technologies

MDPI

Review

Advancements in Semantic Segmentation of 3D Point Clouds for Scene Understanding Using Deep Learning

Hafsa Benallal [1,*], Nadine Abdallah Saab [2], Hamid Tairi [1,*], Ayman Alfalou [2] and Jamal Riffi [1]

[1] L3IA Laboratory, Computer Science Department, Sidi Mohamed Ben Abdellah University, Fes 30000, Morocco; jamal.riffi@usmba.ac.ma
[2] L@bISEN, Light Scattering Learning Team, ISEN Yncrea Ouest, 29200 Brest, France; nadine.abdallah-saab@isen-ouest.yncrea.fr (N.A.S.); ayman.alfalou@isen-ouest.yncrea.fr (A.A.)
* Correspondence: hafsa.benallal@usmba.ac.ma (H.B.); hamid.tairi@usmba.ac.ma (H.T.)

Abstract

Three-dimensional semantic segmentation is a fundamental problem in computer vision with a wide range of applications in autonomous driving, robotics, and urban scene understanding. The task involves assigning semantic labels to each point in a 3D point cloud, a data representation that is inherently unstructured, irregular, and spatially sparse. In recent years, deep learning has become the dominant framework for addressing this task, leading to a broad variety of models and techniques designed to tackle the unique challenges posed by 3D data. This survey presents a comprehensive overview of deep learning methods for 3D semantic segmentation. We organize the literature into a taxonomy that distinguishes between supervised and unsupervised approaches. Supervised methods are further classified into point-based, projection-based, voxel-based, and hybrid architectures, while unsupervised methods include self-supervised learning strategies, generative models, and implicit representation techniques. In addition to presenting and categorizing these approaches, we provide a comparative analysis of their performance on widely used benchmark datasets, discuss key challenges such as generalization, model transferability, and computational efficiency, and examine the limitations of current datasets. The survey concludes by identifying potential directions for future research in this rapidly evolving field.

Keywords: point cloud; 3D semantic segmentation; deep learning; LiDAR; scene reconstruction; survey

check for updates

Academic Editor: Juan Gabriel Avina-Cervantes

Received: 11 June 2025
Revised: 17 July 2025
Accepted: 23 July 2025
Published: 30 July 2025

Citation: Benallal, H.; Abdallah Saab, N.; Tairi, H.; Alfalou, A.; Riffi, J. Advancements in Semantic Segmentation of 3D Point Clouds for Scene Understanding Using Deep Learning. *Technologies* **2025**, *13*, 322. https://doi.org/10.3390/technologies13080322

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Three-dimensional point clouds are represented in a three-dimensional coordinate system and are typically generated by 3D scanners or photogrammetry software. These systems, including LiDAR, stereo vision, and Time-of-Flight (ToF) sensors, capture numerous points on the external surfaces of objects or scenes. Each point corresponds to a real-world coordinate, providing a detailed and accurate geometric representation of the environment [1,2].

In addition to geometric data, 3D point clouds can also capture color or appearance information, depending on the technology used. For example, LiDAR sensors can capture reflectance values, which provide information about the properties of the surfaces being scanned, while RGB-D sensing technology can capture color data along with depth data, resulting in a detailed 3D point cloud that combines both geometric and appearance attributes.

LiDAR technology is commonly used to create point clouds for tasks such as 3D object detection, scene reconstruction, and semantic scene completion. These point clouds generate a geometric representation of the environment, which is useful in applications such as augmented reality, robotic navigation, and building information modeling (BIM). However, compared to other sensors such as RGB-D or stereo cameras, LiDAR data is often sparser and less detailed, especially when capturing data from a distance. Three-dimensional point clouds can be represented in several forms. These include point-based [3,4], voxel [5,6], range-view [7,8], and multi-view fusion [9–11], among others. A 3D model generated from a collected point cloud often represents an incomplete scene. This issue can be addressed by inferring the missing parts using semantic understanding. Each object (point, voxel, etc.) is interpreted based on its semantic meaning to generate a complete 3D model that is visually and semantically plausible [12].

Segmentation facilitates tasks such as classification, object recognition, and scene reconstruction. It represents a major challenge for researchers, as all research was initially directed toward 2D image segmentation, and then it began to deal with 3D point cloud segmentation. Point cloud segmentation is an essential phase in processing 3D point clouds, aiming to give each point a label within a specific category, thereby grouping points with similar features into homogeneous regions (see Figure 1). Three-dimensional point cloud segmentation is generally categorized into three main types: (i) part segmentation; (ii) instance segmentation; (iii) semantic segmentation. Part segmentation focuses on decomposing a single object into its constituent components, a task crucial in applications such as robotics, where it enables the understanding and manipulation of articulated or composite objects [13–15]. In augmented and virtual reality (AR/VR), it enhances user interaction through precise recognition and tracking of object parts [16]. Similarly, in computer-aided design and manufacturing (CAD/CAM), it supports shape analysis and procedural modeling for improved design and fabrication workflows [17].

Instance segmentation involves identifying and segmenting individual object instances within a 3D scene, even when they possess similar geometric or visual features. This is typically achieved by leveraging spatial, color, and texture cues, and training models on annotated datasets to learn inter-object variations [18,19]. Classical methods, such as SegCloud [20], extend 2D convolutional operations to 3D voxel grids. However, such regular representations often fail to capture the irregular and sparse nature of point clouds. Recent deep learning architectures specifically tailored for point-based data offer improved performance but remain constrained by their limited capacity to handle large-scale point sets efficiently [21].

Despite the major advances enabled by deep learning, this field still faces many challenges, not the least of which are the massive size of the data, its irregular structure, and the absence of a regular grid as in 2D. The first approaches attempted to adapt 2D convolutional networks to voxelized 3D representations, but these solutions often suffered from high memory consumption and loss of fine detail.

This review follows in the footsteps of several existing works while clearly standing out for its exclusive focus on 3D semantic segmentation. Indeed, while previous studies, such as [22–24], address point cloud segmentation or classification comprehensively (including part, semantic, or instance segmentation), others [25–27] cover a wide variety of methods and datasets. Additionally, Refs. [1,28–31] analyze deep learning methods for the 3D semantic segmentation task. In contrast, our study adopts a focused approach by exclusively analyzing recent contributions specifically dedicated to deep learning-based 3D semantic segmentation. It proposes a detailed taxonomy, highlights methodological trends, evaluates performance on reference benchmarks, and identifies current limitations as well as research prospects, as illustrated in Figure 2.

The main contributions of this study can be summarized as follows:

- We present a chronological and categorized view of the deep learning methods most relevant to 3D semantic segmentation.
- We synthesize these approaches through a clear classification based on data representation (points, voxels, views, etc.) and algorithmic paradigm (convolutional, graphs, transformers, etc.).
- We carry out a rigorous comparative analysis of the results (mIoU) obtained on the main public datasets: SemanticKITTI, ScanNet, S3DIS, etc.
- We highlight the challenges still open and areas for future research, particularly in terms of robustness, domain adaptation and computational efficiency.

The remainder of the article is structured as follows: Section 2 introduces the various representations of 3D data. Section 3 details the methodological approaches according to the proposed taxonomy. Section 4 presents the reference datasets. Section 5 presents the evaluation metrics. Section 6 provides a critical analysis of the results, and Section 7 concludes with a synthesis and outlook.

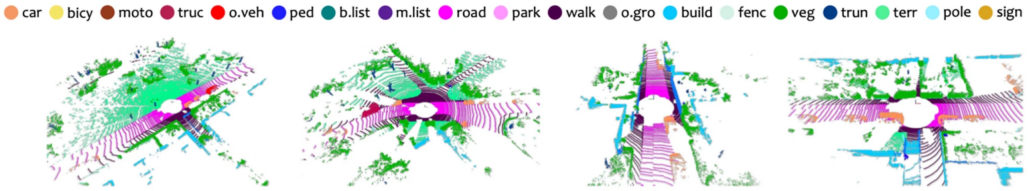


Figure 1. Example of visualization of 3D semantic segmentation on point cloud data, showing various object classes such as vehicles, road elements, and vegetation, each identified by unique colors for clear differentiation [32].

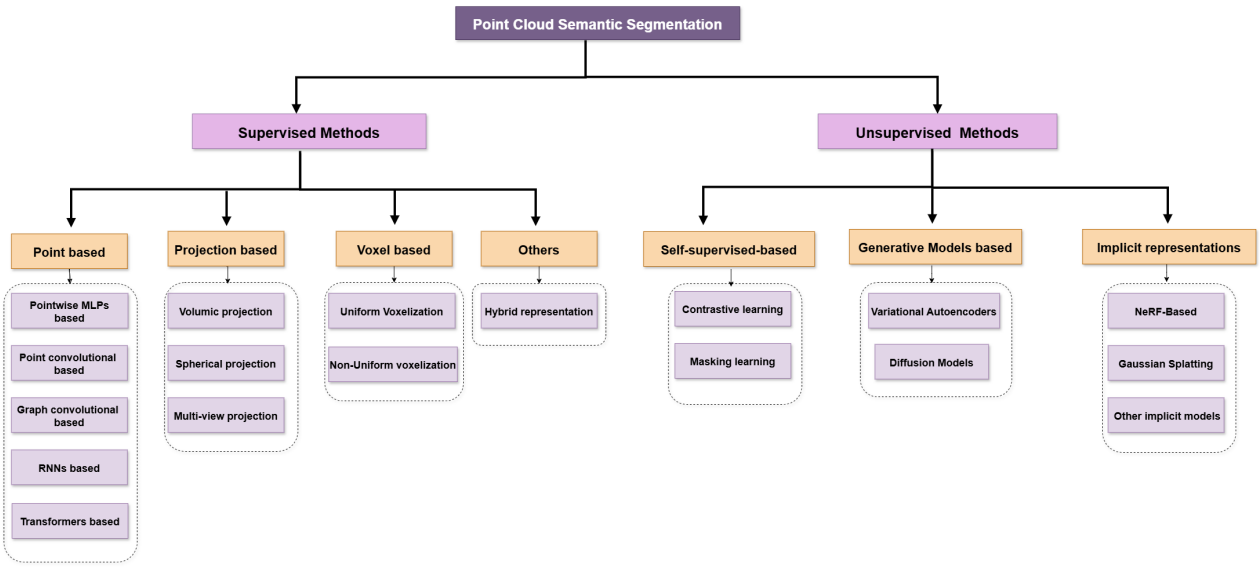


Figure 2. Overview of deep learning methods for 3D point cloud semantic segmentation. The proposed taxonomy is structured into two main categories: (1) supervised methods, including point-based, projection-based, voxel-based, hybrid, and multi-modal fusion approaches; (2) unsupervised methods, encompassing self-supervised techniques and generative models, such as NeRF-based models, Gaussian Splatting, and neural fields.

2. Data Representation

A 3D point cloud can be mathematically represented as a set of 3D points:

$$P_i = \{p_i, i = 1, ..., n\},$$

where each point p_i is a vector of its (x, y, z) coordinates, along with optional additional feature channels such as RGB color, normal vectors, and more. For simplicity and clarity, unless otherwise specified, we consider only the (x, y, z) coordinates as the feature channels for each point. In semantic segmentation tasks, the input can represent a single object for part segmentation or a sub-volume of a 3D scene for object segmentation.

Traditional methods for feature extraction from 3D point clouds typically depend on manually engineered features [33,34]. Due to the irregular and unstructured nature of point clouds, integrating them into deep learning architectures typically requires transformations into alternative formats. Common approaches include converting point clouds into 3D voxel grids or collections of 2D images.

However, these transformations often lead to unnecessarily large datasets and may compromise the data's intrinsic invariance properties. For example, volumetric representations are limited by resolution constraints caused by data sparsity and insufficient information, as well as the high computational cost of 3D convolutions [35]. To address these challenges, various methods have been proposed. For instance, FPNN [36] tackles data sparsity issues but faces scalability challenges when applied to large point clouds. MultiviewCNNs [37] leverage 2D convolutional neural networks (CNNs) by projecting 3D point clouds into multiple 2D views, enabling robust segmentation using the 2D CNN's representational capabilities. This approach leverages the robust representation capabilities of 2D CNNs and utilizes the information from different views of the object. Additionally, feature-based deep neural networks DNNs [38]combine traditional shape descriptors (e.g., curvatures and normals) with deep learning methods for segmentation. By merging handcrafted features with automated deep learning, this hybrid method demonstrates effectiveness in 3D part segmentation.

Several approaches have also been developed to directly process raw 3D point clouds, with segmentation being a particularly computationally intensive task. Consistent with prior studies [29], we categorize 3D semantic segmentation techniques into four groups: point-based, projection-based, voxel-based, and other methods. This taxonomy is illustrated in Figure 2.

3. Methods for 3D Semantic Segmentation

In this field, researchers proposed a wide range of techniques, which are generally categorized into supervised and unsupervised learning paradigms. Supervised approaches rely on annotated data and are typically organized according to the type of input representation used. In contrast, unsupervised techniques do not require labeled data; instead, they automatically identify inherent structures and spatial features within the raw 3D data. By leveraging patterns such as geometric shapes, spatial arrangements, and local feature similarities, these methods aim to segment and classify the 3D environment without manual annotations. This makes unsupervised learning particularly useful when labeled datasets are limited or unavailable.

3.1. Supervised Approaches

3.1.1. Point-Based Methods

PointNets/MLP Methods

MLP-based approaches in 3D semantic segmentation typically involve training a neural network on extensive datasets of 3D volumes and their corresponding labels. These networks learn to predict class labels by analyzing the local context of each point. To enhance the model's performance, preprocessing techniques such as data augmentation or feature normalization are often applied. These techniques can improve the quality of the input data, increase the size of the training dataset, and extract more relevant features [20,39].

