Deep Learning For Information Fusion In Point Cloud Semantic Segmentation For Autonomous Driving: A Survey

Zaipeng Duan[1,2], Xuzhong Hu[1,2], Junfeng Ding[1,2], Wu Lv[3], Ruixiang Zhao[3], Delie Ming[1] and Jie Ma[1,2*]

[1*]School of Huazhong University of Science and Technology, Wuhan, 430074, China.
[2*]National Key Laboratory of Science and Technology on Multispectral Information Processing, Wuhan, 430074, China.
[3]CSSC MARINE TECHNOLOGY CO, Beijing, 100070, China.

*Corresponding author(s). E-mail(s): majie@hust.edu.cn;
Contributing authors: d202181001@hust.edu.cn;
D202281111@hust.edu.cn; djfenghust@hust.edu.cn;
18911990785@163.com; zhaoruixiang12@126.com;
mingdelie@hust.edu.cn;

Abstract

Autonomous driving has made rapid progress in recent years. Under complex and dynamic driving scenarios, autonomous vehicles widely employ a variety of sensors, such as LiDAR and cameras, to integrate diverse information and harness complementary features, ensuring robust and precise environmental perception. However, so far there has been no crucial review of deep learning-based information fusion for point cloud semantic segmentation. To fill this gap and motivate future research, we comprehensively evaluate recent progress in deep learning-based point cloud semantic segmentation with information fusion in autonomous driving. We introduce an innovative classification method, which categorizes these methods into three major groups based on the fusion of information between different representations within a single sensor and multi-sensor information fusion. Furthermore, we conduct a systematic review of existing research results and comparative analyses of their performance

on benchmark datasets. Finally, we summarize the current challenges in point cloud semantic segmentation and offer insights into future directions and trends.

Keywords: Autonomous driving, Point cloud semantic segmentation, Deep learning, Artificial intelligence

1 Introduction

The rapid rise of autonomous driving technology has triggered a revolutionary change in various domains such as transportation, traffic safety, and mobility. However, to achieve autonomous driving, intelligent vehicles must achieve highly accurate environmental perception in highly dynamic environments. Environmental perception, as a fundamental component of autonomous driving technology, is crucial for the interaction of autonomous vehicles [1] and intelligent robots [2–5] with the external environment. Specifically, it enables intelligent driving vehicles to better mimic the perception of human drivers, improving the safety of autonomous cars [6–9]. In this context, point cloud semantic segmentation has become a critical technology in autonomous driving. It aims to assign each point in 3D point cloud data to its corresponding semantic categories, such as road, pedestrian, traffic signs, and others. This task not only aids vehicles in better understanding their surroundings but also enhances the accuracy and safety of decision-making in autonomous driving systems.

In recent years, deep neural networks have gradually dominated the field of point cloud semantic segmentation due to their powerful feature extraction capabilities. In autonomous driving, early deep learning-based 3D semantic

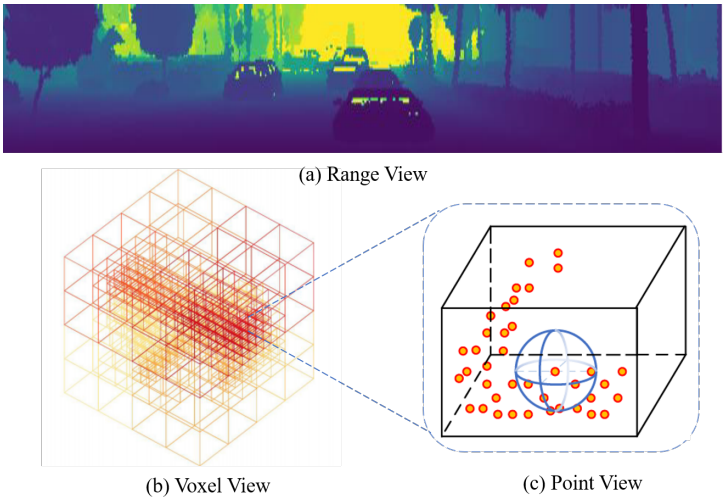


Fig. 1 Traditional semantic segmentation methods.

