometrically homogeneous elements. Cheng et al. [26] encoded a superpoint-level representation with non-local operation at the neighborhood level. Deng et al. [27] proposed an iterative algorithm to generate superpoints by combining geometry-based and color-based region growing methods. Similarly, geometry-based superpoints have been proven to leverage large-scale point clouds and act as priors in weakly supervised learning [28–31]. We will review these deep learning networks in detail in the following sections.

4 Learning-based segmentation with structured data

Unlike geometry-based methods that use hand-crafted features, learning-based approaches automatically extract latent features. However, point cloud feature learning remains difficult due to its unstructured and unordered data format. Since 3D scenes can be represented in several forms, such as multi-view images and voxels, it is natural to transform point clouds into structured data formats.

4.1 Image-based methods

The success and evolution of convolutional neural networks (CNNs) in the image domain has influenced the development of point cloud feature learning. It is common to transform 3D data into a structured multi-view representation [32, 33]. The idea is to use virtual cameras to capture scene point clouds from different angles, resulting in multi-view RGB images along with corresponding depth images. CNN is then applied to perform feature

extraction on the RGB images, and the results of downstream tasks are projected back into the 3D space. For instance, the MVCNN [32] virtually scans a 3D object from 12 angles to obtain rendered images, and features are extracted from each image via the CNNs. These features are fused into a descriptor, which is used for object classification. SnapNet [33] also applies CNNs to multi-view images from the point clouds. Unlike MVCNN, SnapNet selects a set of appropriate snapshots of the point cloud to generate an RGB image and corresponding depth map. CNN is employed to optimize the RGB-D inputs and produce pixel labels that are back-projected to corresponding 3D points. However, these networks have some limitations for segmenting indoor scenes. Semantic segmentation based on multiple views has drawbacks, such as the need to choose viewpoints and the number of scans. Moreover, the process of projection and back-projection inevitably leads to some loss of information, especially structural features.

4.2 Voxel-based methods

Converting point clouds into voxel-based representations is another approach for addressing the challenges of regularizing 3D data [34–37]. VoxNet [34] generates 3D bounding boxes on point cloud segments and transforms them into volumetric grids to represent spatial occupancy. A 3D CNN is used to predict labels directly from the occupancy grid. SEGCloud [36] utilizes a voxel-based CNN to segment indoor scenes. The idea is to pre-process the input point cloud by dividing the indoor scene into voxels. A 3D fully convolutional network is then employed to generate voxel labels. These voxel labels are then interpolated using trilinear interpolation to assign voxel labels to the corresponding points. VV-Net [35] utilizes a radial basis function-based variational autoencoder network for point cloud processing. Unlike binary voxel-based representations, VV-Net provides richer representations of point clouds. The voxelization of scanned point clouds faces certain challenges. There is a risk of losing fidelity and fine-grained details at low resolutions, while the computational and memory requirements become excessively demanding at high resolutions. Although there have been efforts to mitigate these problems, such as reducing memory consumption and computation [38, 39], voxel-based representations still struggle to handle large-scale scene segmentation in general. Recently, OctFormer [40] attempted to partition input point clouds into local windows using octrees, and used sparse octree attention to enhance segmentation performance.

5 Learning-based segmentation with point clouds

Recent studies have explored the direct application of deep learning techniques, as an alternative to multi-view and voxel-based approaches, to raw scanned point clouds.

While point cloud data can be obtained directly from scanning devices, their irregular data format presents challenges to traditional CNNs. To address this issue, PointNet [41] has emerged as a pioneering approach in point-based learning. One limitation of PointNet as a benchmark for point cloud learning is that it does not exploit the structural information within local neighborhoods of points. In response, subsequent studies have made advances by enhancing the use of sampling methods and feature extraction techniques. For instance, some works have improved the sampling method by incorporating farthest point sampling [42] or random sampling [43]. These modifications aim to enhance both the feature extraction capabilities and computational efficiency of the network. These advances built upon the foundation of PointNet address its limitations by incorporating structural information from local neighboring points and refining the feature extraction process. Some researchers choose an alternative approach to PointNet and design particular convolution operations on point clouds [44–49]. As a result, these architectures offer improved performance and capabilities for downstream 3D scene understanding tasks such as point cloud semantic segmentation and point cloud instance segmentation. A brief timeline of learning-based point cloud segmentation methods is shown in Fig. 1. These methods include SGPN [50], PAT [51], PointWeb [52], GSPN [53], ASIS [54], KP-Cov [55], 3D-BoNet [56], 3D-MPA [57], PointGroup [58], MPRM [59], PGCNet [60], WSSPK [61], DyCo3D [62], PSD [63], SSTNet [64], Stratified Transformer [65], HybridCR [66], SoftGroup [67], SegGroup [68], SQN [69], 3D-WSIS [70], and Mask3D [71].

5.1 Point cloud semantic segmentation

Point cloud semantic segmentation, which is a fundamental task in 3D indoor scene understanding, aims to partition a scene into multiple subsets. Based on the semantic meanings of the individual points, our objective is to assign each point in the scene to a specific category label. Semantic segmentation methods can be categorized according to the amount of monitoring information they rely on. Depending on the availability of annotated data, these methods can be categorized into fully supervised methods and weakly supervised methods.

5.1.1 Fully supervised semantic segmentation

Deep learning-based point cloud semantic segmentation requires large-scale data for training and typically relies on dense annotations. Current fully annotated public datasets make fully supervised point cloud learning possible.

PointNet. Qi et al. [41] introduced the PointNet network architecture. This network comprises three key components: the multi-layer perceptron (MLP) module, the max pooling structure, and the feature fusion structure. The MLP module enables the extraction of point cloud features through weight sharing. The max pooling structure,

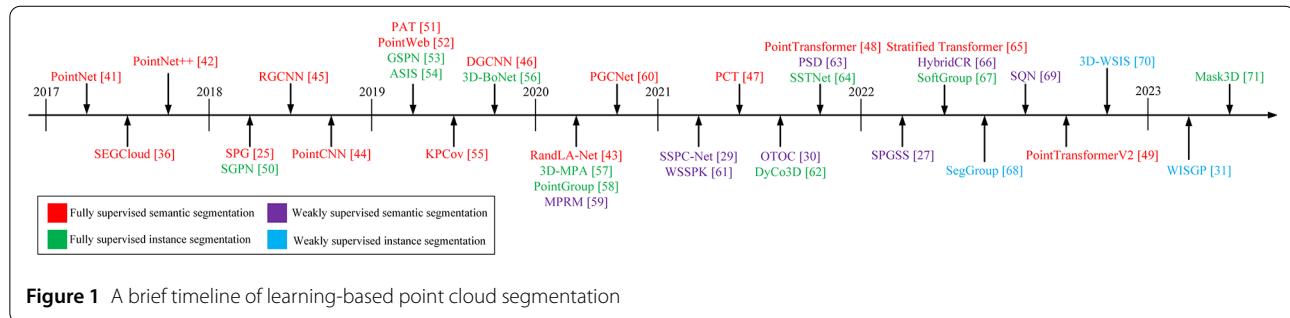


Figure 1 A brief timeline of learning-based point cloud segmentation

which employs a symmetric function, selects the maximum feature value within a group of points and serves as the global feature representation. This design addresses the problem of irregularity in the data. The feature fusion structure combines the local features and the global features obtained from the maximum pooling operation. These merged features are utilized as input, and the MLP predicts labels for each point. Moreover, PointNet incorporates the T-Net structure, which facilitates the learning of an efficient rotation matrix. PointNet has demonstrated effectiveness in tasks such as semantic segmentation and object classification, making it a fundamental network architecture in this area.