PointNet [3], one of the pioneering deep learning models, processes unstructured point clouds directly without requiring voxelization or image projections. The architecture of PointNet employs symmetric functions (e.g., max pooling) to extract global features from point clouds, followed by MLPs to perform classification or segmentation tasks. Despite its simplicity, PointNet captures the permutation invariance of point sets and has demonstrated excellent results on various 3D point cloud tasks, laying the foundation for numerous subsequent models. To address limitations of PointNet in capturing local structures, PointNet++ [40] introduced a hierarchical learning architecture that recursively applies PointNet on nested partitions of the input space. It clusters local regions using a combination of sampling and grouping strategies (such as farthest point sampling and k-nearest neighbors), allowing the network to capture fine-grained geometric patterns across different spatial scales.

Following PointNet++, several works have extended MLP-based methods to improve spatial context encoding. For instance, DGCNN [41] constructs a dynamic graph of nearest neighbors in feature space and applies edge convolutions to better capture geometric relationships. Similarly, PointCNN [42] proposes a xconv operator that learns a canonical ordering of points to apply convolution-like operations. More recent studies such as PointMLP [43] revisit MLPs and propose pure MLP-based networks without attention or convolution mechanisms, achieving competitive results by focusing on affine transformations and residual learning. Moreover, works like PointASNL [44] introduce adaptive sampling and noise-aware learning to further enhance robustness and feature extraction from noisy or sparse point clouds. These developments show that MLPs, when combined with effective neighborhood aggregation and hierarchical design, can remain powerful components in 3D semantic segmentation pipelines.

ConvNet-Based Methods

Convolutional neural network (CNN)-based methods for 3D semantic segmentation leverage structured representations of point clouds, such as voxels or multi-view images, to exploit the spatial locality intrinsic to convolution operations. Early voxel-based approaches, including VoxNet [45], ShapeNet [35], and 3D U-Net [46], transform point clouds into regular 3D grids, enabling the direct application of 3D convolutions. However, the voxelization process often leads to high computational cost and memory inefficiency due to the sparsity of 3D space. To address these limitations, sparse convolutional architectures such as SparseConvNet [47] and MinkowskiNet [6] introduce efficient operations that selectively process occupied voxels, thus allowing for scalable learning on high-resolution inputs.

In parallel, multi-view CNN methods project 3D point clouds onto several 2D views, allowing the use of mature 2D CNN architectures. For example, MVCNN [37] aggregates features across multiple rendered images, while 2D-3D FusionNet [48] combines RGB image features and 3D voxel features in a unified framework to improve semantic understanding. Beyond structured inputs, some CNN-based models operate directly on unstructured point sets. PointConv [49] extends standard convolutions to unordered point clouds by approximating continuous convolution kernels using density and weight functions, ensuring permutation invariance. Similarly, PointSIFT [50] introduces orientation-aware operators to capture fine-grained geometric structures, drawing inspiration from classic 2D descriptors [51]. PointWeb [52] models interactions between neighboring points through adaptive feature adjustment mechanisms, enhancing local context encoding. Another notable model, PointCNN [42], learns to permute and weight points for convolution, exploiting local spatial correlations, although achieving full invariance remains a challenge

Graph-Based Methods

This approach treats the point cloud as a graph, where the points serve as nodes and the edges represent connections between these points. These edges can be based on spatial proximity or other features such as color or texture. A graph convolutional neural network (GCNN) is employed to propagate information from the nodes to edges and vice versa, enabling the network to learn the relationships between different parts of the point cloud and predict the semantic class of each point. ECC [53], one of the earliest methods, extends the scope of edge filters to capture features that elucidate the connections between a point and its neighboring points. Super point Graph (SPG) [54] is a large-scale framework tested on various datasets, containing millions of interior and exterior points. SPG provides a compact yet contextually rich graph representation of the point cloud, highlighting the relationships between different parts of the object.

'One Thing One Click' [55] proposes an annotation approach where, instead of labeling all points, it is sufficient to label only one point per object. This method employs self-learning to model the similarity between nodes in the graph, producing pseudo-labels to guide iterative learning. Following the principles of PointNet [3], DGCN [41] adopts a two-stage approach: first, it learns features to extract super points from the point cloud, and then it uses these super points as the nodes of a graph to perform semantic segmentation. Attention-based GCNs [56,57] use attention mechanisms to weigh the importance of different points in the point cloud, allowing the network to concentrate on the most pertinent parts of the data. Kpconv [58] takes as input ray neighborhoods and treats them with spatially localized weights through a small set of kernel points.

RNN-Based Methods

Recurrent neural networks (RNNs) are a type of deep learning model that is well-suited for processing sequential data. They have been used in various applications, including 3D semantic segmentation. The RNN has the capacity to discern the temporal relationships between the points and utilize this information for the classification or segmentation of the points. In this context, an RNN-based architecture for 3D semantic segmentation generally comprises multiple components, including the following:

- An encoder: Extract features from the 3D point cloud. This can be achieved by converting the point cloud into a voxel grid and applying a 3D CNN to extract features from the voxel grid. Alternatively, it can be accomplished using a PointNet++ architecture, which processes point clouds directly, leveraging a Multi-Layer Perceptron (MLP) to extract features from each point.
- Recurrent layers: These layers process the features extracted by the encoder and learn temporal dependencies between points in the point cloud. This can be performed using a simple RNN, such as an LSTM or GRU, or a more complex architecture like a bidirectional RNN or a multi-layer RNN.
- A decoder: This component predicts the semantic class of each point in the point cloud based on the features and dependencies learned by the recurrent layers. This can be achieved using a simple MLP or a more complex architecture, such as a fully connected Conditional Random Field (CRF) [59].
- A loss function: This component is used to train the model by comparing the predicted semantic class of each point to the ground-truth semantic class. Common loss functions used in 3D semantic segmentation include cross-entropy loss or dice loss.

Finally, the architecture can be fine-tuned with a larger dataset to improve the accuracy of the model. One notable RNN-based architecture for 3D semantic segmentation is PointRNN [60], which models temporal dependencies within point sequences by updating hidden representations associated with individual points. To enhance local context

modeling, a point pyramid pooling module is employed for capturing multi-scale neighborhood features. Building on this, 3P-RNN [61] introduces a point pyramidal pooling module followed by a recurrent neural network to better exploit contextual information at varying spatial densities. Similarly, R-PointNet [62] integrates recurrent structures into the PointNet framework to improve spatial context aggregation. In addition to pure RNN-based models, hybrid architectures that combine recurrent units with 3D convolutional layers such as PointRCNN [63] have demonstrated strong performance in object detection and semantic segmentation. More recently, state-space models have emerged as efficient alternatives to traditional RNNs and transformers. Approaches such as PointMamba [64], PointRWKV [65], and Mamba3D [66] leverage structured sequence modeling to improve segmentation efficiency and scalability. For instance, Point Cloud Mamba (PCM) [67] introduces a serialization technique that preserves spatial locality while converting point clouds into one-dimensional sequences, achieving state-of-the-art results on datasets such as S3DIS. Although RNN-based architectures can be computationally expensive, particularly for large point clouds, and less robust than CNNs when dealing with missing or incomplete data, they are advantageous when temporal context information is important.

Transformer-Based Methods

First introduced in 2017 [68], transformers have revolutionized deep learning by enabling efficient modeling of sequential data through self-attention mechanisms. Unlike recurrent or convolutional neural networks, transformers capture token dependencies in parallel, independent of sequence order. This enables superior performance across various tasks, including computer vision [69], anomaly detection, neural machine translation [70], and text generation. They are also effective at learning hierarchical features from large and complex datasets [23,71].

Transformer architectures consist of stacked layers combining feedforward and self-attention mechanisms, enhanced by residual connections, layer normalization, and positional encoding. Originally developed for natural language processing (NLP), they have demonstrated strong performance in vision tasks as well, often outperforming or matching convolutional neural networks (CNNs) [72–75]. Hybrid models such as TransUNet [76] and DETR [77] exemplify the successful integration of transformer designs into visual understanding.

Their application has recently expanded to unstructured 3D point clouds [78,79], where the self-attention mechanism proves well-suited due to its ability to model spatial dependencies without relying on regular data structures. Transformer-based 3D segmentation models such as PCT [80], Pointformer [81], and Stratified Transformers [82] have further enhanced local feature aggregation and global context understanding.

Recent advances in point cloud segmentation have introduced several improvements to transformer-based architectures. PTV2 [83] incorporates Grouped Vector Attention and Partition-based Pooling, alongside comparative evaluations of FPS-kNN and Grid-kNN strategies, achieving efficient large-scale point cloud processing with high segmentation accuracy. PTV3 [84] simplifies the architecture and enhances computational efficiency through a serialization-based approach, replacing complex attention modules with streamlined processing. It introduces sorting-based serialization and patch interaction strategies (e.g., Shift Dilation, Shift Patch, Shift Order), and employs pre-training, a novel block design, and a mix of Layer and Batch Normalization for improved performance.

A survey on transformer-based visual segmentation [23,71,79,85] categorizes key advances, including vision transformer variants, CNN-transformer hybrids, and self-supervised learning approaches. It reviews meta-architectures and segmentation methods tested on benchmarks.

3.1.2. Projection-Based Methods

Projection-based approaches project 3D point clouds onto 2D planes or modalities, facilitating the use of traditional 2D computer vision techniques. The resulting 2D representation allows for efficient semantic segmentation. Among the most prominent are multi-view, spherical, and volumetric representations, each offering distinct advantages depending on the target application and sensor characteristics.

Multi-View Representation

Multi-view methods generate multiple 2D projections of 3D data from different viewpoints. For instance, MvCNN [37] uses a network to extract features from multiple 2D views of 3D shapes, representing the point cloud as a set of 2D images captured from various angles. Similarly, RangeFormer [86] leverages traditional range-view representations and introduces a Scalable Training from Range-view (STR) strategy, achieving superior segmentation performance.

Spherical Representation

Spherical methods map 3D points onto a sphere's surface or use spherical coordinates. Spherical CNNs [87] generalize traditional convolutions to spherical surfaces, while [88] employs attention mechanisms to focus on critical regions. SqueezeSegV1 [8] introduces an end-to-end pipeline with CNNs and conditional random fields (CRF). Subsequent versions, SqueezeSegV2 [89] and SqueezeSegV3 [90], enhance robustness to noise through a Context Aggregation Module (CAM) and adaptive spatial convolutions (SAC). Rangenet++ [91] combines rotating LiDAR sensors with CNNs for post-processing.

Recent advancements, such as SphereFormer [92], address challenges like data sparsity by introducing radial window self-attention and dynamic feature selection, improving performance on LiDAR datasets.

Volumetric Representation

The methods based on volumetric representations are treated by few researchers because of the lack of detail in each voxel and the difficulty of performing the calculations needed to achieve this in practice. VoxNet [45] is an early method that addresses the problem of exploiting point cloud information. Each point cloud is represented by a grid of voxels. They proposed an architecture that exploits the volumetric shapes of the occupancy grid and integrates them into a supervised 3D convolutional neural network. OctNet [93] enhanced performance by using a hybrid grid and octree structure to partition point clouds hierarchically. Each octree represents a point cloud partition along a regular grid. Based on OctNet, OCNN [94] introduced a method that induces 3D-CNNs to extract features from octrees.

3.1.3. Voxel-Based Methods

Uniform Voxilization

Volumetric representations transform irregular 3D point clouds into structured voxel grids, enabling the application of convolutional operations analogous to those used in 2D image processing. These methods discretize the 3D space into volumetric elements, or voxels, and are broadly categorized into uniform and non-uniform voxelization techniques depending on the spatial resolution and data sparsity [95]. Early approaches such as VoxNet [45] introduced the concept of encoding point clouds into fixed-size voxel grids, followed by 3D convolutional neural networks (3D CNNs) for object classification. However, uniform voxelization leads to exponential memory consumption as spatial resolution increases, making it impractical for large-scale scenes or fine-grained segmentation. To al-

leviate these limitations, SEGCloud [20] proposed a hybrid pipeline combining 3D fully convolutional neural networks (3D-FCNNs) operating on voxelized inputs with a tri-linear interpolation step to project predictions back to the original point cloud. The model further incorporates a fully connected conditional random field (FC-CRF) to refine the segmentation boundaries, addressing the discretization artifacts introduced during voxelization.

Non-Uniform Voxilization

Non-uniform voxelization strategies offer a more efficient representation by adapting voxel resolution to local point density. OctNet [93] introduced an octree-based hierarchical structure to partition the 3D space recursively. This significantly reduces memory usage and computational cost by allocating higher resolution only in regions with dense geometric information. Despite its advantages, OctNet still performs redundant operations in sparsely populated areas. To overcome this, Ref. [47] introduced Sparse Submanifold Convolutions (SSC) through the SparseConvNet framework, which restricts computation strictly to non-empty voxels and avoids the propagation of activations into empty space. This approach set a new standard in 3D semantic segmentation due to its ability to process high-resolution voxel grids efficiently.

Building upon SparseConv, MinkowskiEngine [6] further optimized sparse convolutional operations, supporting both 3D and 4D (spatiotemporal) data with generalized sparse tensors and providing a scalable backbone for modern 3D semantic segmentation models such as SPVCNN [4] and TPVFormer [96].