segmentation methods can be categorized into three fundamental representations: 1) Point-based methods [10–13]; 2) Projection-based methods [14–17]; 3) Voxel-based methods [18–21] as shown in Fig 1. However, due to the complexity of autonomous driving scenarios, single-representation semantic segmentation struggles to ensure stable perception across all situations. To fully leverage the strengths of different representations and mitigate their respective weaknesses in fine-grained segmentation tasks, the trend has shifted towards information fusion in semantic segmentation. Specifically, point-based methods take each point from the raw data as input and perform semantic segmentation by aggregating point cloud features within their local neighborhoods using point-wise MLP (Multilayer Perceptron) or point convolution methods. While these methods preserve 3D geometric features, they rely on costly neighborhood point lookups and computations. Projection-based methods project the point cloud onto different viewing planes and perform segmentation through 2D image processing. Although this reduces computational costs, it may lead to the loss of geometric information. Voxel-based methods transform point clouds into voxel data and apply 3D convolutions for segmentation. Despite their significant effectiveness, these methods have higher storage and computational costs. Furthermore, point cloud density varies with distance in autonomous driving scenarios, making it challenging for existing algorithms to segment distant targets. On the other hand, RGB images provide texture and contextual information that complements point cloud data. By combining the information from point clouds and RGB images, this approach can synergize their strengths and mitigate their individual limitations.

Currently, there have been some reviews [22–27] that summarize and analyze deep learning-based point cloud semantic segmentation. Nevertheless, in recent years, semantic segmentation of point clouds has made significant progress by incorporating information fusion, especially in the context of autonomous driving. It's worth noting that most of the previously mentioned reviews predominantly concentrate on 3D semantic segmentation methods that rely on single information. Until now, there has been a noticeable lack of a comprehensive review that systematically encompasses the latest progress in information fusion-driven point cloud semantic segmentation, with a specific focus on its applicability to autonomous driving. Therefore, in this paper, considering the importance of sensor information fusion, we categorize some representative methods into three classes: methods based on multi-view fusion, methods based on multi-modal representation fusion, and methods based on multi-sensor fusion. We summarize their strengths and weaknesses, compare the overall performance of the representative methods, and discuss outstanding issues, and future perspectives. We believe that this review is beneficial for the perception modules in autonomous driving. In sum up, the contributions of this paper are listed as follows:

1. We introduce an innovative classification framework for information fusion in point cloud semantic segmentation for autonomous driving perception

tasks, which includes methods based on multi-view fusion, methods based on multi-modal representation fusion, and methods based on multi-sensor fusion.

2. In this paper, we organize and review the methods based on information fusion approaches and perform a comparative analysis of representative methods.

3. We conduct a detailed analysis of the remaining challenges and propose several potential research directions for information fusion-based semantic segmentation methods. These findings provide inspiration for future research efforts.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of semantic segmentation information fusion, including commonly used data formats (point and image) as inputs for downstream tasks, as well as widely used open-access datasets and evaluation metrics. Unlike the image branch, the LiDAR branch has multiple representations. Section 3 provides a detailed description of information fusion methods, introducing an innovative classification that categorizes all current fusion efforts into three main categories that differ from traditional methods. Section 4 conducts an in-depth analysis of the existing challenges in the field of point cloud semantic segmentation and outlines future research directions. Finally, Section 5 summarizes the contents of this paper.

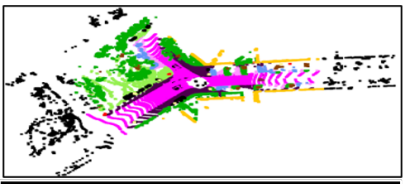
第二节给定了语义分割综述，第三节给定信息融合方法，第四届给定了现在的点云分割挑战

2 BACKGROUND

In this section, we first introduce the representations of point and image data as shown in Fig 2. For the image branch, most methods tend to preserve the original format of the data that serves as input to downstream modules [28]. However, the point branch is highly data format dependent [29], and different representations can significantly influence the design of downstream modules. These representations typically fall into categories such as unstructured point cloud, point cloud projection images, and voxel-based representations. Following that, we will discuss commonly used open datasets and evaluation metrics.

2.1 Image Representation

As the most commonly used data acquisition sensor in 2D and 3D object detection and semantic segmentation tasks, a monocular camera provides RGB images with rich texture information [30, 31]. Specifically, these images typically have a resolution of $H * W$, where H represents the height and W represents the width. Each image pixel (u, v), has a multi-channel feature vector $f(u, v) = (\mathrm{R}, \mathrm{G}, \mathrm{B}, ...)$, which typically contains color components captured by the camera in the red, green, and blue channels or other manually designed features. However, due to the lack of depth information, performing semantic segmentation directly on objects in 3D space is challenging.