PointNet++. PointNet++ [42] introduces a set of abstraction structures consisting of sampling layers, grouping layers, and PointNet layers. This hierarchical design enables the extraction of multi-scale features from point clouds. By stacking multiple layers of this feature extraction structure, PointNet++ can be applied for tasks such as point cloud classification and segmentation.

PointCNN. PointCNN [44] transforms the input points into a latent representation. This transformation, known as x-conv, is implemented using MLPs. This transformation allows for the application of traditional convolution, which is effective in capturing local and global patterns in regular data domains.

GCN-based methods. Recent studies have explored the application of a graph convolutional network (GCN) to point clouds, recognizing that the points and their neighboring points can form a graph structure [25, 46, 72]. The objective is to extract local geometric structure information while preserving permutation invariance. This is achieved by constructing a spatial or spectral adjacency graph using the features of vertices and their neighbors. DGCNN [46] employs MLPs to aggregate edge features, which consist of nodes and their spatial neighbors. The features of the nodes are then updated based on the edge features. RGCNN [45] considers the features of data points in a point cloud as graph signals and uses spectral-based graph convolution for point cloud classification and segmentation. The spectral-based graph convolution operation is defined using the approximation of the Chebyshev polynomial. Furthermore, the Laplacian matrix of the

graph is updated at each layer of the network based on the learned depth features. This allows for the extraction of local structural information while accounting for the unordered nature of the data. DGCNN and RGCNN demonstrate different approaches to the use of GCN. DGCNN focuses on edge feature aggregation and node feature updates, while RGCNN uses spectral-based graph convolution and updates the Laplacian matrix based on learned depth features. SPG [25] is a deep learning framework specifically designed for the task of semantic segmentation in large-scale point clouds with millions of points. The framework introduces the concept of a superpoint graph (SPG), which effectively captures the inherent organization of 3D point clouds. By dividing the scanned scene into geometrically homogeneous elements, SPG provides a compact representation that captures the contextual relationships between different object parts within the point cloud. Leveraging this rich representation, GCN is employed to learn and infer semantic segmentation labels. The combination of the SPG structure and GCN enables the capture of contextual relationships, resulting in accurate semantic segmentation of complex and voluminous point cloud data. PointWeb [52] designs an adaptive feature extraction module to find the interaction between densely connected neighbors. Unlike most point-based deep learning methods, PGCNet [60] incorporates geometric information as a prior and uses surface patches for data representation. The idea behind this method is that man-made objects can be decomposed into a set of geometric primitives. The PGCNet framework first extracts surface patches from indoor scene point clouds using the region growing method. With surface patches and their geometric features as input, a GCN-based network is designed to explore patch features and contextual information. Specifically, a dynamic graph U-Net module, which employs dynamic edge convolution, aggregates hierarchical feature embeddings. Taking advantage of the surface patch representation, PGCNet can achieve competitive semantic segmentation performance with much less training.

Transformer-based methods. The Transformer technique has revolutionized natural language processing

(NLP) and 2D vision [73, 74], inspiring the application of attention-based networks in 3D space. PCT [47] extracts point cloud features through the attention mechanism. It solves the unordered problem by merging the spatial position encoding and the input embedding to represent each point. PAT [51] uses a parameter-efficient group self-attention operation and Gumbel–Softmax-based sampling to replace multi-head self-attention and furthest point sampling. Point Transformer [48] directly incorporates local attention between each point and its neighboring points, effectively addressing the memory cost problem. Point Transformer V2 [49] further improves the previous version by replacing the original attention with group vector attention with grouped weight encoding. Stratified Transformer [65] improves long-range context capture by stratified sampling. Refer to Ref. [48] for an illustration of the Transformer-based structure. All these Transformer-based networks can serve as the powerful backbone for various point cloud understanding tasks.

In recent years, Transformer-based backbones have proven to be more effective in exploiting features than other structures, while cost of computing has increased. The identification of efficient and powerful learning networks for point cloud semantic segmentation is worthy of further exploration.

5.1.2 Weakly supervised semantic segmentation

Despite tremendous progress, there are still limitations to the widespread adoption of fully supervised semantic segmentation methods. Extensive and precise annotations are required for fully supervised training, while current point cloud data are still scarce and difficult to annotate. To cope with limited annotated data, researchers have explored weakly supervised learning for semantic segmentation.

One strategy is to utilize only a small fraction of labeled points. Xu and Lee [75] proposed a weakly supervised network for semantic point cloud segmentation. This is achieved by three carefully-designed branches. The Siamese branch enhances the training samples by encouraging consistency between the original predictions and corresponding augmented predictions. The inexact branch suppresses the activation of negative categories for any given point. Using spatial and color constraints, the smooth branch ensures that spatially connected points with similar colors have the same labels. SQN [69] encodes a set of hierarchical latent representations and retrieves subsets according to spatial positions. These representations are fed into MLPs to predict semantic labels. DAT [76] incorporates dual adaptive transformations using an adversarial strategy to leverage local consistency and structural smoothness. These methods require adaptive and high-quality sampling, which is difficult when the input data scales vary.

Another line of methods is to group input point clouds into sub-clouds or superpoints. Based on a classification network trained with sub-cloud level labels, Wei et al. [59] presented a multi-path region mining network to produce point-level pseudo labels for fully supervised training. By constructing a SPG, SSPC-Net [29] achieves point cloud semantic segmentation through a semi-supervised graph neural network (GNN). The features of superpoints are extracted as input for the generation and propagation of pseudo labels. A coupled attention mechanism is employed to enhance the extraction of discriminative contextual features. Deng et al. [27] used the superpoint as a constraint and a guide for pseudo label propagation. This framework consists of a superpoint generation module for superpoint generation, a pseudo-label optimization module for the identification of pseudo labels with low confidence, a superpoint feature aggregation module for feature extraction, and an edge prediction module for edge constraints. Refer to Ref. [27] for an illustration of the superpoint-based structure. Although these methods have achieved instance segmentation with much fewer annotations, they highly depend on the grouping quality.