More recently, hybrid approaches have emerged to combine the strengths of voxel-based and point-based methods. For instance, KPConv [58] leverages deformable kernel point convolutions directly on point clouds while optionally integrating voxel features for global context. Such methods demonstrate that while voxel-based representations offer computational tractability and structural regularity, they benefit from complementary modalities for improved accuracy and efficiency.

3.1.4. Other Representations

Hybrid approaches in 3D semantic segmentation aim to combine the strengths of various learning paradigms, such as CNNs, RNNs, and graph-based models, to improve accuracy and robustness [97,98]. Graph-based methods excel at capturing spatial relationships between points, while CNNs are effective at extracting hierarchical features, making their integration a promising direction for context-aware segmentation. Early works also explored RGB-depth fusion [99], although they often struggled to fully exploit the correlation between depth and color information.

A major advancement in this area lies in multi-modal fusion strategies that combine LiDAR data with high-resolution RGB images to enrich 3D representations with photometric and semantic context. FusionNet [100], adopts a multi-branch architecture incorporating both early and late fusion strategies to align LiDAR and RGB features across different stages of processing, enhancing spatial coherence and semantic consistency. Moreover, 2DPASS (2D Projection-Assisted Semantic Segmentation) [11] further advances this line by establishing dense point-to-pixel correspondences through LiDAR-to-image projection. Its dual-branch encoder independently processes 3D and 2D features, which are then fused via a point-to-pixel alignment module, followed by a cross-modality refinement mechanism that facilitates mutual enhancement. This enables the network to capitalize on the geometric precision of LiDAR and the semantic richness of images, particularly improving performance in challenging regions such as object boundaries.

Similarly, AF2S3Net [101] introduces a unified attention-based framework that combines point-based, voxel-based, and 2D image features. The architecture leverages a 2D-3D

attention module to capture inter-modality dependencies, enabling the network to jointly reason over spatial structure and contextual semantics. By dynamically attending to relevant features across modalities, AF2S3Net significantly boosts segmentation accuracy in complex scenes.

Beyond these, several hybrid models adopt structural fusion strategies. VMNet [102] integrates voxel and mesh representations using both Euclidean and geodesic cues. PV-RCNN++ [103] refines point-voxel representations using image-based contextual features from multiple views. Meta-RangeSeg [104] projects 3D point clouds onto 2D spherical images to facilitate spatio-temporal learning. Encoder–decoder architectures like JS3C [105] and TORNADO-Net [106] also incorporate multi-view projections or semantic priors to guide 3D understanding.

Finally, cross-modal knowledge distillation has emerged as an effective approach for transferring semantic information from pre-trained 2D vision models to 3D domains. Methods such as PointDC [107], ProtoTransfer [108], CrossPoint [109], and Three Pillars [110] exploit this paradigm to boost generalization while reducing reliance on labeled 3D data. Likewise, PartDistill [111] employs vision–language models to enable fine-grained part segmentation, reflecting the growing impact of self-supervised and foundation model-driven fusion approaches.

3.2. Unsupervised Approaches

Unsupervised learning has gained increasing attention in 3D semantic segmentation, driven by the high cost and difficulty of obtaining large-scale annotated point cloud datasets. These methods, which do not rely on human-labeled supervision, can be broadly categorized into self-supervised approaches, generative model-based methods, and implicit representation learning.

3.2.1. Self-Supervised Learning

Self-supervised learning (SSL) has become a crucial paradigm in 3D vision, particularly for domains where labeled data is scarce or expensive to obtain. By formulating pretext tasks that exploit the intrinsic structure of 3D point clouds, these methods enable the learning of semantically meaningful representations without manual annotations. Among the most prominent SSL strategies, contrastive learning has shown exceptional promise. It aims to learn embeddings where different augmentations or spatial transformations of the same point cloud (positive pairs) are pulled together in feature space, while unrelated samples (negative pairs) are pushed apart. A seminal work in this area, PointContrast [2], introduced a framework for contrasting anchor points and their transformed views within and across scans. This method significantly improved the generalization of 3D networks to downstream tasks like semantic segmentation and scene completion by learning both local geometry and global context.

Extending contrastive paradigms to other modalities, DepthContrast [112] proposed contrastive training over depth images, transferring learned representations to LiDAR-based tasks. This cross-modality capability highlights the versatility of SSL in bridging 2D and 3D domains.

Another influential branch of SSL in 3D involves masking-based pretraining, inspired by the success of BERT in natural language processing. These methods introduce structured corruption to the input (e.g., masking out parts of the point cloud) and task the model with reconstructing the missing geometry, thereby encouraging it to capture high-level semantic information. Point-BERT [113] represents a pioneering work in this space, proposing a tokenizer that encodes point cloud patches into discrete tokens using a pretrained DGCNN.

The model then reconstructs masked tokens via a Transformer-based decoder, capturing detailed part-level features.

Following this, Point-MAE [114] further refined this approach by employing a more efficient asymmetric encoder–decoder architecture, improving scalability while preserving reconstruction accuracy. The use of masked auto-encoding has also been extended to multi-view and range-view data, as seen in works like MV-AR [115] a self-supervised pretraining method for LiDAR-based 3D object detection that incorporates voxel and point level masking and reconstruction. It introduces Reversed-Furthest-Voxel-Sampling to handle uneven LiDAR point distributions. Additionally, a new data efficient benchmark on the Waymo dataset is developed to better evaluate pretraining effectiveness.

Moreover, hybrid models like GeoTransformer [116] combine geometric reasoning with masked modeling, demonstrating that incorporating spatial priors during pretraining leads to improved robustness and data efficiency. Across all these directions, the learned representations not only yield strong performance on object classification but also transfer effectively to fine-grained semantic segmentation, scene flow estimation, and 3D object detection validating the generalization power of SSL frameworks.

Taken together, self-supervised strategies whether contrastive, reconstructive, or hybrid have demonstrated that 3D models can learn rich spatial and semantic priors without explicit supervision. These representations are foundational for reducing dependence on annotated data and for advancing practical deployment of 3D vision systems in robotics, autonomous driving, and AR/VR applications.

3.2.2. Generative Methods

Generative models have emerged as a compelling class of unsupervised learning techniques in 3D computer vision, offering the ability to capture and model the complex underlying distributions of geometric data such as point clouds, meshes, and volumetric grids. Among these, Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) were the earliest to be extended from 2D vision to the 3D domain. One of the foundational works, 3D-GAN by [117], demonstrated the feasibility of synthesizing 3D voxelized shapes using adversarial training. This model employs a volumetric convolutional generator and discriminator, trained to reconstruct plausible 3D objects from noise, effectively learning object-level priors without the need for labeled data.

Building upon voxel-based representations, researchers sought to address the inefficiency and sparsity of voxel grids by turning to point cloud-based generative methods. In this context, PointFlow [118] introduced a significant innovation by modeling point clouds as samples from a continuous distribution using normalizing flows. Unlike voxel methods that suffer from quantization artifacts, PointFlow leverages a probabilistic latent space to generate fine-grained and diverse point clouds, demonstrating strong unsupervised capabilities in capturing shape variations.

Recent years have witnessed the rise of diffusion models, which surpass GANs in stability and fidelity, becoming a new paradigm for generative learning. A notable example is Point-E [119], which adapts diffusion modeling to generate sparse 3D point clouds conditioned on complex inputs such as textual prompts or RGB images. Point-E excels in zero-shot generation, offering semantically meaningful shapes without explicit category supervision. Although primarily developed for generative tasks, the pre-trained latent spaces of these models have been shown to be useful for downstream applications, such as semantic segmentation and shape classification, through fine-tuning or by transferring learned geometric priors.

Other generative frameworks, such as VAE-PointNet and 3D-VAE-GAN, combine latent encoding with adversarial losses to enhance shape fidelity and realism. These hybrid

models provide better control over the latent space, enabling shape interpolation, morphing, and attribute manipulation. Additionally, implicit function-based generative models, such as Occupancy Networks [120] and DeepSDF [121], model continuous 3D geometry as a function in space, allowing for high-resolution reconstructions without voxel discretization.

The generative modeling paradigm not only enriches shape synthesis but also advances self-supervised representation learning, where the objective is to learn task-agnostic features from raw 3D data. These learned representations can be fine-tuned with minimal supervision, thus bridging the gap between generation and recognition. As such, the field is witnessing a convergence between generative modeling and discriminative learning in unsupervised 3D vision, with generative models providing powerful priors for a range of downstream tasks, including segmentation, classification, and scene understanding.

3.2.3. Implicit Representation

Implicit representation learning has recently emerged as a powerful alternative to discrete 3D representations (e.g., voxels, point clouds, meshes), offering a continuous and compact means to model complex 3D geometry and semantics. These methods, often parameterized by neural fields, learn to approximate continuous functions that map spatial coordinates (and optionally viewing directions) to scene properties such as occupancy, radiance, signed distance, or semantic class probabilities. Their ability to generalize across varying resolutions and capture fine geometric details makes them particularly attractive for 3D semantic segmentation and scene understanding.

One of the most influential breakthroughs in this field is the Neural Radiance Field (NeRF) [122], which models volumetric scenes by learning a continuous function that maps 3D coordinates and viewing directions to RGB color and density. Although NeRF was initially proposed for novel view synthesis from sparse RGB images, subsequent extensions have introduced semantic supervision into the framework. [123,124] inject semantic information by associating each spatial point not only with radiance but also with a probability distribution over semantic classes, learned through multi-task training or weak labels. These models demonstrate how implicit functions can be adapted for dense semantic inference, even under sparse or noisy supervision.

A related development is Gaussian Splatting [125], which represents a scene not through a neural network, but via a set of 3D anisotropic Gaussians with learned properties such as opacity, position, and orientation. This method achieves real-time rendering performance while maintaining high visual fidelity, and recent adaptations have shown that Gaussian primitives can encode semantic class distributions when supervised appropriately, enabling their use in segmentation and mapping tasks.

Other families of implicit models, such as Occupancy Networks [120] and Signed Distance Function (SDF)-based approaches like DeepSDF [121], further expand the utility of neural fields by modeling binary occupancy or surface boundaries. These models have been used in unsupervised shape completion, multi-object segmentation, and category-level scene decomposition. Their strength lies in their ability to reconstruct continuous surfaces from sparse or partial 3D data, such as single-view scans or incomplete LiDAR frames, making them particularly relevant for autonomous driving and robotic perception where occlusion is prevalent.

Emerging methods like MonoSDF [126] and VolRecon [127] combine implicit geometry learning with volumetric rendering and depth priors from monocular images, showing that neural fields can bridge the gap between 2D and 3D semantic learning. Furthermore, works like CodeNeRF [128] explore latent-conditioned NeRFs, enabling scene-level semantic reasoning across different environments by conditioning the radiance field on latent codes that encode scene-level context.

Collectively, these methods illustrate a growing trend toward neural representations that unify geometry, appearance, and semantics in a compact, resolution-agnostic framework. By leveraging implicit functions, these approaches open new directions for unsupervised and weakly supervised 3D semantic segmentation, offering scalability, memory efficiency, and the potential to learn from incomplete, sparse, or ambiguous 3D data.

4. Benchmark Datasets

Figure 3 presents a variety of widely used 3D semantic segmentation datasets, each reflecting different scene types and sensor modalities. These include indoor datasets like S3DIS [129] and ScanNet [130], outdoor datasets such as Semantic3D [131] and SemanticKITTI [32], and multi-sensor driving data in NuScenes [132]. Together, they cover diverse conditions—indoor/outdoor, static/dynamic, and varying sensor inputs—making them essential benchmarks for evaluating the robustness and generalization of segmentation methods.

The following datasets below are commonly used for evaluating the performance of 3D semantic segmentation models:

S3DIS [129], the Stanford Large-Scale 3D Indoor Spaces (S3DIS) dataset, is a benchmark for 3D semantic segmentation of indoor scenes. It contains 6 large-scale indoor areas, including 271 rooms and more than 6 million square feet of space. Each point in the dataset is labeled with one of 13 semantic classes. This dataset is widely used for evaluating the performance of various 3D semantic segmentation methods. It is one of the most widely used datasets in the field of semantic segmentation, and is frequently used in academic research and industry projects. The dataset is publicly available for research purposes and can be downloaded from the official website. S3DIS is challenging due to its large scale and the variety of objects and layouts present in the scenes. Furthermore, the dataset contains missing data, noise, and other challenges, making it good for evaluating the robustness of various 3D semantic segmentation methods.

The Semantic3D dataset [131] is a point cloud of 4.8 million 3D points acquired from the static terrestrial laser scanner. Each point has 3D coordinates, RGB information, and intensity. Semantic3D includes eight classes of urban and rural areas. It contains scenes like castles, churches, parts of streets, villages, soccer fields, etc. This dataset provides denser point clouds and contains a higher number of labeled points than other datasets. Semantic3D dataset is challenging due to the large scale, high level of detail, and the variety of objects and layouts present in the scenes. Additionally, it is commonly utilized to assess the performance of various 3D semantic segmentation methods and to compare their results with other methods.

The NuScenes dataset [132] contains 1000 scenes, each 20 s long, with data captured at a rate of 20 Hz. The data is annotated with 3D bounding boxes for objects like vehicles, pedestrians, and cyclists. The dataset also includes semantic segmentation labels for the LiDAR point clouds, as well as key points for objects like vehicles and pedestrians.