	RGB Image	Point Cloud
Permutation	Ordered	Orderless
Data Type	Discrete	Continuous
Dimension	2D	3D
Data Structure	Regular	Irregular
Resolution	High	Low

Fig. 2 Comparison between image representation and point cloud representation.

2.2 Point Cloud Representation

In 3D point cloud semantic segmentation, point cloud data is commonly represented in three main forms: unstructured point cloud, point cloud projection images, and voxel-based representations.

2.2.1 Unstructured Point Cloud

In autonomous driving, the acquisition of point cloud data primarily relies on light detection and ranging (LiDAR) technology. LiDAR is an active sensor that measures the time-of-flight (ToF) of emitted laser beams and calculates the distance between objects and the sensor, generating sparse three-dimensional point cloud data. Point cloud data is a large collection of three-dimensional information points that can express the distribution and surface characteristics of the target space in the same spatial reference frame. Each point carries its own coordinate position and intensity information (x, y, z, i). To process LiDAR data effectively, some methods directly perform point-based feature extraction and aggregation [10, 11]. However, in autonomous driving scenarios, point cloud data can be massive, and unorganized point representations may suffer from redundancy or speed limitations. Therefore, many researchers attempt to convert point clouds into voxel or two-dimensional projection representations for downstream processing.

2.2.2 Point Cloud Projection Images

Significant advances in deep learning for semantic image segmentation have prompted researchers to explore the transformation of three-dimensional point clouds to the image plane and subsequently process the data using Convolutional Neural Network (CNN) models. Specifically, they attempt to project

LiDAR data into two common types of representations in the image space, namely Range View (RV) and Bird's Eye View (BEV). Given the coordinate transformation matrix and camera intrinsic matrix between LiDAR and the camera, we can map 3D points (x, y, z) to pixel points (u, v) through a spherical transformation. RV, which has a similar format to camera images, is often used as an additional channel for feature fusion. Unlike directly projecting LiDAR information onto the frontal view image space, BEV mapping provides an elevated view of the scene from above. In semantic segmentation, these are commonly applied in multi-task scenarios involving object detection and semantic segmentation. Although projection-based methods achieve a trade-off between accuracy and computational cost by reducing the dimensionality and computational complexity of point clouds, this intermediate representation inevitably introduces discretization errors and occlusion issues, resulting in a loss of geometric information.

2.2.3 Voxel-based representations

Some work utilizes 3D CNNs by discretizing 3D space into 3D voxels, where the input is a point cloud P within the range of $L_x * L_y * L_z$, transformed into regular voxels with a resolution of H*W*D. The resolution is controlled by the voxel parameters s=(s_x, s_y, s_z) (length/width/height), where (H, W, D) = $(L_x/s_x, L_y/s_y, L_z/s_z)$. Recent work may consider a more reasonable discretization method, namely the voxelization of a cylindrical volume, significantly reducing the redundancy of unstructured point clouds. Although it allows the use of 3D sparse convolution techniques [32], all points within the voxel grid are considered as the same class, which is not conducive to the semantic segmentation of small and adjacent objects.

2.3 Challenges In Information Fusion

Despite significant advances in point-based, projection-based, and voxel-based methods, there are several challenges in information fusion between point cloud data, spherical projection images, and voxel data: 1) Mismatched Data Resolutions: Point cloud data may have non-uniform distributions, leading to inconsistent resolutions. The resolution of spherical projection images depends on point density on the sphere's surface, while voxel representation's resolution is determined by voxel size. 2) Inconsistent Data Dimensions: Point cloud collections are three-dimensional data, with each point containing three coordinates and intensity. Spherical projection images are two-dimensional and contain color and similar information, while voxel representation can be seen as a three-dimensional data structure. 3) Data Misalignment: Due to the limited horizontal angular resolution of LiDAR, there can be a many-to-one relationship between adjacent points and pixels during spherical projection [33].