As an alternative strategy to the above methods, learning from unlabeled points can act as a pre-training pretext. Drawing inspiration from recent developments in contrastive learning for self-supervised tasks, Jiang et al. [77] introduced guided point contrastive learning, which improves feature representation in the semi-supervised network. Augmented point clouds generated from input point clouds are fed into an unsupervised branch for backbone network training. The backbone network, classifier, and projector are shared with the supervised branch to produce semantic scores. By incorporating self-supervised learning, Zhang et al. [61] proposed a two-component network for weakly supervised point cloud semantic segmentation. Prior knowledge is learned from large-scale unlabeled points via a self-supervised network. Together with a sparse label propagation mechanism, the prior information is transferred to a weakly supervised network for label prediction. Zhang et al. [63] proposed a perturbed self-distillation framework for point cloud semantic segmentation tasks. The core of this framework is to maintain consistency between the perturbed branch and the original branch, bridging the information between labeled and unlabeled data. The consistency constraints are imposed to establish a graph topology among all the points. Besides, the semantic context of labeled points is used to monitor the overall understanding of the point clouds. One thing one click [30] performs semantic segmentation with one annotated point per object. A self-training approach with label propagation is integrated into this framework. With such sparse supervision, the semantic and geometric similarities are learned to generate and update pseudo labels. HybridCR [66] uses a hybrid contrastive regularization for finding the similarity in local neighborhoods and

global contexts. PointMatch [78] learns consistent representations from sparse annotations by improving the quality of pseudo labels. This is achieved by introducing super point information. Recently, Liu et al. [79] combined active learning with self-training to enhance instance segmentation performance by selecting points to be annotated. While pre-training is promising, it still requires large amounts of data for training, and it can be difficult to fine-tune models from other tasks.

5.2 Point cloud instance segmentation

Instance segmentation involves the identification and labeling of each object as a separate instance. Image-based instance segmentation can be divided into two distinct categories: detection-based methods and detection-free methods. Detection-based approaches first predict the localization of each object to generate proposals and then obtain pre-pixel instance masks [8]. For instance, YOLO [80] predicts semantic classes and the target bounding boxes for different image grids to complete image segmentation. Detection-free methods rely on the semantic segmentation results and then use clustering techniques to obtain instance labels. In particular, PFN [81] designs a framework that trains three sub-tasks, i.e., semantic segmentation, instance location, and instance count for each category. The final instance-level segmentation results are obtained by clustering. With the rise of deep learning in 3D data and the availability of large-scale annotated point cloud datasets, there has been increasing attention on deep learning based segmentation of 3D point cloud instances.

5.2.1 Fully supervised instance segmentation

Full supervised point cloud instance segmentation requires point-level instance labels. Similar to image-based instance segmentation, fully supervised methods can also be categorized into detection-based and detection-free methods. Refer to Ref. [53] and Ref. [50] for details.

Detection-based methods first predict 3D bounding boxes and then produce point-level instance masks. GSPN [53] adopts an analysis-by-synthesis strategy and produces object proposals. A region-based PointNet is designed to refine proposals and generate instance segmentation. 3D-BoNet [56] performs end-to-end regression of 3D bounding boxes and predicts point-level masks for all instances. It consists of a backbone network, followed by two parallel branches. One branch is dedicated to bounding box regression, while the other branch focuses on point mask prediction. GICN [82] approximates the instance center of each object as a Gaussian distribution. This Gaussian distribution is then sampled to generate candidates, which are subsequently used to generate corresponding bounding boxes and instance masks.

Detection-free methods first predict point-level semantic labels and then group points into instances. SGPN [50]

is an early deep learning framework for point cloud instance segmentation. Using PointNet++ as the backbone, the SGPN predicts group proposals based on a similarity matrix. ASIS [54] produces point-level instance labels via joint training with semantic supervision. Similarly, JSNet [83] and JSIS3D [84] also benefit from training instance and semantic segmentation simultaneously. Liang et al. [85] used a GNN based on attention-based neighbor search to obtain discriminative features under both semantic and instance supervision. Mean-shift post-processing was then employed to cluster embeddings for final predictions. SoftGroup [67] is a two-step framework that consists of bottom-up grouping and top-down refinement. Given the input point clouds, a soft grouping module is used to produce instance proposals on the basis of semantic scores and offset vectors. While most detection-free methods require post-processing, such as center voting or non-maximum suppression, Mask3D [71] utilizes a Transformer-based module to directly predict instance masks. Semantic and geometric information is encoded into point features through a stacked Transformer decoder, which provides an instance heatmap that indicates the similarities among the point clouds. Recently, SPFormer [86] has been developed to directly predict instances in an end-to-end manner based on superpoint cross-attention. Superpoint features are aggregated from point clouds and used as input to the Transformer decoder.

In recent years, detection-based methods, which attempt to perform the instance segmentation task by a separate detection step, have received less attention than detection-free approaches that aim for an end-to-end solution. Moreover, different backbones with varying levels of annotation have been explored. However, the accuracy of instance segmentation is still low, and the generality of existing methods lacks strong empirical evidence.

5.2.2 Weakly supervised instance segmentation

While fully supervised point cloud instance segmentation can suffer from performance degradation when dense annotations are unavailable, weakly supervised frameworks attempt to classify points into objects with small numbers of labels.

Liao et al. [87] proposed a semi-supervised framework for point cloud instance segmentation by using a bounding box for supervision. The input point clouds were decomposed into subsets by bounding box proposals. Semantic information and consistency constraints were used to produce final instance masks. Hou et al. [88] designed a pre-training method that could gracefully fine-tune an instance segmentation network. To further enhance feature exploitation, the spatial information was integrated into contrastive pre-training. Tang et al. [70] grouped point clouds into superpoints and explored inter-superpoint

spatial and semantic relationships. The final instance segmentation was completed by clustering with volume constraints. To address the issue of ambiguous labels due to intersections among bounding boxes, WISGP [31] divided points into two distinct sets. The univocal set consists of points with clear instance labels, while the equivocal set comprises points with uncertain belongings. Geometric representations such as polygon meshes and superpoints were employed to propagate univocal labels to connected equivocal points. Pseudo labels were assigned to the remaining equivocal points based on an instance segmentation network. The model was retrained with all labeled points to produce final instance segmentation results. One Thing One Click++ [89] expanded the previous self-training framework for weakly supervised 3D instance segmentation. A 3D-Unet and a Relation Net were employed to aggregate features and learn pairwise similarities. The initial pseudo labels generated by annotations were iteratively updated to refine the final outputs. To further alleviate dependency on annotations, FreePoint [90] explored unsupervised point cloud instance segmentation. A multi-cut algorithm was used to group point clouds into coarse instance masks based on point features that consist of coordinates, colors, normals, and self-supervised deep features. This grouping generated pseudo labels for weakly supervised network training. The framework can be integrated into supervised semantic instance segmentation as an unsupervised pre-training pretext. The aforementioned weakly supervised methods have achieved significant improvements in recent years, but they still face difficulties in handling unbalanced data.