The PartNet dataset [133] contains over 50,000 3D shapes across 24 object categories, such as chairs, tables, cars, and airplanes, each with over 20 object parts. This dataset has been manually annotated with fine-grained part labels, providing a high-quality benchmark for evaluating 3D semantic segmentation models. The PartNet dataset includes a wide range of object categories and shapes, making it suitable for training models that can generalize to a variety of real-world environments. The fine-grained part labels in the dataset also allow for the evaluation of models on more detailed and specific tasks, such as identifying specific parts of an object. In recent years, the PartNet dataset has been used in various research papers on 3D semantic segmentation and 3D object understanding,

making it a widely accepted benchmark for evaluating the performance of various models and algorithms.

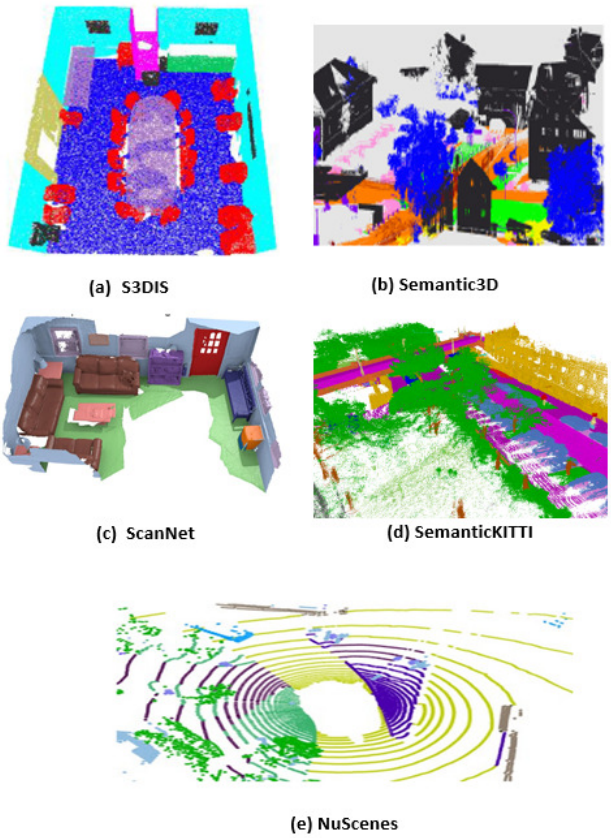


Figure 3. Examples of datasets for semantic segmentation: (**a**) S3DIS [129]; (**b**) Semantic3D [131]; (**c**) ScanNet [130]; (**d**) SemanticKITTI [32]; (**e**) NuScenes [132].

SemanticKITTI [32] is a dataset on a massive scale based on the KITTI [134] vision benchmark. This dataset contains more than 23,000 frames of camera data, along with ground-truth labels for each frame. The dataset includes 19 semantic classes, i.e., 43,000. scans, which are divided into two parts: the sequences from 00 to 10, which allow for the use of semantic interpretation of the scenes, e.g., semantic segmentation; and the rest of the sequences, which are used for testing.

The ModelNet40 dataset [35] is widely used for point cloud treatment due to its diverse categories, clean shapes, and well-constructed synthetic dataset. The dataset contains models of 40 object categories, such as airplanes, chairs, tables, cars, and so on, with a total of 9843 models. Each model is represented by a set of 3D points and is manually annotated with its category label. The dataset is formed by more than 12,000 meshes, and the points of the datasets are taken from the meshed surfaces, after a shift and scaling in a unit sphere. ModelNet is widely used in research on 3D object recognition, 3D shape prediction, and other related tasks.

Other Datasets

There are other datasets of high quality that are not widely used, including ScanNet [130], which includes 1500 scans and 2.5 million images acquired in over 700 distinct locations, and ParisLille-3D [135], which has high quality but is not widely used.

Table 1 highlights a diverse set of benchmark datasets widely adopted for 3D point cloud semantic segmentation, encompassing both real-world and synthetic environments. These datasets differ significantly in terms of acquisition modalities, sensor types, anno-

tation granularity, and application domains ranging from indoor scenes captured with RGB-D sensors to extensive outdoor environments acquired via mobile laser scanning. For indoor segmentation tasks, datasets such as S3DIS and ScanNet are particularly relevant due to their comprehensive coverage of building interiors and rich semantic annotations. Conversely, outdoor segmentation is typically addressed using large-scale datasets like Semantic3D and SemanticKITTI, which offer billions of points and multiple annotated classes such as terrain, vegetation, etc. The selection of a dataset is intrinsically tied to the target application, as each offers distinct contextual features and technical challenges. However, the absence of standardized representations across datasets remains a key limitation, hindering the fair evaluation and application of deep learning models. Establishing consistent benchmarks would not only facilitate more robust comparisons but also accelerate progress in real-world deployments across industrial and urban environments. Despite the high cost of data collection and annotation, publicly available datasets remain indispensable for model training, validation, and reproducibility in the field of 3D semantic segmentation.

Table 1. Representative datasets for 3D point cloud semantic segmentation. Abbreviations: RWE = Real-World Environment, SE = Synthetic Environment, IS = Indoor Segmentation, OS = Outdoor Segmentation, OC = Object Classification, MP = Million Points, TF = Thousand Frames, MLS = Mobile Laser Scanning.

Dataset	Year	Type	Application	Size	Sensor
S3DIS [129]	2016	RWE	IS	273 MP	Matterport
ScanNet [130]	2017	RWE	IS	242 MP	RGB-D
Semantic3D [131]	2017	RWE	OS	4000 MP	MLS
SemanticKITTI [32]	2019	RWE	OS	4549 MP	MLS
NuScenes [132]	2020	RWE	OS	341 TF	Velodyne HDL-32E
ModelNet40 [35]	2015	SE	OC	12.3 TN	-
ScanNet [130]	2017	RWE	IS	242 MP	RGB-D
ParisLille-3D [135]	2018	RWE	OS	1430 MP	MLS

5. Evaluation Metrics

The authors assess the effectiveness of each method to evaluate their scalability to larger and higher-dimensional datasets. To validate the superiority of the proposed methods, several semantic segmentation evaluation metrics are employed. Among these, Mean Intersection over Union (mIoU), Overall Accuracy (OAcc), Mean Accuracy (mAcc), and execution time are the most commonly used metrics to evaluate performance.

- Mean Intersection over Union (mIoU): to check the effectiveness of semantic segmentation methods, mIoU is used as an indicator of the degree of similarity between the intersection and union of two ensembles, it can be calculated mathematically by the ratio of the intersection to the union of two sets: ground truth and predicted output. Given $k+1$ classes, including empty classes, mIoU is quickly calculated as shown by the formula.

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{1}$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} IoU_i \tag{2}$$

where TP, FP, and FN correspond to the number of true positive, false positive, and false negative predictions for each of the k + 1 classes.

- Overall Accuracy (OAcc): Also known as OA, it is a simple metric calculated by the ratio between the correct accuracy of the model and the total number of samples.

$$OAcc_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{3}$$

- Mean Accuracy (mAcc): Represents the average class accuracy, which is a direct per-class application of OAcc, averaged over the total number of classes k.

$$mAcc = \frac{1}{k+1} \sum_{i=0}^{k} OAcc_i \tag{4}$$

- Execution time: The computing time of the algorithms must be seriously considered. With acquisition systems becoming increasingly resolved and accurate, and with large amounts of data being stored, the performance of semantic segmentation methods is a central issue. The authors determine the efficiency of each of their methods in order to evaluate the scalability of their methods to larger and higher-dimensional datasets. But it remains a big challenge due to the use of different datasets, as well as the fact that each method is executed on different performance machines (ram capacity, GPU, etc.).
- Dice Similarity Coefficient (DSC): The Dice Similarity Coefficient (DSC) is a widely used metric for measuring the similarity between two sets of data, particularly in image segmentation tasks. It quantifies the degree of overlap between a predicted segmentation mask and the ground truth mask, providing a robust evaluation of segmentation performance.
  Mathematically, the Dice coefficient is calculated by comparing pixels on a per-pixel basis between the predicted segmentation and the corresponding ground truth. The principle is straightforward: the number of common pixels (i.e., pixels present in both masks) is counted using the intersection of the two images, which is then multiplied by two to account for both sets. This value is then divided by the total number of pixels across both images, ensuring a normalized similarity score between 0 and 1. A DSC of 1 indicates a perfect match, while a score of 0 signifies no overlap.

$$\text{DSC} = \frac{2 * TP_i}{2 * TP_i + FP_i + FN_i} \tag{5}$$

When choosing a metric for 3D semantic segmentation, it is important to consider the specific task and the characteristics of the data. It is also a good idea to consider using multiple metrics to obtain a more comprehensive understanding of the performance of your model. The performance of machine learning methods can vary significantly depending on the task and dataset. Therefore, it is crucial to specify these when comparing methods.

6. Analysis and Discussion

The evolution of 3D semantic segmentation has been strongly shaped by advancements in 2D computer vision, the emergence of specialized neural architectures, and the proliferation of annotated benchmark datasets. This section presents a critical and comparative analysis of state-of-the-art methods, focusing on how foundational vision paradigms have been reinterpreted for 3D spatial data, and how architectural innovations, particularly hybrid networks and multi-sensor fusion, address the challenges of structural complexity, scale variation, and modality diversity. Methodological performance is examined through standardized metrics such as mean accuracy (mAcc), overall accuracy (oAcc), and mean intersection over union (mIoU), using datasets like SemanticKITTI, Seman-

tic3D, and nuScenes. In addition to numerical results, we evaluate how design choices, data representations, and environmental context impact segmentation performance and model transferability. This discussion aims to identify key trends, expose critical limitations, and offer insight into the design trade-offs shaping current research directions in 3D semantic segmentation.

6.1. From 2D to 3D: Transferring Vision Paradigms

Semantic segmentation was initially conceptualized as a dense pixel-wise classification task, with the Fully Convolutional Network (FCN) framework laying the groundwork for later advancements. Over time, innovations in encoder–decoder structures, multi-scale processing [136–139], non-local context modeling [140–142], and precise boundary delineation [143–146] have significantly boosted segmentation accuracy. More recently, transformer-based models [23] have extended this trend by introducing global attention mechanisms, thereby offering a principled alternative to CNN-based prediction heads.

The transition from 2D image segmentation to 3D semantic segmentation represents a significant paradigm shift, reflecting both conceptual and technical advances in scene understanding [25,147]. While 2D convolutional neural networks have performed well on structured image grids, their application to point clouds is hindered by the irregular, non-Euclidean nature of 3D data. Early solutions such as SqueezeSeg [8] sought to bridge this gap by projecting point clouds onto 2D surfaces and leveraging 2D CNN architectures. These approaches evolved into more sophisticated variants, SqueezeSegV2 [89] and SqueezeSegV3 [90], which introduced context modules, adaptive convolutions, and improved normalization to enhance segmentation quality.

Following these initial strategies, hybrid models such as PolarNet [148], RangeNet++ [91], and RandLA-Net [149] integrated 2D projections with point-wise networks like PointNet [3], enabling efficient yet expressive semantic segmentation. While these models benefit from incorporating spatial priors, they are still susceptible to information loss during projection and often struggle to capture intricate local geometries.

To overcome such limitations, subsequent architectures have advanced point-based learning through the development of permutation-invariant convolutions, adaptive neighborhood selection, and context-aware transformations. Notable examples include PointCNN [42], PointSIFT [50], and PointConv [49], each of which preserves the geometric fidelity of raw point clouds while enabling richer feature interactions.

Recent trends have also moved beyond purely architectural innovations, with a growing interest in integrating semantic priors from vision–language models trained on large-scale 2D datasets. Approaches such as CLIP2Scene [5] and CLIP-S^4 [150] adapt pretrained vision–language embeddings to the 3D domain, enabling few-shot or zero-shot semantic understanding. These methods transfer semantic richness from 2D images into point cloud representations, offering notable advantages in terms of generalization and label efficiency. Nevertheless, challenges remain particularly with modality alignment, projection inconsistency, and occlusion handling, underscoring the need for domain-specific adaptations to effectively exploit such cross-modal signals.

6.2. Architectural Integration Techniques : From Hybrid Architectures to Multi-Sensor Fusion in 3D Semantic Segmentation

Hybrid architectures are becoming more popular in 3D semantic segmentation because they combine various data, such as raw point clouds, voxelized volumes, and image-based projections, to improve both detailed local information and overall understanding of the scene. This architectural hybridity is exemplified in several recent models that operationalize such multi-representation strategies to optimize efficiency and accuracy across diverse 3D environments. Models such as FusionNet [100] facilitate efficient neighborhood

searches in large-scale point clouds, while mini PointNet employs voxel-based aggregation through Multi-Layer Perceptrons (MLPs) for compact feature learning. 3D-MiniNet [151] introduces a projection-based module that transmits 3D information to a 2D Fully Convolutional Neural Network (FCNN), effectively leveraging both spatial and semantic priors. SPVNAS [4] uses sparse convolutions that are improved through Neural Architecture Search (NAS) to enhance segmentation performance for high-dimensional 3D video data.

The integration of transformer-based modules into hybrid models has further enhanced the capability to capture both local and global contextual information. Point-BERT [113] employs a Masked Point Modeling (MPM) task to pre-train point cloud transformers, demonstrating significant improvements in downstream tasks and transferability across datasets. Mask3D [152] leverages transformer decoders to directly predict instance masks from 3D point clouds, achieving state-of-the-art performance on benchmarks like ScanNet and S3DIS. SoftGroup [153] introduces a soft grouping mechanism that allows each point to be associated with multiple classes, mitigating errors from hard semantic predictions and improving instance segmentation accuracy.