Furthermore, in the complex and dynamic environments of autonomous driving, a single sensor may struggle to provide consistent perception across

all scenarios. The fusion of LiDAR and cameras can enhance environmental perception. Specifically, three-dimensional point clouds contain structural information but lack texture details, while RGB images offer rich color and texture information but lack structural cues. However, there are substantial modality differences between RGB images and point clouds: 1) Scale Disparity: Point cloud data and images represent object sizes differently. Point cloud coordinates typically represent positions in actual space, whereas pixel values in images do not directly correspond to real-world sizes. 2) Data Density Difference: Point cloud data is often sparse, with neighboring points having significant distances between them, even in high-density scenarios. In contrast, image data can have rich pixel information, resulting in higher data density. 3) Feature Dimensionality Mismatch: Point cloud data typically has higher feature dimensions compared to image color channels and textures. Point clouds can carry additional information such as intensity, reflectance, normals, etc., while image features mainly include color and texture.

These challenges need to be addressed to effectively integrate and utilize the complementary information from LiDAR and cameras for semantic segmentation in autonomous driving applications.

2.4 Evaluation Metrics

To evaluate the performance of 3D semantic segmentation algorithms, it is necessary to use common objective evaluation metrics to ensure the fairness of algorithm evaluation. The experimental performance evaluation criteria for semantic segmentation algorithms mainly include the following aspects: accuracy, time complexity, and space complexity. Accuracy is the most critical metric among them. Most existing literature uses the mean Intersection over Union (mIoU) as the evaluation metric for semantic segmentation results. The specific calculation is shown in Equation 1. Assuming $k+1$ is the number of predicted categories (including a background class), i represents the true value, j represents the predicted value, and P_{ij} represents the number of points predicted as j when the true value is i.

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ij} - p_{ii}} \tag{1}$$

Complexity is another important metric for evaluating model performance, including both temporal and spatial complexity. With the rapid development of point cloud semantic segmentation technology and the increase in data computing power, the models used in this technology have become increasingly complex. In fact, tasks such as pedestrian detection and autonomous driving require real-time and efficient semantic segmentation networks. Therefore, real-time performance (runtime) and spatial complexity (number of network parameters) requirements are essential.

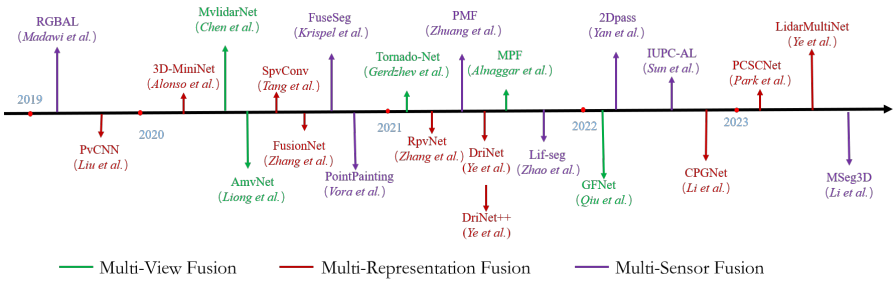


Fig. 3 Timeline of information fusion in point cloud semantic segmentation.

2.5 Open-access Datasets

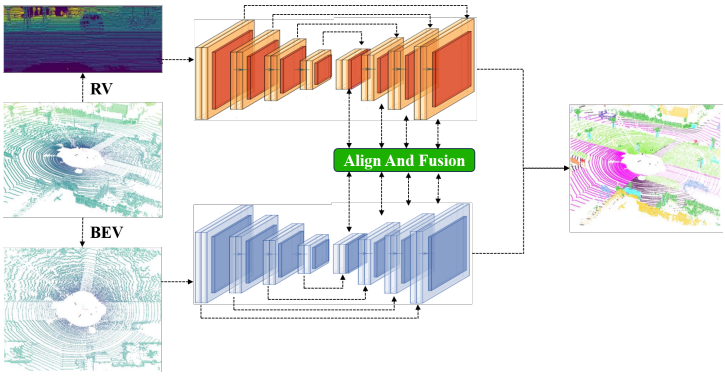
To validate the effectiveness of algorithms proposed by researchers for semantic segmentation of point clouds, a large and reliable dataset is essential. With the development of deep learning in 3D semantic segmentation, the role of 3D datasets has become increasingly important. Currently, to promote research in 3D point cloud semantic segmentation, many research institutions have provided some reliable and open 3D datasets such as the SemanticKITTI dataset [34] and the nuScenes dataset [35].