6 Learning-based segmentation with multi-modality

Recent advances in foundation models of 2D vision and NLP have inspired the exploration of multi-modality methods in 3D models [12, 91–98]. For instance, Peng et al. [97] proposed a zero-shot approach that co-embeds point features with images and text. Rozenberszki et al. [12] presented a language-grounded method by discovering the joint embedding space of point features and text features. Liu et al. [91] transferred the knowledge from 2D to 3D for part segmentation. Wang et al. [92] trained a multi-modal model that learns from vision, language, and geometry to improve 3D semantic scene understanding. Xue et al. [93] introduced a unified representation of images, text, and 3D point clouds by aligning them during pre-training. Ding et al. [94] distilled knowledge from vision-language models for 3D scene understanding tasks. Zeng et al. [95] aligned 3D representations to open-world vocabularies via a cross-modal contrastive objective. Zhang et al. [98] performed text-scene paired semantic understanding with language-assisted learning. How to facilitate and adapt multi-modalities with point clouds for better

scene understanding is worth exploring. These methods utilize rich information from vision and text, enabling a more comprehensive representation of the indoor scene. However, these approaches require high computational resources, and pre-training is highly dependent on limited multi-modal datasets.

7 Performance evaluation

7.1 Evaluation metrics

The widely adopted evaluation metrics for indoor point cloud semantic segmentation include overall accuracy (OA), mean intersection over union (mIoU), and mean accuracy (mAcc).

The standard evaluation metric for indoor point cloud instance segmentation is mean average precision with IoU thresholds (mAP) from 0.5 to 0.95. In particular, mAP@50 is AP score with IoU thresholds of 0.5. Additionally, mean precision (mPrec) and mean recall (mRec) are frequently used criteria on the S3DIS dataset.

7.2 Results on public datasets

Semantic segmentation results. Tables 1 and 2 present indoor point cloud semantic segmentation results of different methods on S3DIS Area 5 and ScanNet v2, respectively. We can observe that the state-of-the-art methods outperform the pioneering work of PointNet [41] with more than 20% mIoU gains. Transformer-based methods [48, 49, 65] have been the dominant methods in recent years, following the great success in NLP and image understanding. Meanwhile, several weakly supervised methods show the possibility of achieving semantic segmentation with fewer data, reaching more than 65% of mIoU on S3DIS Area 5 and 70% on ScanNet. These results are encouraging, although there is still a gap between fully supervised and weakly supervised approaches. It is desirable to further improve the ability to extract features from limited annotated data.

Instance segmentation results. Tables 3 and 4 present indoor point cloud instance segmentation results of different methods on S3DIS Area 5 and ScanNet v2, respectively. Detection-free methods have received more attention than detection-based methods because they attempt to complete the instance segmentation task in an end-to-end manner. Several networks [31, 68, 70, 87, 89] have started to learn instance information from limited annotation. These results clearly show that there is still room for improvement in point cloud instance segmentation using weakly supervised learning.

8 Discussion

Point cloud segmentation is a crucial task in 3D indoor scene understanding. With the availability of 3D datasets, deep learning-based segmentation methods have gained significant attention and have contributed to the progress

Table 1 Semantic segmentation results of different methods on S3DIS Area 5

| Methods | OA | mIoU | mAcc |
|-----------------------------|------|------|------|
| <i>Fully supervised</i> | | | |
| SEGCloud [36] | – | 48.9 | 57.4 |
| PointNet [41] | – | 41.1 | 49.0 |
| PointCNN [44] | 85.9 | 57.3 | 63.9 |
| DGCNN [46] | – | 57.3 | 63.9 |
| SPG [25] | 86.4 | 58.0 | 66.5 |
| PGCNet [60] | 86.2 | 53.6 | 63.9 |
| PointWeb [52] | 87.0 | 60.3 | 66.6 |
| PCT [47] | – | 61.3 | 67.7 |
| PAT [51] | – | 60.1 | 70.8 |
| KPConv [55] | – | 67.1 | 72.8 |
| PointTransformer [48] | 90.8 | 70.4 | 76.5 |
| PointTransformerV2 [49] | 91.1 | 71.6 | 77.9 |
| Stratified Transformer [65] | 91.5 | 72.0 | 78.1 |
| <i>Weakly supervised</i> | | | |
| Deng et al. [27] (10%) | – | 51.5 | – |
| SSPC-Net [29] (0.01%) | – | 51.5 | – |
| OTOC [30] (0.02%) | – | 50.1 | – |
| Zhang et al. [61] (1%) | – | 61.8 | – |
| PSD [63] (1%) | – | 63.5 | – |
| SQN [69] (1%) | – | 63.6 | – |
| HybridCR [66] (1%) | – | 65.3 | – |

Table 2 Semantic segmentation results of different methods on ScanNet v2 validation set and test set

| Methods | Val mIoU | Test mIoU |
|-----------------------------|----------|-----------|
| <i>Fully supervised</i> | | |
| PointNet [42] | 53.5 | 55.7 |
| PointCNN [44] | – | 48.4 |
| RandLA-Net [43] | – | 64.5 |
| KPConv [55] | 69.2 | 68.6 |
| PointTransformer [48] | 70.6 | – |
| Stratified Transformer [65] | 74.3 | 73.7 |
| PointTransformerV2 [49] | 75.4 | 75.2 |
| <i>Weakly supervised</i> | | |
| MPRM [59] (sub-cloud) | – | 41.1 |
| Zhang et al. [61] (1%) | – | 51.1 |
| PSD [63] (1%) | – | 54.7 |
| SQN [69] (0.1%) | – | 56.9 |
| HybridCR [66] (1%) | 56.9 | 56.8 |
| OTOC [30] (0.02%) | 70.5 | 69.1 |

in this field. However, obtaining accurate segmentation results often requires dense annotations, which is a laborious and costly process. In order to mitigate the reliance on extensive annotations and enable learning from limited labeled data, the research focus has shifted towards weakly supervised approaches in recent years. By exploring weakly supervised frameworks, researchers aim to achieve satisfactory segmentation results while minimizing the annotation efforts and costs involved. Despite the rapid development of point cloud segmentation, existing frameworks still face several challenges.