Despite these notable advancements, several challenges remain in achieving robust generalization across diverse datasets. Many models exhibit performance degradation when applied to datasets with different characteristics, highlighting limited transferability without fine-tuning. Additionally, prevalent issues such as class imbalance and limited scene complexity in existing datasets constrain comprehensive evaluation of model robustness (see Table 2). Furthermore, standard metrics like mean Intersection over Union (mIoU) may inadequately reflect model sensitivity to noise, varying point density, or real-time processing constraints. In response to these challenges, emerging trends such as Gaussian Splatting (GS) [154] offer promising avenues for 3D representation and segmentation. Techniques like Segment Any 3D Gaussians (SAGA) [155] enable real-time, multi-granularity segmentation using 3D Gaussian representations, while FlashSplat [156] provides an optimal solver for 3D-GS segmentation via linear programming. GaussianCut [154] further introduces interactive graph-cut algorithms facilitating user-guided segmentation in complex scenes. Although these approaches demonstrate strong potential, challenges related to memory consumption and semantic integration persist.

Table 2. Comparison of various representations of LiDAR point clouds

View	Complexity	Representative
Raw Points	$\Theta(N.d)$	RandLA-Net [149], KPConv [58]
Range View	$\Theta(\frac{H.W}{r^2}.d)$	SqueezeSeg [8], RangeNet++ [91]
Bird's Eye View	$\Theta(\frac{H.W}{r^2}.d)$	PolarNet [148]
Voxel (Dense)	$\Theta(\frac{H.W.L}{r^3}.d)$	PVCNN [4]
Voxel (Sparse)	$\Theta(N.d)$	MinkowskiNet [6], SPVNAS [4]
Voxel (Cylinder)	$\Theta(N.d)$	Cylinder3D [157]
Multi-View	$\Theta((N+\frac{H.W}{r^2}).d)$	AMVNet [9], RPVNet [10]

Complementing these developments, multi-modal fusion strategies, particularly those combining LiDAR and RGB data [158–162], have gained traction in improving segmentation accuracy. By integrating complementary sensor modalities, models achieve enhanced robustness and precision in complex environments. However, such fusion introduces additional computational overhead and necessitates careful data calibration and synchronization to maintain effectiveness.

6.3. Benchmarking of Algorithms

Exploitation of local features: The extraction of local geometric features plays a vital role in 3D semantic segmentation, particularly for capturing fine details such as object

boundaries and surface variations. Early methods like PointNet [3] relied on shared MLPs and global max pooling but lacked mechanisms to model local neighborhoods. This limitation was addressed in PointNet++ [40], which introduced hierarchical grouping based on spatial proximity to learn local features at multiple scales.

To better capture local structure, several strategies have emerged. k-Nearest neighbor (k-NN) search is widely used to define local regions, enabling networks to incorporate spatial context [163]. Convolution-based models, such as PointConv [49] and KPConv [58], generalize the concept of convolution to point clouds by applying learnable filters over local neighborhoods. Graph-based approaches, such as DGCNN [41] and RandLA-Net [149], model points as graph nodes and exploit dynamic connectivity to enhance local feature learning.

However, local methods often suffer from sensitivity to point density, noise, and occlusion, especially in outdoor environments. Adaptive solutions, such as PAConv [164], mitigate this by learning flexible receptive fields. Meanwhile, hybrid strategies like ShellNet [165] combine local and global features through hierarchical aggregation, offering more robust representations. Ultimately, local features must be integrated with global context to ensure semantic consistency. Recent architectures balance both aspects using hierarchical encoding, dilated convolutions, or transformers, aiming to preserve fine-grained detail while maintaining global coherence.

Density in LiDAR point clouds: LiDAR point clouds often exhibit variable point density, with points near the sensor being denser than those farther away. However, existing methods, such as those proposed in [6,10,11,47,101], do not account for this variability. To address this issue, several methods have been proposed. Point-based methods, such as those in [89–91,166], utilize point features and positions as input and aggregate information from neighbors, providing precise information about each point but potentially being computationally expensive.

View-based methods, such as those in [32,91,166], transform the point cloud into a distance or bird's-eye view to facilitate the extraction of 3D features. Voxel-based methods, such as those in [6,47,167], divide the 3D space into regular voxels and apply sparse convolutions. These methods are often more efficient than point-based methods but can lose information due to voxel resolution.

Three-dimensional semantic segmentation heavily relies on the quality and quantity of available training data. Hence, differences in the performance of the methods mentioned may stem from variations in the datasets used for training and evaluation. All of these methods focus on capturing local information in the LiDAR point cloud, and several approaches have been proposed to enhance their efficiency. For instance, some methods, such as RPVNet and 2DPASS, combine the characteristics of different modalities to improve 3D perception. However, despite incorporating 2D information, these methods fail to match the performance of those proposed in [86,92].

High-quality LiDAR point cloud processing requires good representation, as evidenced by Table 3, where point-based and view-based approaches dominate. These methods typically involve representing the point cloud as a set of points or features in a 3D coordinate system. Voxel-based methods, which divide the point cloud into a regular grid of 3D voxels and represent the occupancy or density of points within each voxel, are also widely used. Additionally, multi-view fusion, which combines information from multiple perspectives or sensor modalities to improve accuracy or completeness, is an effective approach.

Table 3. Reported results for semantic segmentation task on the large-scale outdoor SemanticKITTI benchmark (Note: graph-based (G), hybrid (H), point-based (P), RNN-based (R), transformer-based (T), voxel-based (V)).

Date of Publication	Methods	Mean IoU (%)	View
2023	UniSeg [168]	75.2	P + V + T
2023	SphereFormer [92]	74.8	T
2023	RangeFormer [91]	73.3	P
2022	2DPASS [11]	72.9	H
2022	PTv2 [83]	71.2	T
2022	PVKD [169]	71.2	V
2021	AF2S3Net [101]	70.8	P + V
2020	Cylinder3D [157]	68.9	V
2020	SPVNAS [4]	66.4	P + V
2020	JS3C-Net [105]	66	P + V
2022	GFNet [170]	65.4	P
2020	KPRNet [171]	63.1	H
2017	PointNet++ [40]	20.1	P
2017	SPGraph [54]	17.4	G
2017	PointNet [3]	14.6	P
2023	RangeViT [172]	64.0	T

Challenges and Advances in 3D Segmentation: This review draws upon over 180 peer-reviewed works published between 2017 and 2025, selected based on their citation impact, benchmark performance (SemanticKITTI, nuScenes, S3DIS), and contribution to novel architectures or learning paradigms. Both seminal methods and recent advances were retained to provide a comprehensive yet focused analysis of the field.

To synthesize these findings, Table 4 presents a comparative overview of representative methods, highlighting key trade-offs between segmentation accuracy (mIoU), computational efficiency, noise robustness, real-time applicability, and practical deployment potential. Transformer-based approaches such as UniSeg, SphereFormer, and 2DPASS achieve high mIoU scores often exceeding 72%, demonstrating strong semantic reasoning capabilities. However, their significant computational demands and lack of real-time feasibility limit their use in embedded or time-sensitive scenarios. In contrast, lightweight models like RandLA-Net and PolarNet offer a better balance between accuracy and efficiency, making them more suitable for real-time applications. Hybrid and multi-modal architectures, including AF2S3Net and FusionNet, provide compelling compromises by integrating complementary representations or sensor inputs, albeit at the cost of increased architectural complexity. These insights underscore the importance of designing scalable, resource-efficient models that generalize well across diverse operational contexts.

Three-dimensional semantic segmentation has progressed significantly due to innovations in deep learning, yet several persistent challenges continue to hinder practical deployment. One of the main limitations lies in the restricted generalization of models beyond their training datasets. As shown in Table 2, while architectures such as KPConv or SPVCNN achieve high performance on specific benchmarks, their accuracy tends to drop significantly when applied to different domains or sensor configurations. This issue highlights the limited transferability of most pre-trained models, which often require substantial fine-tuning or re-training to adapt to different sensor setups, environments, or semantic taxonomies.

Table 4. Detailed comparison of the main 3D semantic segmentation methods based on their performance (mIoU), evaluation dataset, computational cost, robustness to noise, ability to operate in real time, and practical applicability.

Method	mIoU (%)	Dataset	Computational Cost	Noise Robustness	Real-Time	Practical Applicability
PointNet [3]	47.6	S3DIS	Low	Low	Yes	Low
PointNet++ [40]	54.5	S3DIS	Moderate	Moderate	No	Moderate
KPConv [58]	70.6	S3DIS	High	High	No	High
RandLA-Net [149]	70.0	SemanticKITTI	Low	Moderate	Yes	High
PointCNN [42]	65.4	S3DIS	High	Moderate	No	Moderate
SPVNAS [4]	66.4	SemanticKITTI	Moderate	High	Yes	High
Cylinder3D [157]	68.9	SemanticKITTI	Moderate	High	No	High
RangeNet++ [91]	65.5	SemanticKITTI	Low	Moderate	Yes	Moderate
PolarNet [148]	54.3	SemanticKITTI	Low	Moderate	Yes	Moderate
SPGraph [54]	62.1	Semantic3D	High	High	No	Moderate
MinkowskiNet [6]	67.0	ScanNet	High	High	No	High
PTV1 [78]	70.4	S3DIS	High	High	No	Moderate
PTV2 [83]	72.7	S3DIS	High	High	No	Moderate
PTV3 [84]	72.3	S3DIS	High	High	No	Moderate
AF2S3Net [101]	70.8	SemanticKITTI	Moderate	High	Yes	High
2DPASS [11]	72.9	SemanticKITTI	High	High	No	High
SphereFormer [92]	74.8	SemanticKITTI	High	High	No	High
UniSeg [168]	75.2	SemanticKITTI	High	High	No	High
FusionNet [100]	66.3	SemanticKITTI	Moderate	High	No	High
SalsaNext [173]	59.5	SemanticKITTI	Moderate	Moderate	Yes	Moderate
RangeViT [172]	64.0	SemanticKITTI	High	Moderate	No	Moderate
FlashSplat [156]	71.0	ScanNet	Moderate	High	Yes	High
GaussianCut [154]	69.5	ScanNet	High	High	No	Moderate
Mask3D [152]	74.0	ScanNet	High	High	No	High
Point-BERT [113]	66.6	S3DIS	High	High	No	Moderate
SPFormer [174]	54.9	ScanNet	High	Moderate	No	Low
LatticeNet [175]	64.0	ScanNet	High	Moderate	No	Moderate
SegCloud [20]	61.3	Semantic3D	Moderate	Moderate	No	Moderate
OctNet [93]	50.7	Semantic3D	High	Moderate	No	Low
VMNet [102]	52.9	SemanticKITTI	Moderate	Moderate	No	Moderate

Point-based methods exploit raw 3D geometry without intermediate representation, maintaining fidelity but suffering from computational inefficiencies due to neighborhood constructions such as kNN. Their high sensitivity to density and scale variations makes large-scale deployment difficult without sophisticated partitioning strategies. Meanwhile, voxel-based methods offer regular data structures for convolutional processing but incur significant memory costs due to voxel quantization and sparsity. SparseConvNet and KPConv have optimized memory locality and feature resolution, but the need for GPU memory and training time remains a barrier for embedded or real-time systems.

Hybrid models have emerged as a promising trade-off (see Table 5). Architectures like 2DPASS combine point-based and voxel-based pipelines with 2D projections to leverage both fine-grained geometry and contextual RGB information. Hybrid models that integrate transformer and convolutional components, such as AF2S3Net, enhance long-range contextual understanding while preserving local spatial detail. However, these improvements often require complex training pipelines and substantial computational resources.

Multi-modal fusion strategies, particularly those integrating LiDAR and RGB, enhance segmentation accuracy by combining geometric and visual cues. However, they require careful calibration and are sensitive to misalignment and increased computational complexity, which limits their scalability in dynamic environments.

The emergence of neural implicit models such as NeRF and Gaussian Splatting introduces a paradigm shift. Originally designed for view synthesis and reconstruction, these methods are being extended to semantic segmentation by learning continuous 3D fields. Though still experimental, they offer high fidelity and data continuity, bypassing the limitations of voxel or point discretizations. Nevertheless, challenges remain regarding the incorporation of semantic labels, training stability, and inference speed for real-time tasks.

Table 5. Performance assessment of projected images, point clouds segmentation methods for semantic representation. on S3DIS (Area5), ScanNet (6-fold), Semantic3D (reduced-8) and SemanticKITTI (xyz only). Note: the '-' symbol means that the results are not available. (mAcc: average class accuracy, oAcc: overall accuracy, mIoU: mean Intersection-over-Union).