Table 3 Instance segmentation results of different methods on S3DIS Area 5

| Methods | mAP | mAP@50 | mPrec | mRec |
|--------------------------|------|--------|-------|------|
| <i>Fully supervised</i> | | | | |
| SGPN [50] | – | – | 36.0 | 28.7 |
| ASIS [54] | – | – | 55.3 | 42.4 |
| 3D-BoNet [56] | – | – | 57.5 | 40.2 |
| 3D-MPA [57] | – | – | 63.1 | 58.0 |
| PointGroup [58] | – | 57.8 | 61.9 | 62.1 |
| DyCo3D [62] | – | – | 64.3 | 64.2 |
| SSTNet [64] | 42.7 | 59.3 | 65.6 | 64.2 |
| SoftGroup [67] | 51.6 | 66.1 | 73.6 | 66.6 |
| Mask3D [71] | 56.6 | 68.4 | 68.7 | 66.3 |
| <i>Weakly supervised</i> | | | | |
| WISGP [31] (3D Box) | 37.2 | 51.0 | 44.3 | 56.7 |
| SegGroup [68] (0.02%) | 21.0 | 29.8 | 47.2 | 34.9 |
| 3D-WSIS [70] (0.02%) | 23.3 | 33.0 | 50.8 | 38.9 |

Table 4 Instance segmentation results of different methods on ScanNet v2 validation set and test set

| Methods | Val mAP | Val mAP@50 | Test mAP | Test mAP@50 |
|--------------------------|---------|------------|----------|-------------|
| <i>Fully supervised</i> | | | | |
| SGPN [50] | – | – | 4.9 | 14.3 |
| GSPN [53] | 19.3 | 33.8 | – | 30.6 |
| 3D-BoNet [56] | – | – | 25.3 | 48.8 |
| 3D-MPA [57] | 35.5 | 59.1 | 35.5 | 61.1 |
| PointGroup [58] | 34.8 | 56.7 | 40.7 | 63.6 |
| DyCo3D [62] | 35.4 | 57.6 | 39.5 | 64.1 |
| SSTNet [64] | 49.4 | 64.3 | 50.6 | 69.8 |
| SoftGroup [67] | 46.0 | 67.6 | 50.4 | 76.1 |
| Mask3D [71] | 55.2 | 73.7 | 56.6 | 78.0 |
| <i>Weakly supervised</i> | | | | |
| WISGP [31] (3D Box) | 35.2 | 56.9 | – | – |
| SPIB [87] (0.16%) | – | 38.6 | – | – |
| SegGroup [68] (0.02%) | 23.4 | 43.4 | 24.6 | 44.5 |
| 3D-WSIS [70] (0.02%) | 28.1 | 47.2 | 25.1 | 47.0 |
| OTOC++ [89] (0.02%) | – | – | 32.6 | 52.9 |

8.1 Datasets and representations

The size of annotated point cloud data is still limited compared to that of image datasets. Although acquiring point clouds becomes affordable, annotating point clouds is still a time-consuming task. Since both fully supervised and pre-training [99, 100] require a large amount of data, larger datasets with more diverse scenes are desired to advance learning-based point cloud segmentation. Therefore, an efficient and user-friendly annotation method for large datasets is needed. This might be achieved by unsupervised approaches with geometric priors. The recently developed datasets, such as the ScanNet200 dataset [12], have drawn increasing attention to imbalanced learning [101] in point cloud segmentation.

Existing point cloud segmentation methods use different data formats, including point clouds, RGB-D images, voxels, and geometric primitives. Each data format has its advantages and drawbacks in different 3D scene understand-

ing tasks. On the basis of point-based networks, we can now directly process point clouds for training and reasoning. Obviously, not all points are needed for scene perception. For indoor scene point cloud data, finding a better representation is still a promising direction of research.

8.2 Data efficiency and multi-modality

Data-efficient learning frameworks are highly desirable because they alleviate the burden of collecting extensive dense annotations for training the model. Although current weakly supervised point cloud segmentation methods can achieve competitive performance with fully supervised learning, there are still gaps to be filled. More importantly, the generality and robustness of these data-efficient methods are not convincing, as they mainly test on public datasets with limited sizes rather than on open-world scenes. Therefore, further exploration of generalist models is the trend for the future.

One promising route is to integrate other modalities, such as images and natural languages. Previous works [37, 102, 103] have explored the combination of 2D images and 3D point clouds for better understanding of scenes. Recent developments in foundation models of 2D vision and NLP have served as inspiration sources for investigating multi-modalities in 3D data [12, 91–98]. While these methods achieve incredible results in different 3D tasks, adapting knowledge from other modalities to indoor point cloud segmentation is still challenging. In addition, collecting adequate multi-modal pre-training data can be costly. How to facilitate and adapt multi-modalities with point clouds for a better understanding of indoor scenes is worth exploring.

8.3 Learning methods for dynamic scene segmentation

Current learning-based indoor point cloud segmentation methods are mostly designed for static scenes. Indoor objects can be moved around in real-world scenarios, allowing for a more comprehensive representation of the indoor scene. Moreover, annotating such dynamic scenes is even more costly than annotating 3D point clouds. 4D representation learning has become the core of dynamic feature exploitation. Recent work [104, 105] has explored 4D feature extraction and distillation to improve downstream tasks such as scene segmentation. Transferring such information to varying scales of indoor scenes is still challenging. The development of learning methods for dynamic scene segmentation is an interesting prospect for further investigation.

9 Conclusion

Point cloud segmentation plays a key role in 3D vision and intelligence. This paper aims to provide a concise overview of point cloud segmentation techniques for understanding 3D indoor scenes. First, we present public

3D point cloud datasets, which are the foundation of point cloud segmentation research, especially for deep learning-based methods. Second, we review representative approaches for indoor scene point cloud segmentation, including geometry-based and learning-based methods. Geometry-based approaches extract geometric information and can be combined with learning-based methods. Learning-based methods can be divided into structured data-based and point-based methods. We mainly consider point-based semantic and instance segmentation frameworks, including fully supervised networks and weakly supervised networks. Finally, we discuss the open problems in the field and outline future research directions. We expect that this review can provide insights into the field of indoor scene point cloud segmentation and stimulate new research.

Abbreviations

CAD, computer-aided design; CNN, convolutional neural network; GCN, graph convolutional network; MLP, multi-layer perceptron; NLP, natural language processing; RANSAC, random sample consensus; SPG, superpoint graph.

Acknowledgements

The authors express their gratitude to the anonymous reviewers and the editor for their invaluable feedback, which greatly improved the quality of this manuscript.

Author contributions

All authors contributed to the idea for the article. Literature search and analysis were performed by YS. YS prepared the manuscript initially, and all authors participated in the comment and modification of the paper. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (No. 61972458) and Zhejiang Provincial Natural Science Foundation of China (No. LZ23F020002).

Data availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Competing interests

The authors declare no competing interests.

Author details

¹School of Information Science and Technology, Zhejiang Shuren University, Hangzhou, 310015, China. ²School of Information Science and Technology, Hangzhou Normal University, Hangzhou, 311121, China.

Received: 5 July 2023 Revised: 23 April 2024 Accepted: 24 April 2024

Published online: 07 June 2024

References

1. Rusu, R. B., Marton, Z. C., Blodow, N., Dolha, M. E., & Beetz, M. (2008). Towards 3D point cloud based object maps for household environments. *Robotics and Autonomous Systems*, 56(11), 927–941.
2. Zhu, Y., Mottaghi, R., Kolve, E., Lim, J. J., Gupta, A., Li, F., et al. (2017). Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *Proceedings of the IEEE international conference on robotics and automation* (pp. 3357–3364). Piscataway: IEEE.
3. Wirth, F., Quehl, J., Ota, J., & Stiller, C. (2019). Pointatme: efficient 3D point cloud labeling in virtual reality. In *Proceedings of the 2019 IEEE intelligent vehicles symposium* (pp. 1693–1698). Piscataway: IEEE.

4. Li, J., Gao, W., Wu, Y., Liu, Y., & Shen, Y. (2022). High-quality indoor scene 3D reconstruction with RGB-D cameras: a brief review. *Computational Visual Media*, 8(3), 369–393.
 5. Nguyen, A., & Le, B. (2013). 3D point cloud segmentation: a survey. In *Proceedings of the 6th IEEE conference on robotics, automation and mechatronics* (pp. 225–230). Piscataway: IEEE.
 6. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440). Piscataway: IEEE.
 7. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 770–778). Piscataway: IEEE.
 8. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2961–2969). Piscataway: IEEE.
 9. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., & Bennamoun, M. (2020). Deep learning for 3D point clouds: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12), 4338–4364.
 10. Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. (2017). ScanNet: richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5828–5839). Piscataway: IEEE.
 11. Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., et al. (2016). 3D semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1534–1543). Piscataway: IEEE.
 12. Rozenberszki, D., Litany, O., & Dai, A. (2022). Language-grounded indoor 3D semantic segmentation in the wild. In S. Avidan, G. J. Brostow, M. Cissé, et al. (Eds.), *Proceedings of the 17th European conference on computer vision* (pp. 125–141). Cham: Springer.
 13. Anand, A., Koppula, H. S., Joachims, T., & Saxena, A. (2013). Contextually guided semantic labeling and search for three-dimensional point clouds. *The International Journal of Robotics Research*, 32(1), 19–34.
 14. Lai, K., Bo, L., & Fox, D. (2014). Unsupervised feature learning for 3D scene labeling. In *Proceedings of the IEEE international conference on robotics and automation* (pp. 3050–3057). Piscataway: IEEE.
 15. Mattausch, O., Panozzo, D., Mura, C., Sorkine-Hornung, O., & Pajarola, R. (2014). Object detection and classification from large-scale cluttered indoor scans. *Computer Graphics Forum*, 33(2), 11–21.
 16. Hu, S.-M., Cai, J.-X., & Lai, Y.-K. (2018). Semantic labeling and instance segmentation of 3D point clouds using patch context analysis and multiscale processing. *IEEE Transactions on Visualization and Computer Graphics*, 26(7), 2485–2498.
 17. Nan, L., Xie, K., & Sharf, A. (2012). A search-classify approach for cluttered indoor scene understanding. *ACM Transactions on Graphics*, 31(6), 1–10.
 18. Li, Y., Dai, A., Guibas, L., & Nießner, M. (2015). Database-assisted object retrieval for real-time 3D reconstruction. *Computer Graphics Forum*, 34(2), 435–446.
 19. Shi, Y., Long, P., Xu, K., Huang, H., & Xiong, Y. (2016). Data-driven contextual modeling for 3D scene understanding. *Computers & Graphics*, 55, 55–67.
 20. Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.
 21. Monszpart, A., Mellado, N., Brostow, G. J., & Mitra, N. J. (2015). RAPter: rebuilding man-made scenes with regular arrangements of planes. *ACM Transactions on Graphics*, 34(4), 1–12.
 22. Sun, Y., Miao, Y., Yu, L., & Renato, P. (2018). Abstraction and understanding of indoor scenes from single-view RGB-D scanning data. *Journal of Computer-Aided Design & Computer Graphics*, 30(6), 1046–1054.
 23. Yu, L., Sun, Y., Zhang, X., Miao, Y., & Zhang, X. (2021). Point cloud instance segmentation of indoor scenes using learned pairwise patch relations. *IEEE Access*, 9, 15891–15901.
 24. Huang, S.-S., Ma, Z.-Y., Mu, T.-J., Fu, H., & Hu, S.-M. (2021). Supervoxel convolution for online 3D semantic segmentation. *ACM Transactions on Graphics*, 40(3), 1–15.
 25. Landrieu, L., & Simonovsky, M. (2018). Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4558–4567). Piscataway: IEEE.
 26. Cheng, M., Hui, L., Xie, J., Yang, J., & Kong, H. (2020). Cascaded non-local neural network for point cloud semantic segmentation. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems* (pp. 8447–8452). Piscataway: IEEE.
 27. Deng, S., Dong, Q., Liu, B., & Hu, Z. (2022). Superpoint-guided semi-supervised semantic segmentation of 3D point clouds. In *Proceedings of the international conference on robotics and automation* (pp. 9214–9220). Piscataway: IEEE.
 28. Shi, X., Xu, X., Chen, K., Cai, L., Foo, C. S., & Jia, K. (2021). Label-efficient point cloud semantic segmentation: an active learning approach. arXiv preprint. [arXiv:2101.06931](https://arxiv.org/abs/2101.06931).
 29. Cheng, M., Hui, L., Xie, J., & Yang, J. (2021). SSPC-Net: semi-supervised semantic 3D point cloud segmentation network. In *Proceedings of the 33rd AAAI conference on artificial intelligence* (pp. 1140–1147). Palo Alto: AAAI Press.
 30. Liu, Z., Qi, X., & Fu, C.-W. (2021). One thing one click: a self-training approach for weakly supervised 3D semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1726–1736). Piscataway: IEEE.
 31. Du, H., Yu, X., Hussain, F., Armin, M. A., Petersson, L., & Li, W. (2023). Weakly-supervised point cloud instance segmentation with geometric priors. In *Proceedings of IEEE/CVF winter conference on applications of computer vision* (pp. 4271–4280). Piscataway: IEEE.
 32. Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3D shape recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 945–953). Piscataway: IEEE.
 33. Boulch, A., Guerry, J., Le Saux, B., & Audebert, N. (2018). SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Computers & Graphics*, 71, 189–198.
 34. Maturana, D., & Scherer, S. (2015). VoxNet: a 3D convolutional neural network for real-time object recognition. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems* (pp. 922–928). Piscataway: IEEE.
 35. Meng, H.-Y., Gao, L., Lai, Y.-K., & Manocha, D. (2019). VV-Net: voxel VAE net with group convolutions for point cloud segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8500–8508). Piscataway: IEEE.
 36. Tchapmi, L., Choy, C., Armeni, I., Gwak, J., & Savarese, S. (2017). SEGCloud: semantic segmentation of 3D point clouds. In *Proceedings of international conference on 3D vision* (pp. 537–547). Piscataway: IEEE.
 37. Hou, J., Dai, A., & Nießner, M. (2019). 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4421–4430). Piscataway: IEEE.
 38. Riegler, G., Osman Ulusoy, A., & Geiger, A. (2017). Octnet: learning deep 3D representations at high resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3577–3586). Piscataway: IEEE.
 39. Wang, P.-S., Liu, Y., Guo, Y.-X., Sun, C.-Y., & Tong, X. (2017). O-CNN: octree-based convolutional neural networks for 3D shape analysis. *ACM Transactions on Graphics*, 36(4).
 40. Wang, P.-S. (2023). OctFormer: octree-based transformers for 3D point clouds. *ACM Transactions on Graphics*, 42(4), 1–11.
 41. Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 652–660). Piscataway: IEEE.
 42. Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). Pointnet++: deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 30th international conference on neural information processing systems* (pp. 5099–5108). Red Hook: Curran Associates.
 43. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., et al. (2020). RandLA-Net: efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11108–11117). Piscataway: IEEE.
 44. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., & Chen, B. (2018). PointCNN: convolution on x-transformed points. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 828–838). Red Hook: Curran Associates.
 45. Te, G., Hu, W., Zheng, A., & Guo, Z. (2018). RGCNN: regularised graph CNN for point cloud segmentation. In *Proceedings of the 26th ACM international conference on multimedia* (pp. 746–754). New York: ACM.
 46. Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, 38(5).