Methods	S3DIS			ScanNet			Semantic3D			SemanticKITTI		
	mAcc	oAcc	mIoU	mAcc	oAcc	mIoU	mAcc	oAcc	mIoU	mAcc	oAcc	mIoU
SqueezeSeg [8]	-	-	-	-	-	-	-	-	-	-	-	-
RangNet++ [91]	-	-	-	-	-	-	-	-	-	-	-	-
OctNet [93]	39	68.9	26.3	26.4	76.6	18.1	71.3	80.7	50.7	-	-	-
SegCloud [20]	-	-	-	-	-	-	88.1	61.3	-	-	-	-
RangNet53 [20]	-	-	-	-	-	-	88.1	61.3	-	-	52.2	-
MI-Net [176]	-	-	-	-	-	-	-	-	-	-	-	-
TangentConv [177]	62.2	82.5	52.8	55.1	80.1	40.9	80.7	89.3	66.4	-	-	40.9
PointNet [3]	67.1	81.0	54.5	-	-	-	-	-	-	-	-	20.1
PointNet++ [40]	67.1	81.0	54.5	-	-	-	-	-	-	-	-	20.1
PointWeb [52]	76.2	87.3	66.7	-	-	-	-	-	-	-	-	-
PointSIFT [50]	-	-	70.23	70.2	-	41.5	-	-	-	-	-	-
RSNet [178]	66.5	-	56.5	-	-	-	-	-	-	-	-	-
KPConv [58]	79.1	-	70.6	-	-	-	-	92.9	74.6	-	-	-
PointCNN [42]	75.6	88.1	65.4	-	-	-	-	-	-	-	-	-
PointConv [49]	-	-	-	-	-	-	-	-	-	-	-	-
RandLA-Net [149]	82.0	88.0	70.0	-	-	-	-	94.8	77.4	-	-	53.9
PolarNet [148]	-	-	-	-	-	-	-	-	-	-	-	54.3
DGCN [41]	-	-	-	-	-	-	-	-	-	-	-	-
SPG [54]	73.0	85.5	62.1	-	-	-	-	94.0	73.2	-	-	17.4
SPLATNet [179]	-	-	-	-	-	39.3	-	-	-	-	-	18.4
LatticeNet [175]	-	-	-	-	-	64.0	-	-	-	-	-	52.9
VMNet [102]	-	-	-	-	-	-	64.0	-	-	-	-	52.9
LaserMix [180]	-	-	73.0	-	-	61.4	-	-	-	-	-	60.8
PTV1 [78]	90.8	76.5	70.4	-	-	-	-	-	-	-	-	-
PTV2 [83]	91.6	78.0	72.7	-	-	-	-	-	-	-	-	-
PTV3 [84]	78.4	91.4	72.3	-	-	-	-	-	-	-	-	-
PointTransformer [78]	76.5	90.8	70.4	-	-	-	-	-	-	-	-	-

In terms of learning paradigms, supervised training remains dominant but increasingly impractical due to annotation cost and domain specificity. Semi-supervised and self-supervised learning strategies, including contrastive learning like PointContrast, have emerged as scalable alternatives. These methods leverage unlabelled data and auxiliary tasks to learn transferable features, facilitating open-set segmentation and continual learning. However, their robustness in highly cluttered or unstructured scenes is still under investigation.

Dataset limitations further complicate the landscape. Most public datasets exhibit strong class imbalance, with overrepresented static classes (roads, buildings, etc.) and scarce dynamic ones (pedestrians, cyclists, etc). This skew impacts model evaluation and real-world applicability. Moreover, current datasets rarely span diverse geographic regions, weather conditions, or sensor types. Few benchmarks account for multi-modal data, temporal dynamics, or application-specific scenarios like cultural heritage or climate monitoring. These gaps hinder progress on generalization, domain adaptation, and cross-sensor robustness.

Finally, deployment constraints must be considered. Methods optimized for accuracy often lack real-time feasibility or require specialized hardware. While transformer models achieve state-of-the-art results, they are resource intensive. Efficient models via pruning, quantization, and knowledge distillation must be pursued to enable embedded

applications. Lightweight architectures such as MobileNet, ShuffleNet, and EfficientNet, though underexplored in 3D contexts, show promise for constrained environments.

To guide future research, several open challenges and questions must be addressed. How can 3D semantic segmentation models be made more robust to cross-domain variations in datasets and sensor modalities? Are hybrid architectures, which combine voxel, point, and image-based representations, a scalable solution, or do they introduce excessive complexity and computational overhead? Which methods achieve the most favorable trade-off between segmentation accuracy, inference speed, and memory efficiency, especially for real-time or embedded systems? Finally, how can segmentation systems be better aligned with the evolving requirements of real-world applications such as autonomous driving, mobile robotics, and large-scale urban mapping?

7. Conclusions and Future Work

This paper provides a comprehensive survey of recent advances in 3D semantic segmentation based on deep learning techniques. We first categorize existing methods into point-based, voxel-based, projection-based, hybrid, and transformer-based approaches. A comparative analysis across these categories is conducted using standardized benchmarks and evaluation metrics, supported by detailed quantitative tables. Furthermore, we assess the role of dataset characteristics and representation strategies in shaping model performance.

Our findings highlight that semantic segmentation remains a critical component in 3D vision applications such as autonomous driving, robotics, and urban mapping. However, the irregular and sparse nature of point clouds poses fundamental challenges in feature extraction, representation learning, and scalability. Additionally, the high dimensionality of 3D data results in substantial computational overhead during both training and inference. Incorporating multiple representations and sensor modalities can mitigate these issues and lead to more efficient and accurate models.

Promising future research directions include the following: (i) improving segmentation accuracy through more expressive architectures; (ii) enabling scalable learning on large-scale, real-world datasets; (iii) incorporating rich geometric and appearance cues (e.g., texture, color, reflectance); (iv) developing lightweight and real-time algorithms suitable for embedded systems; (v) advancing multi-modal fusion techniques (e.g., LiDAR-RGB-D) for robust performance in diverse environments; (vi) addressing scene complexity and occlusion to enhance generalization and applicability in dynamic and cluttered scenarios.

In general, advancing 3D semantic segmentation will require not only algorithmic innovation but also better benchmark diversity, efficient deployment strategies, and deeper integration with upstream and downstream perception tasks.

Author Contributions: Conceptualization, H.B., N.A.S., and J.R.; methodology, H.B., N.A.S., and J.R.; validation, J.R. and N.A.S.; formal analysis, H.B., N.A.S., and J.R.; investigation, J.R. and N.A.S.; resources, H.B.; data curation, H.B.; writing—original draft preparation, H.B.; writing—review and editing, J.R. , N.A.S., A.A., and H.T.; visualization, H.B.; supervision, H.B., J.R., N.A.S., and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are sourced from multiple datasets, including SemanticKITTI, nuScenes, and Semantic3D, which are openly available to the research community. 1. SemanticKITTI: The SemanticKITTI dataset can be accessed by visiting the official SemanticKITTI

website (https://www.semantic-kitti.org/, accessed on 10 July 2025). Researchers can follow the provided instructions to request access and download the dataset. SemanticKITTI comprises high-resolution LiDAR point clouds and camera images collected from an autonomous vehicle in various urban scenes. 2. nuScenes: The nuScenes dataset is available for researchers to access by visiting the official nuScenes website (https://www.nuscenes.org/, accessed on 10 July 2025). Instructions on how to request access and download the dataset can be found on the website. The nuScenes dataset includes diverse sensor data such as LiDAR point clouds, camera images, radar data, and metadata, collected from autonomous vehicles. 3. Semantic3D: The Semantic3D dataset can be obtained by visiting the official Semantic3D website (http://www.semantic3d.net/, accessed on 10 July 2025). Researchers can follow the provided instructions to download the dataset. Semantic3D consists of large-scale LiDAR point cloud data captured in urban and rural areas, along with corresponding semantic annotations. The availability and usage of these datasets are subject to the terms and conditions set forth by their respective creators.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

References

1. Bello, S.A.; Yu, S.; Wang, C.; Adam, J.M.; Li, J. Review: Deep Learning on 3D Point Clouds. *Remote Sens.* **2020**, *12*, 1729. [CrossRef]
2. Xie, S.; Gu, J.; Guo, D.; Qi, C.R.; Guibas, L.; Litany, O. PointContrast: Unsupervised Pre-training for 3D Point Cloud Understanding. In *Computer Vision–ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, **2020**; Volume 12348, pp. 574–591. [CrossRef]
3. 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**. [CrossRef]
4. Tang, H. ; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; Han, S. Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution. In *Computer Vision–ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland , 2020; Volume 12373, pp. 685–702. [CrossRef]
5. Chen, R. ; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; Wang, W. CLIP2Scene: Towards Label-efficient 3D Scene Understanding by CLIP. *arXiv* **2023**, arXiv:2301.04926. [CrossRef]
6. Choy, C.; Gwak, J.; Savarese, S. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3070–3079. [CrossRef]
7. Aksoy, E.E.; Baci, S.; Cavdar, S. Salsanet: Fast Road and Vehicle Segmentation in Lidar Point Clouds for Autonomous Driving. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 926–932. [CrossRef]
8. 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. [CrossRef]
9. Liong, V.E.; Nguyen, T.N.T.; Widjaja, S.; Sharma, D.; Chong, Z.J. AMVNet: Assertion-based Multi-View Fusion Network for LiDAR Semantic Segmentation. *arXiv* **2020**, arXiv:2012.04934. [CrossRef]
10. Xu, J.; Zhang, R.; Dou, J.; Zhu, Y.; Sun, J.; Pu, S. RPVNet: A Deep and Efficient Range-Point-Voxel Fusion Network for LiDAR Point Cloud Segmentation. *arXiv* **2021**, arXiv:2103.12978. [CrossRef]
11. Yan, X.; Gao, J.; Zheng, C.; Zheng, C.; Zhang, R.; Cui, S.; Li, Z. 2DPASS: 2D Priors Assisted Semantic Segmentation on LiDAR Point Clouds. In *Computer Vision–ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer Nature: Cham, Switzerland, 2022; Volume 13688, pp. 677–695. [CrossRef]
12. Roldão, L.; De Charette, R.; Verroust-Blondet, A. 3D Semantic Scene Completion: A Survey. *Int. J. Comput. Vis.* **2022**, *130*, 1978–2005. [CrossRef]
13. 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, 17–24 June 2023; pp. 21736–21746. [CrossRef]
14. Aleotti, J.; Caselli, S. A 3D Shape Segmentation Approach for Robot Grasping by Parts. *Robot. Auton. Syst.* **2012**, *60*, 358–366. [CrossRef]
15. Mo, K.; Guerrero, P.; Yi, L.; Su, H.; Wonka, P.; Mitra, N.; Guibas, L.J. StructureNet: Hierarchical Graph Networks for 3D Shape Generation. *arXiv* **2019**, arXiv:1908.00575. [CrossRef]

16. Kareem, A.; Lahoud, J.; Cholakkal, H. PARIS3D: Reasoning-Based 3D Part Segmentation Using Large Multimodal Model. In *Computer Vision–ECCV 2024*; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Springer Nature: Cham, Switzerland, 2025; Volume 15130, pp. 466–482. [CrossRef]
17. Zhou, F.; Zhang, Q.; Zhu, H.; Liu, S.; Jiang, N.; Cai, X.; Qi, Q.; Hu, Y. Attentional Keypoint Detection on Point Clouds for 3D Object Part Segmentation. *Appl. Sci.* **2023**, *13*, 12537. [CrossRef]
18. Zhou, T.; Wang, W. Prototype-Based Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 6858–6872. [CrossRef]
19. Yasir, S.M.; Ahn, H. Deep Learning-Based 3D Instance and Semantic Segmentation: A Review. *arXiv* **2024**, arXiv:2406.13308. [CrossRef]
20. 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. [CrossRef]
21. Landrieu, L.; Simonovsky, M. Segmentation Sémantique à Grande Echelle Par Graphe de Superpoints. In Proceedings of the RFIAP, Paris, France, 26–28 June 2018. Available online: https://hal.science/hal-01939229v1 (accessed on 10 July 2025).
22. Nguyen, A.; Le, B. 3D Point Cloud Segmentation: A Survey. In Proceedings of the 2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM), Manila, Philippines, 12–15 November 2013; pp. 225–230. [CrossRef]
23. Li, X. ; Ding, H.; Yuan, H.; Zhang, W.; Pang, J.; Cheng, G.; Chen, K.; Liu, Z.; Loy, C.C. Transformer-Based Visual Segmentation: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10138–10163. [CrossRef] [PubMed]
24. Grilli, E.; Menna, F.; Remondino, F. A Review of Point Clouds Segmentation and Classification Algorithms. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 339. [CrossRef]
25. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv* **2017**, arXiv:1704.06857. [CrossRef]
26. Yu, H.; Yang, Z.; Tan, L.; Wang, Y.; Sun, W.; Sun, M.; Tang, Y. Methods and Datasets on Semantic Segmentation: A Review. *Neurocomputing* **2018**, *304*, 82–103. [CrossRef]
27. Lateef, F.; Ruichek, Y. Survey on Semantic Segmentation Using Deep Learning Techniques. *Neurocomputing* **2019**, *338*, 321–348. [CrossRef]
28. Xie, Y.; Tian, J.; Zhu, X.X. Linking Points With Labels in 3D: A Review of Point Cloud Semantic Segmentation. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 38–59. [CrossRef]
29. Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; Bennamoun, M. Deep Learning for 3D Point Clouds: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 4338–4364. [CrossRef]
30. Vodrahalli, K.; Bhowmik, A.K. 3D Computer Vision Based on Machine Learning with Deep Neural Networks: A Review. *J. Soc. Inf. Disp.* **2017**, *25*, 676–694. [CrossRef]
31. Ioannidou, A.; Chatzilari, E.; Nikolopoulos, S.; Kompatsiaris, I. Deep Learning Advances in Computer Vision with 3D Data: A Survey. *ACM Comput. Surv.* **2018**, *50*, 1–38. [CrossRef]
32. 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9296–9306. [CrossRef]
33. Hackel, T.; Wegner, J.D.; Schindler, K. Fast Semantic Segmentation of 3D Point Clouds with Strongly Varying Density. *Isprs Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2016**, *III-3*, 177–184. [CrossRef]
34. Chua, C.S.; Jarvis, R. Point Signatures: A New Representation for 3d Object Recognition. *Int. J. Comput. Vis.* **1997**, *25*, 63–85. [CrossRef]
35. 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 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1912–1920. [CrossRef]
36. Li, Y.; Pirk, S.; Su, H.; Qi, C.R.; Guibas, L.J. FPNN: Field Probing Neural Networks for 3D Data. *arXiv* **2016**, arXiv:1605.06240. [CrossRef]
37. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-View Convolutional Neural Networks for 3D Shape Recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 945–953. [CrossRef]
38. Guo, K.; Zou, D.; Chen, X. 3D Mesh Labeling via Deep Convolutional Neural Networks. *ACM Trans. Graph.* **2015**, *35*, 1–12. [CrossRef]
39. Uy, M.A.; Pham, Q.H.; Hua, B.S.; Nguyen, D.T.; Yeung, S.K. Revisiting Point Cloud Classification: A New Benchmark Dataset and Classification Model on Real-World Data. *arXiv* **2019**, arXiv:1908.04616. [CrossRef]
40. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *arXiv* **2017**, arXiv:1706.02413. [CrossRef]