47. Guo, M.-H., Cai, J.-X., Liu, Z.-N., Mu, T.-J., Martin, R. R., & Hu, S.-M. (2021). PCT: point cloud transformer. *Computational Visual Media*, 7(2), 187–199.
48. Zhao, H., Jiang, L., Jia, J., Torr, P. H., & Koltun, V. (2021). Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 16259–16268). Piscataway: IEEE.
49. Wu, X., Lao, Y., Jiang, L., Liu, X., & Zhao, H. (2022). Point transformer v2: grouped vector attention and partition-based pooling. In *Proceedings of the 35th international conference on neural information processing systems* (pp. 33330–33342). Red Hook: Curran Associates.
50. Wang, W., Yu, R., Huang, Q., & Neumann, U. (2018). SGPN: similarity group proposal network for 3D point cloud instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2569–2578). Piscataway: IEEE.
51. Yang, J., Zhang, Q., Ni, B., Li, L., Liu, J., Zhou, M., et al. (2019). Modeling point clouds with self-attention and Gumbel subset sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3323–3332).
52. Zhao, H., Jiang, L., Fu, C.-W., & Jia, J. (2019). PointWeb: enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5565–5573). Piscataway: IEEE.
53. Yi, L., Zhao, W., Wang, H., Sung, M., & Guibas, L. J. (2019). GSPN: generative shape proposal network for 3D instance segmentation in point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3947–3956). Piscataway: IEEE.
54. Wang, X., Liu, S., Shen, X., Shen, C., & Jia, J. (2019). Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4096–4105). Piscataway: IEEE.
55. Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., & Guibas, L. J. (2019). KPConv: flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6411–6420). Piscataway: IEEE.
56. Yang, B., Wang, J., Clark, R., Hu, Q., Wang, S., Markham, A., et al. (2019). Learning object bounding boxes for 3D instance segmentation on point clouds. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 6737–6746). Red Hook: Curran Associates.
57. Engelmann, F., Bokeloh, M., Fathi, A., Leibe, B., & Nießner, M. (2020). 3D-MPA: multi-proposal aggregation for 3D semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9031–9040). Piscataway: IEEE.
58. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.-W., & Jia, J. (2020). PointGroup: dual-set point grouping for 3D instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4867–4876). Piscataway: IEEE.
59. Wei, J., Lin, G., Yap, K.-H., Hung, T.-Y., & Xie, L. (2020). Multi-path region mining for weakly supervised 3D semantic segmentation on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4384–4393). Piscataway: IEEE.
60. Sun, Y., Miao, Y., Chen, J., & Pajarola, R. (2020). PGCNet: patch graph convolutional network for point cloud segmentation of indoor scenes. *The Visual Computer*, 36(10), 2407–2418.
61. Zhang, Y., Li, Z., Xie, Y., Qu, Y., Li, C., & Mei, T. (2021). Weakly supervised semantic segmentation for large-scale point cloud. In *Proceedings of the 33rd AAAI conference on artificial intelligence* (pp. 3421–3429). Palo Alto: AAAI Press.
62. He, T., Shen, C., & van den Hengel, A. (2021). DyCo3D: robust instance segmentation of 3D point clouds through dynamic convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 354–363). Piscataway: IEEE.
63. Zhang, Y., Qu, Y., Xie, Y., Li, Z., Zheng, S., & Li, C. (2021). Perturbed self-distillation: weakly supervised large-scale point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15520–15528). Piscataway: IEEE.
64. Liang, Z., Li, Z., Xu, S., Tan, M., & Jia, K. (2021). Instance segmentation in 3D scenes using semantic superpoint tree networks. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2783–2792). Piscataway: IEEE.
65. Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., et al. (2022). Stratified transformer for 3D point cloud segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8500–8509). Piscataway: IEEE.
66. Li, M., Xie, Y., Shen, Y., Ke, B., Qiao, R., Ren, B., et al. (2022). HybridCR: weakly-supervised 3D point cloud semantic segmentation via hybrid contrastive regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14930–14939). Piscataway: IEEE.
67. Vu, T., Kim, K., Luu, T. M., Nguyen, T., & Yoo, C. D. (2022). Softgroup for 3D instance segmentation on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2708–2717). Piscataway: IEEE.
68. Tao, A., Duan, Y., Wei, Y., Lu, J., & Zhou, J. (2022). SegGroup: seg-level supervision for 3D instance and semantic segmentation. *IEEE Transactions on Image Processing*, 31, 4952–4965.
69. Hu, Q., Yang, B., Fang, G., Guo, Y., Leonardi, A., Trigoni, N., et al. (2022). SQN: weakly-supervised semantic segmentation of large-scale 3D point clouds. In S. Avidan, G. J. Brostow, M. Cissé, et al. (Eds.), *Proceedings of the 17th European conference on computer vision* (pp. 600–619). Cham: Springer.
70. Tang, L., Hui, L., & Xie, J. (2022). Learning inter-superpoint affinity for weakly supervised 3D instance segmentation. In *Proceedings of the 16th Asian conference on computer vision* (Vol. 13841, pp. 176–192). Cham: Springer.
71. Schulte, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., & Leibe, B. (2023). Mask3D: mask transformer for 3D semantic instance segmentation. In *Proceedings of the IEEE international conference on robotics and automation* (pp. 8216–8223). Piscataway: IEEE.
72. Wang, L., Huang, Y., Hou, Y., Zhang, S., & Shan, J. (2019). Graph attention convolution for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10296–10305). Piscataway: IEEE.
73. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Proceedings of the 30th international conference on neural information processing systems*. Red Hook: Curran Associates.
74. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint. [arXiv:2010.11929](https://arxiv.org/abs/2010.11929).
75. Xu, X., & Lee, G. H. (2020). Weakly supervised semantic point cloud segmentation: towards 10x fewer labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13706–13715). Piscataway: IEEE.
76. Wu, Z., Wu, Y., Lin, G., Cai, J., & Qian, C. (2022). Dual adaptive transformations for weakly supervised point cloud segmentation. In S. Avidan, G. J. Brostow, M. Cissé, et al. (Eds.), *Proceedings of the 17th European conference on computer vision* (pp. 78–96). Cham: Springer.
77. Jiang, L., Shi, S., Tian, Z., Lai, X., Liu, S., Fu, C.-W., et al. (2021). Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6423–6432). Piscataway: IEEE.
78. Wu, Y., Cai, S., Yan, Z., Li, G., Yu, Y., Han, X., et al. (2022). PointMatch: a consistency training framework for weakly supervised semantic segmentation of 3D point clouds. arXiv preprint. [arXiv:2202.10705](https://arxiv.org/abs/2202.10705).
79. Liu, G., van Kaick, O., Huang, H., & Hu, R. (2022). Active self-training for weakly supervised 3D scene semantic segmentation. arXiv preprint. [arXiv:2209.07069](https://arxiv.org/abs/2209.07069).
80. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788). Piscataway: IEEE.
81. Liang, X., Lin, L., Wei, Y., Shen, X., Yang, J., & Yan, S. (2017). Proposal-free network for instance-level object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12), 2978–2991.
82. Liu, S.-H., Yu, S.-Y., Wu, S.-C., Chen, H.-T., & Liu, T.-L. (2020). Learning gaussian instance segmentation in point clouds. arXiv preprint. [arXiv:2007.09860](https://arxiv.org/abs/2007.09860).
83. Zhao, L., & Tao, W. (2020). JSNet: joint instance and semantic segmentation of 3D point clouds. In *Proceedings of the 32nd AAAI conference on artificial intelligence* (pp. 12951–12958). Palo Alto: AAAI Press.
84. Pham, Q.-H., Nguyen, T., Hua, B.-S., Roig, G., & Yeung, S.-K. (2019). JSIS3D: joint semantic-instance segmentation of 3D point clouds with multi-task pointwise networks and multi-value conditional random fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8827–8836). Piscataway: IEEE.