41. 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]
42. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. Pointcnn: Convolution on x-Transformed Points. *arXiv* **2018**, arXiv:1801.07791. [CrossRef]
43. 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:2003.00492. [CrossRef]
44. Yan, X.; Zheng, C.; Li, Z.; Wang, S.; Cui, S. PointASNL: Robust Point Clouds Processing using Nonlocal Neural Networks with Adaptive Sampling. *arXiv* **2020**, arXiv:2003.00492. [CrossRef]
45. 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–2 October 2015; pp. 922–928. [CrossRef]
46. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *arXiv* **2016**, arXiv:1606.06650. [CrossRef]
47. Graham, B.; Engelcke, M.; Maaten, L.V.D. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9224–9232. [CrossRef]
48. Kundu, A.; Yin, X.; Fathi, A.; Ross, D.; Brewington, B.; Funkhouser, T.; Pantofaru, C. Virtual Multi-view Fusion for 3D Semantic Segmentation. *arXiv* **2020**, arXiv:2007.13138. [CrossRef]
49. Wu, W.; Qi, Z.; Fuxin, L. Pointconv: Deep Convolutional Networks on 3d Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9621–9630. [CrossRef]
50. 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. [CrossRef]
51. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
52. Zhao, H.; Jiang, L.; Fu, C.W.; Jia, J. PointWeb: Enhancing Local Neighborhood Features for Point Cloud Processing. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5560–5568. [CrossRef]
53. 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. [CrossRef]
54. Landrieu, L.; Simonovsky, M. Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4558–4567. [CrossRef]
55. Liu, Z.; Qi, X.; Fu, C.W. 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, Nashville, TN, USA, 20–25 June 2021; pp. 1726–1736. [CrossRef]
56. Xie, Z.; Chen, J.; Peng, B. Point Clouds Learning with Attention-Based Graph Convolution Networks. *Neurocomputing* **2020**, *402*, 245–255. [CrossRef]
57. Wang, L.; Huang, Y.; Hou, Y.; Zhang, S.; Shan, J. Graph Attention Convolution for Point Cloud Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10296–10305. [CrossRef]
58. Thomas, H.; Qi, C.R.; Deschaud, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. Kpconv: Flexible and Deformable Convolution for Point Clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6411–6420. [CrossRef]
59. Tseng, H.; Chang, P.; Andrew, G.; Jurafsky, D.; Manning, C. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In Proceedings of the Fourth SIGHAN Workshop on Chinese language Processing, Jeju Island, Republic of Korea, 14–15 October 2005. Available online: https://aclanthology.org/I05-3027/ (accessed on 10 July 2025).
60. Fan, H.; Yang, Y. PointRNN: Point Recurrent Neural Network for Moving Point Cloud Processing. *arXiv* **2019**, arXiv:1910.08287. [CrossRef]
61. Ye, X.; Li, J.; Huang, H.; Du, L.; Zhang, X. 3D Recurrent Neural Networks with Context Fusion for Point Cloud Semantic Segmentation. In *Computer Vision–ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 415–430. [CrossRef]
62. Yi, L.; Zhao, W.; Wang, H.; Sung, M.; Guibas, L. GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud. *arXiv* **2018**, arXiv:1812.03320. [CrossRef]
63. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3D Object Proposal Generation and Detection from Point Cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 770–779.

64. Liang, D.; Zhou, X.; Xu, W.; Zhu, X.; Zou, Z.; Ye, X.; Tan, X.; Bai, X. PointMamba: A Simple State Space Model for Point Cloud Analysis. *arXiv* **2024**, arXiv:2402.10739. [CrossRef]
65. He, Q.; Zhang, J.; Peng, J.; He, H.; Li, X.; Wang, Y.; Wang, C. PointRWKV: Efficient RWKV-Like Model for Hierarchical Point Cloud Learning. *arXiv* **2024**, arXiv:2405.15214. [CrossRef]
66. Han, X.; Tang, Y.; Wang, Z.; Li, X. Mamba3D: Enhancing Local Features for 3D Point Cloud Analysis via State Space Model. *arXiv* **2024**, arXiv:2404.14966. [CrossRef]
67. Zhang, T.; Yuan, H.; Qi, L.; Zhang, J.; Zhou, Q.; Ji, S.; Yan, S.; Li, X. Point Cloud Mamba: Point Cloud Learning via State Space Model. *arXiv* **2024**, arXiv:2403.00762. [CrossRef]
68. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
69. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021. [CrossRef]
70. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
71. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *arXiv* **2021**, arXiv:2101.01169. [CrossRef]
72. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [CrossRef]
73. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. *arXiv* **2021**, arXiv:2104.13840. [CrossRef]
74. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12114–12124. [CrossRef]
75. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542. [CrossRef]
76. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306. [CrossRef]
77. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229. [CrossRef]
78. Zhao, H. ; Jiang, L.; Jia, J.; Torr, P.; Koltun, V. Point Transformer. *arXiv* **2020**, arXiv:2012.09164 . [CrossRef] [PubMed]
79. Lu, D. ; Xie, Q.; Wei, M.; Gao, K.; Xu, L.; Li, J. Transformers in 3D Point Clouds: A Survey. *arXiv* **2022**, arXiv:2205.07417. [CrossRef]
80. 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]
81. Pan, X.; Xia, Z.; Song, S.; Li, L.E.; Huang, G. 3D Object Detection with Pointformer. *arXiv* **2020**, arXiv:2012.11409. [CrossRef]
82. Lai, X.; Liu, J.; Jiang, L.; Wang, L.; Zhao, H.; Liu, S.; Qi, X.; Jia, J. Stratified Transformer for 3D Point Cloud Segmentation. *arXiv* **2022**, arXiv:2203.14508. [CrossRef]
83. Wu, X. ; Lao, Y.; Jiang, L.; Liu, X.; Zhao, H. Point Transformer V2: Grouped Vector Attention and Partition-based Pooling. *arXiv* **2022**, arXiv:2210.05666. [CrossRef]
84. Wu, X.; Jiang, L.; Wang, P.S.; Liu, Z.; Liu, X.; Qiao, Y.; Ouyang, W.; He, T.; Zhao, H. Point Transformer V3: Simpler, Faster, Stronger. *arXiv* **2023**, arXiv:2312.10035. [CrossRef]
85. Lahoud, J.; Cao, J.; Khan, F.S.; Cholakkal, H.; Anwer, R.M.; Khan, S.; Yang, M.H. 3D Vision with Transformers: A Survey. *arXiv* **2022**, arXiv:2208.04309. [CrossRef]
86. Kong, L.; Liu, Y.; Chen, R.; Ma, Y.; Zhu, X.; Li, Y.; Hou, Y.; Qiao, Y.; Liu, Z. Rethinking Range View Representation for LiDAR Segmentation. *arXiv* **2023**, arXiv:2303.05367. [CrossRef]
87. Cohen, T.S.; Geiger, M.; Koehler, J.; Welling, M. Spherical CNNs. *arXiv* **2018**, arXiv:1801.10130. [CrossRef]
88. Wang, Y.; Shi, T.; Yun, P.; Tai, L.; Liu, M. PointSeg: Real-Time Semantic Segmentation Based on 3D LiDAR Point Cloud. *arXiv* **2018**, arXiv:1807.06288. [CrossRef]
89. 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. *arXiv* **2018**, arXiv:1809.08495. [CrossRef]
90. Xu, C.; Wu, B.; Wang, Z.; Zhan, W.; Vajda, P.; Keutzer, K.; Tomizuka, M. SqueezeSegV3: Spatially-Adaptive Convolution for Efficient Point-Cloud Segmentation. In *Computer Vision–ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12373, pp. 1–19. [CrossRef]

91. 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. [CrossRef]
92. Lai, X.; Chen, Y.; Lu, F.; Liu, J.; Jia, J. Spherical Transformer for LiDAR-based 3D Recognition. *arXiv* **2023**, arXiv:2303.12766. [CrossRef]
93. Riegler, G.; Ulusoy, A.O.; Geiger, A. OctNet: Learning Deep 3D Representations at High Resolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6620–6629. [CrossRef]
94. 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]
95. He, Y.; Yu, H.; Liu, X.; Yang, Z.; Sun, W.; Anwar, S.; Mian, A. Deep Learning Based 3D Segmentation: A Survey. *arXiv* **2021**, arXiv:2103.05423. [CrossRef]
96. Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; Lu, J. Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction (TPVFormer). *arXiv* **2023**, arXiv:2302.07817. [CrossRef]
97. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223. [CrossRef]
98. Tallamraju, R.; Rajappa, S.; Black, M.J.; Karlapalem, K.; Ahmad, A. Decentralized Mpc Based Obstacle Avoidance for Multi-Robot Target Tracking Scenarios. In Proceedings of the 2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), Philadelphia, PA, USA, 6–8 August 2018; pp. 1–8. [CrossRef]
99. Couprie, C.; Farabet, C.; Najman, L.; LeCun, Y. Indoor Semantic Segmentation Using Depth Information. *arXiv* **2013**, arXiv:1301.3572. [CrossRef]
100. Zhang, F.; Fang, J.; Wah, B.; Torr, P. Deep FusionNet for Point Cloud Semantic Segmentation. In *Computer Vision–ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12369, pp. 644–663. [CrossRef]
101. Cheng, R.; Razani, R.; Taghavi, E.; Li, E.; Liu, B. $(AF)^2$ -S3Net: Attentive Feature Fusion with Adaptive Feature Selection for Sparse Semantic Segmentation Network. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12542–12551. [CrossRef]
102. Hu, Z.; Bai, X.; Shang, J.; Zhang, R.; Dong, J.; Wang, X.; Sun, G.; Fu, H.; Tai, C.L. Vmnet: Voxel-mesh Network for Geodesic-Aware 3d Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15488–15498. [CrossRef]
103. Shi, S.; Jiang, L.; Deng, J.; Wang, Z.; Guo, C.; Shi, J.; Wang, X.; Li, H. PV-RCNN++: Point-Voxel Feature Set Abstraction with Local Vector Representation for 3D Object Detection. *Int. J. Comput. Vis.* **2023**, *131*, 531–551. [CrossRef]
104. Wang, S.; Zhu, J.; Zhang, R. Meta-RangeSeg: LiDAR Sequence Semantic Segmentation Using Multiple Feature Aggregation. *IEEE Robot. Autom. Lett.* **2022**, *7*, 9739–9746. [CrossRef]
105. Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; Cui, S. Sparse Single Sweep LiDAR Point Cloud Segmentation via Learning Contextual Shape Priors from Scene Completion. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 3101–3109. [CrossRef]
106. Gerdzhev, M.; Razani, R.; Taghavi, E.; Bingbing, L. TORNADO-Net: mulTiview tOtal vaRiatioN semAntic Segmentation with Diamond inceptiOn Module. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 9543–9549. [CrossRef]
107. Chen, Z.; Xu, H.; Chen, W.; Zhou, Z.; Xiao, H.; Sun, B.; Xie, X.; Kang, W. PointDC:Unsupervised Semantic Segmentation of 3D Point Clouds via Cross-modal Distillation and Super-Voxel Clustering. *arXiv* **2024**, arXiv:2304.08965. [CrossRef]
108. Tang, P.; Xu, H.M.; Ma, C. ProtoTransfer: Cross-Modal Prototype Transfer for Point Cloud Segmentation. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 3314–3324. [CrossRef]
109. Afham, M.; Dissanayake, I.; Dissanayake, D.; Dharmasiri, A.; Thilakarathna, K.; Rodrigo, R. CrossPoint: Self-Supervised Cross-Modal Contrastive Learning for 3D Point Cloud Understanding. *arXiv* **2022**, arXiv:2203.00680. [CrossRef]
110. Puy, G.; Gidaris, S.; Boulch, A.; Siméoni, O.; Sautier, C.; Pérez, P.; Bursuc, A.; Marlet, R. Three Pillars improving Vision Foundation Model Distillation for Lidar. *arXiv* **2023**, arXiv:2310.17504. [CrossRef]
111. Umam, A.; Yang, C.K.; Chen, M.H.; Chuang, J.H.; Lin, Y.Y. PartDistill: 3D Shape Part Segmentation by Vision-Language Model Distillation. *arXiv* **2023**, arXiv:2312.04016. Version Number: 2. [CrossRef]
112. Zhang, Q.; Litany, O.; Choy, C.; Koltun, V. Self-Supervised Pretraining of 3D Features on Any Point-Cloud. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021. [CrossRef]
113. Yu, X. ; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; Lu, J. Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022. *arXiv* **2022**, arXiv:2111.14819. [CrossRef]