85. Liang, Z., Yang, M., Li, H., & Wang, C. (2020). 3D instance embedding learning with a structure-aware loss function for point cloud segmentation. *IEEE Robotics and Automation Letters*, 5(3), 4915–4922.
86. Sun, J., Qing, C., Tan, J., & Xu, X. (2023). Superpoint transformer for 3D scene instance segmentation. In *Proceedings of the 37th AAAI conference on artificial intelligence* (pp. 2393–2401). Palo Alto: AAAI Press.
87. Liao, Y., Zhu, H., Zhang, Y., Ye, C., Chen, T., & Fan, J. (2021). Point cloud instance segmentation with semi-supervised bounding-box mining. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 10159–10170.
88. Hou, J., Graham, B., Nießner, M., & Xie, S. (2021). Exploring data-efficient 3D scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15587–15597). Piscataway: IEEE.
89. Liu, Z., Qi, X., & Fu, C.-W. (2023). One Thing One Click++: self-training for weakly supervised 3D scene understanding. arXiv preprint. [arXiv:2303.14727](https://arxiv.org/abs/2303.14727).
90. Zhang, Z., Ding, J., Jiang, L., Dai, D., & Xia, G.-S. (2023). FreePoint: unsupervised point cloud instance segmentation. arXiv preprint. [arXiv:2305.06973](https://arxiv.org/abs/2305.06973).
91. Liu, M., Zhu, Y., Cai, H., Han, S., Ling, Z., Porikli, F., et al. (2023). 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* (pp. 21736–21746). Piscataway: IEEE.
92. Wang, Z., Cheng, B., Zhao, L., Xu, D., Tang, Y., & Sheng, L. (2023). VL-SAT: visual-linguistic semantics assisted training for 3D semantic scene graph prediction in point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 21560–21569). Piscataway: IEEE.
93. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., et al. (2023). ULIP: learning a unified representation of language, images, and point clouds for 3D understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1179–1189). Piscataway: IEEE.
94. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., & Qi, X. (2023). PLA: language-driven open-vocabulary 3D scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7010–7019). Piscataway: IEEE.
95. Zeng, Y., Jiang, C., Mao, J., Han, J., Ye, C., Huang, Q., et al. (2023). CLIP2: contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15244–15253). Piscataway: IEEE.
96. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., et al. (2023). Open-vocabulary point-cloud object detection without 3D annotation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1190–1199). Piscataway: IEEE.
97. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al. (2023). Openscene: 3D scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 815–824). Piscataway: IEEE.
98. Zhang, J., Fan, G., Wang, G., Su, Z., Ma, K., & Yi, L. (2023). Language-assisted 3D feature learning for semantic scene understanding. In *Proceedings of the 37th AAAI conference on artificial intelligence* (pp. 3445–3453). Palo Alto: AAAI Press.
99. Xie, S., Gu, J., Guo, D., Qi, C. R., Guibas, L., & Litany, O. (2020). PointContrast: unsupervised pre-training for 3D point cloud understanding. In A. Vedaldi, H. Bischof, T. Brox, et al. (Eds.), *Proceedings of the 16th European conference on computer vision* (pp. 574–591). Cham: Springer.
100. Yang, Y.-Q., Guo, Y.-X., Xiong, J.-Y., Liu, Y., Pan, H., Wang, P.-S., et al. (2023). Swin3D: a pretrained transformer backbone for 3D indoor scene understanding. arXiv preprint. [arXiv:2304.06906](https://arxiv.org/abs/2304.06906).
101. Zhong, Z., Cui, J., Yang, Y., Wu, X., Qi, X., Zhang, X., et al. (2023). Understanding imbalanced semantic segmentation through neural collapse. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19550–19560). Piscataway: IEEE.
102. Yu, P.-C., Sun, C., & Sun, M. (2022). Data efficient 3D learner via knowledge transferred from 2D model. In S. Avidan, G. J. Brostow, M. Cissé, et al. (Eds.), *Proceedings of the 17th European conference on computer vision* (pp. 182–198). Cham: Springer.
103. Kweon, H., & Yoon, K.-J. (2022). Joint learning of 2D-3D weakly supervised semantic segmentation. In *Proceedings of the 35th international conference on neural information processing systems* (pp. 30499–30511). Red Hook: Curran Associates.
104. Chen, Y., Nießner, M., & Dai, A. (2022). 4DContrast: contrastive learning with dynamic correspondences for 3D scene understanding. In S. Avidan, G. J. Brostow, M. Cissé, et al. (Eds.), *Proceedings of the 17th European conference on computer vision* (pp. 543–560). Cham: Springer.
105. Zhang, Z., Dong, Y., Liu, Y., & Yi, L. (2023). Complete-to-partial 4D distillation for self-supervised point cloud sequence representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17661–17670). Piscataway: IEEE.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)