114. Pang, Y.; Liu, W.; Lin, C.Z.; Yu, X.; Zheng, Y.; Gong, M.; Li, Y. Masked Autoencoders for Point Cloud Self-supervised Learning. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022. [CrossRef]
115. Xu, R.; Wang, T.; Zhang, W.; Chen, R.; Cao, J.; Pang, J.; Lin, D. MV-JAR: Masked Voxel Jigsaw and Reconstruction for LiDAR-Based Self-Supervised Pre-Training. *arXiv* **2023**, arXiv:2303.13510. [CrossRef]
116. Qin, Z.; Yu, H.; Wang, C.; Guo, Y.; Peng, Y.; Ilic, S.; Hu, D.; Xu, K. GeoTransformer: Fast and Robust Point Cloud Registration with Geometric Transformer. arXiv **2023**, arXiv:2308.03768. [CrossRef]
117. Wu, J.; Zhang, C.; Xue, T.; Freeman, W.T.; Tenenbaum, J.B. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Barcelona, Spain, 5–10 December 2016. [CrossRef]
118. Yang, G.; Chen, X.; Zhao, H.; Xu, Z.; Zhang, L.; Huang, J.; Huang, J.; Niethammer, M.; Fidler, S. PointFlow: 3D Point Cloud Generation with Continuous Normalizing Flows. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
119. Nichol, A.; Dhariwal, P.; Chan, W.; Ramachandran, A. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv* **2022**, arXiv:2212.08751. [CrossRef]
120. Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; Geiger, A. Occupancy Networks: Learning 3D Reconstruction in Function Space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
121. Park, J.J.; Florence, P.; Straub, J.; Newcombe, R.; Lovegrove, S. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
122. Mildenhall, B.; Srinivasan, P.; Tancik, M.; Barron, J.; Ramamoorthi, R.; Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020. [CrossRef]
123. Zhi, S.; Laidlow, T.; Leutenegger, S.; Davison, A.J. In-Place Scene Labelling and Understanding with Implicit Scene Representation. *arXiv* **2021**, arXiv:2103.15875. [CrossRef]
124. Wang, R.; Zhang, S.; Huang, P.; Zhang, D.; Yan, W. Semantic Is Enough: Only Semantic Information for NeRF Reconstruction. In Proceedings of the 2023 IEEE International Conference on Unmanned Systems (ICUS), Beijing, China, 13 October 2023; pp. 906–912. [CrossRef]
125. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph. (TOG)* **2023**, *42*, 1–13. [CrossRef]
126. Zheng, Y.; Li, X.; Xu, J.; Zhu, Y.; Jiang, Y.; Wang, Y.; Zhang, Z.; Bao, H.; Wang, G. MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023. [CrossRef]
127. Xu, H.; Chen, W.; Zhang, Z.; Su, H.; Yu, F. VolRecon: Volume Rendering of Signed Ray Distance Fields for Generalizable Multi-view 3D Reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
128. Jang, W.; Agapito, L. CodeNeRF: Disentangled Neural Radiance Fields for Object Categories. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 12929–12938. [CrossRef]
129. 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1534–1543. [CrossRef]
130. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2432–2443. [CrossRef]
131. Hackel, T.; Savinov, N.; Ladicky, L.; Wegner, J.D.; Schindler, K.; Pollefeys, M. Semantic3D.Net: A New Large-scale Point Cloud Classification Benchmark. *Isprs Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2017**, *IV-1/W1*, 91–98. [CrossRef]
132. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11618–11628. [CrossRef]
133. Mo, K.; Zhu, S.; Chang, A.X.; Yi, L.; Tripathi, S.; Guibas, L.J.; Su, H. Partnet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3d Object Understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 909–918. [CrossRef]
134. 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. [CrossRef]

135. 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]
136. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Context Contrasted Feature and Gated Multi-Scale Aggregation for Scene Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2393–2402. [CrossRef]
137. Shuai, B.; Ding, H.; Liu, T.; Wang, G.; Jiang, X. Toward Achieving Robust Low-Level and High-Level Scene Parsing. *IEEE Trans. Image Process.* **2019**, *28*, 1378–1390. [CrossRef]
138. Li, X.; Zhao, H.; Han, L.; Tong, Y.; Tan, S.; Yang, K. Gated Fully Fusion for Semantic Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11418–11425. [CrossRef]
139. Li, X.; Zhang, L.; Cheng, G.; Yang, K.; Tong, Y.; Zhu, X.; Xiang, T. Global Aggregation Then Local Distribution for Scene Parsing. *IEEE Trans. Image Process.* **2021**, *30*, 6829–6842. [CrossRef]
140. Zhang, L.; Li, X.; Arnab, A.; Yang, K.; Tong, Y.; Torr, P.H.S. Dual Graph Convolutional Network for Semantic Segmentation. *arXiv* **2019**, arXiv:1909.06121. [CrossRef]
141. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. *arXiv* **2017**, arXiv:1711.07971. [CrossRef]
142. Yuan, Y.; Chen, X.; Chen, X.; Wang, J. Segmentation Transformer: Object-Contextual Representations for Semantic Segmentation. *arXiv* **2019**, arXiv:1909.11065. [CrossRef]
143. Kirillov, A.; Wu, Y.; He, K.; Girshick, R. Pointrend: Image Segmentation as Rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. [CrossRef]
144. Ding, H.; Jiang, X.; Liu, A.Q.; Thalmann, N.M.; Wang, G. Boundary-Aware Feature Propagation for Scene Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Glasgow, UK, 23–28 August 2019; pp. 6818–6828. [CrossRef]
145. Li, X.; Li, X.; Zhang, L.; Cheng, G.; Shi, J.; Lin, Z.; Tan, S.; Tong, Y. Improving Semantic Segmentation via Decoupled Body and Edge Supervision. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 435–452. [CrossRef]
146. He, H.; Li, X.; Cheng, G.; Shi, J.; Tong, Y.; Meng, G.; Prinet, V.; Weng, L. Enhanced Boundary Learning for Glass-like Object Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 15839–15848. [CrossRef]
147. 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]
148. Zhang, Y.; Zhou, Z.; David, P.; Yue, X.; Xi, Z.; Gong, B.; Foroosh, H. PolarNet: An Improved Grid Representation for Online LiDAR Point Clouds Semantic Segmentation. *arXiv* **2020**, arXiv:2003.14032. [CrossRef]
149. 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. [CrossRef]
150. He, W.; Jamonnak, S.; Gou, L.; Ren, L. CLIP-S^4: Language-Guided Self-Supervised Semantic Segmentation. *arXiv* **2023**, arXiv:2305.01040. [CrossRef]
151. Alonso, I.; Riazuelo, L.; Montesano, L.; Murillo, A.C. 3D-MiniNet: Learning a 2D Representation From Point Clouds for Fast and Efficient 3D LIDAR Semantic Segmentation. *IEEE Robot. Autom. Lett.* **2020**, *5*, 5432–5439. [CrossRef]
152. Schult, J.; Engelmann, F.; Hermans, A.; Litany, O.; Tang, S.; Leibe, B. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. *arXiv* **2022**, arXiv:2210.03105. [CrossRef]
153. Vu, T.; Kim, K.; Luu, T.M.; Nguyen, X.T.; Yoo, C.D. SoftGroup for 3D Instance Segmentation on Point Clouds. *arXiv* **2022**, arXiv:2203.01509. [CrossRef]
154. Jain, U.; Mirzaei, A.; Gilitschenski, I. GaussianCut: Interactive segmentation via graph cut for 3D Gaussian Splatting. *arXiv* **2024**, arXiv:2411.07555. [CrossRef]
155. Cen, J.; Fang, J.; Yang, C.; Xie, L.; Zhang, X.; Shen, W.; Tian, Q. Segment Any 3D Gaussians. *arXiv* **2023**, arXiv:2312.00860. [CrossRef]
156. Shen, Q.; Yang, X.; Wang, X. FlashSplat: 2D to 3D Gaussian Splatting Segmentation Solved Optimally. *arXiv* **2024**, arXiv:2409.08270. [CrossRef]
157. Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Li, W.; Ma, Y.; Li, H.; Yang, R.; Lin, D. Cylindrical and Asymmetrical 3D Convolution Networks for LiDAR-Based Perception. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 6807–6822. [CrossRef]
158. Madawy, K.E.; Rashed, H.; Sallab, A.E.; Nasr, O.; Kamel, H.; Yogamani, S. RGB and LiDAR fusion based 3D Semantic Segmentation for Autonomous Driving. *arXiv* **2019**, arXiv:1906.00208. [CrossRef]
159. Zhang, X.; He, L.; Chen, J.; Wang, B.; Wang, Y.; Zhou, Y. Multiattention Mechanism 3D Object Detection Algorithm Based on RGB and LiDAR Fusion for Intelligent Driving. *Sensors* **2023**, *23*, 8732. [CrossRef] [PubMed]

160. Dimitrovski, I.; Spasev, V.; Kitanovski, I. Deep Multimodal Fusion for Semantic Segmentation of Remote Sensing Earth Observation Data. *arXiv* **2024**, arXiv:2410.00469. [CrossRef]
161. Liu, S.; Zhang, J.; Wang, B.; Liu, J. A Cross-Modal Feature Fusion Model Based on ConvNeXt for RGB-D Semantic Segmentation. *Mathematics* **2023**, *11*, 1828. [CrossRef]
162. Li, J.; Xiao, J.; Lu, Q.; Xu, Q.; Deng, J.; He, Y.; Zheng, N. Deep Learning for LiDAR-Only and LiDAR-Fusion 3D Perception: A Survey. *Intell. Robot.* **2021**, *1*, 171–190. [CrossRef]
163. Li, Y.; Chen, H.; Cui, Z.; Timofte, R.; Pollefeys, M.; Chirikjian, G.; Van Gool, L. Towards Efficient Graph Convolutional Networks for Point Cloud Handling. *arXiv* **2021**, arXiv:2104.05706. [CrossRef]
164. Xu, M.; Ding, R.; Zhao, H.; Qi, X. PAConv: Position Adaptive Convolution with Dynamic Kernel Assembling on Point Clouds. *arXiv* **2021** , arXiv:2103.14635. [CrossRef]
165. Zhang, Z.; Hua, B.S.; Yeung, S.K. ShellNet: Efficient Point Cloud Convolutional Neural Networks using Concentric Shells Statistics. *arXiv aa* **2019**, arXiv:1908.06295. [CrossRef]
166. Razani, R.; Cheng, R.; Taghavi, E.; Bingbing, L. Lite-Hdseg: Lidar Semantic Segmentation Using Lite Harmonic Dense Convolutions. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp.9550–9556.
167. Graham, B.; van der Maaten, L. Submanifold Sparse Convolutional Networks. *arXiv* **2017**, arXiv:1706.01307. [CrossRef]
168. Liu, Y.; Chen, R.; Li, X.; Kong, L.; Yang, Y.; Xia, Z.; Bai, Y.; Zhu, X.; Ma, Y.; Li, Y.; et al. Uniseg: A Unified Multi-Modal Lidar Segmentation Network and the Openpcseg Codebase. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 21662–21673. [CrossRef]
169. Hou, Y.; Zhu, X.; Ma, Y.; Loy, C.C.; Li, Y. Point-to-Voxel Knowledge Distillation for Lidar Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8479–8488. [CrossRef]
170. Qiu, H.; Yu, B.; Tao, D. GFNet: Geometric Flow Network for 3D Point Cloud Semantic Segmentation. *arXiv* **2022**, arXiv:2207.02605. [CrossRef]
171. Kochanov, D.; Nejadasl, F.K.; Booij, O. KPRNet: Improving Projection-Based LiDAR Semantic Segmentation. *arXiv* **2020**, arXiv:2007.12668. [CrossRef]
172. Ando, A.; Gidaris, S.; Bursuc, A.; Puy, G.; Boulch, A.; Marlet, R. RangeViT: Towards Vision Transformers for 3D Semantic Segmentation in Autonomous Driving. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 5240–5250. [CrossRef]
173. Cortinhal, T.; Tzelepis, G.; Erdal Aksoy, E. SalsaNext: Fast, Uncertainty-Aware Semantic Segmentation of LiDAR Point Clouds. In *Advances in Visual Computing*; Bebis, G., Yin, Z., Kim, E., Bender, J., Subr, K., Kwon, B.C., Zhao, J., Kalkofen, D., Baciu, G., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12510, pp. 207–222. [CrossRef]
174. Sun, J.; Qing, C.; Tan, J.; Xu, X. Superpoint Transformer for 3D Scene Instance Segmentation. *arXiv* **2022**, arXiv:2211.15766. [CrossRef]
175. Rosu, R.A.; Schütt, P.; Quenzel, J.; Behnke, S. LatticeNet: Fast Point Cloud Segmentation Using Permutohedral Lattices. *arXiv* **2019**, arXiv:1912.05905. [CrossRef]
176. Li, S.; Chen, X.; Liu, Y.; Dai, D.; Stachniss, C.; Gall, J. Multi-Scale Interaction for Real-Time LiDAR Data Segmentation on an Embedded Platform. *IEEE Robot. Autom. Lett.* **2022**, *7*, 738–745. [CrossRef]
177. Tatarchenko, M.; Park, J.; Koltun, V.; Zhou, Q.Y. Tangent Convolutions for Dense Prediction in 3d. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3887–3896. [CrossRef]
178. 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. [CrossRef]
179. 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. [CrossRef]
180. Kong, L.; Ren, J.; Pan, L.; Liu, Z. LaserMix for Semi-Supervised LiDAR Semantic Segmentation. *arXiv* **2023**, arXiv:2207.00026. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.