SemanticKITTI: This dataset was developed by a research group from the University of Bonn, Germany, in 2019. It is a large-scale dataset of outdoor scenes based on LiDAR sensors. The SemanticKITTI dataset consists of 21 sequences from different scenes, totaling 43552 fully annotated LiDAR scans with a complete 360° field of view. It includes annotations for 19 object categories, and sequences 00-07 and 11-21 are used for online testing. RGB images corresponding to the LiDAR scans are provided for the entire dataset.

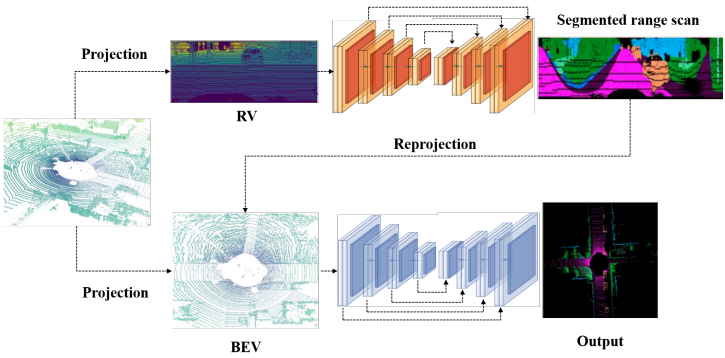
nuScenes: This dataset was developed by the Motional team in 2020 for autonomous driving. It is a large-scale public dataset that includes sensors such as 6 cameras, 1 LiDAR, 5 mm-wave radars, GPS, and IMU. The nuScenes dataset contains 1,000 different scenes, including annotations for 17 object categories. The data is split into 700/150/150 scenes for training, validation, and testing, respectively.

3 Point Cloud Semantic Segmentation Methods Based On Information Fusion

In this section, we will explore three-dimensional point cloud semantic segmentation techniques in autonomous driving, with a focus on mitigating sensor information loss. We divide these techniques into three main approaches: Multi-View Fusion, Multi-Representation Fusion, and Multi-Sensor Fusion. As shown in Fig 3, we provide an overview of representative point cloud semantic segmentation methods, with different colors representing different approaches. These methods use information fusion to utilize various representations of



(a) Feature Fusion



(b) Result Fusion

Fig. 4 Classification of multi-view fusion methods.

point clouds and RGB images, aiming to improve the accuracy and robustness of semantic segmentation. Specifically, multi-view fusion methods project point cloud data onto various view planes and fuse information from multiple views to obtain comprehensive semantic information. Multi-representation fusion methods utilize diverse data formats to represent point clouds, such as spherical projection maps or voxel data and then merge features from these different representations. Lastly, multi-sensor fusion methods combine data from various sensors, including LiDAR and camera, to overcome the limitations of individual sensors and enhance the robustness of semantic segmentation.

3.1 Multi-View Fusion

The projection-based methods transform three-dimensional point cloud data into two-dimensional image data such as Range View (RV) and Bird's Eye View (BEV). However, this transformation inevitably leads to information loss, i.e., the depth information is lost in the range view, and the height information is lost in the bird's eye view, as shown in Table 1. To reduce this information loss, researchers have attempted to fuse rich complementary information between

Table 1 Advantages And Disadvantages Of Projection Views.

Projection Methods	Advantages	Disadvantages
Spherical projection	Reduce point cloud dimensions	Loss of depth information
Bird's eye view projection	Reduce computational costs	Loss of height information

Table 2 Classification of multi-view fusion methods.

Fusion hierarchy	Methods	Network	Advantages	Disadvantages
Feature fusion	MVlidarNet[36],GFNet[37] TornadoNet[40]	2D	Multi-view Information Fusion Reduce Information Loss	Discretization Error
Decision fusion	MPF[39] AMVNet[38]	2D	High Robustness	Discretization Error

different projection views. This fusion can be categorized into feature fusion and decision fusion, as shown in Table 2. The typical model architecture of multi-view fusion methods is shown in Fig 4.

To exploit the complementary information between the range feature map and the BEV feature map, Chen et al. [36] introduced a two-stage semantic segmentation network called MVlidarNet. This method first obtains initial predictions from the range map, then reprojects them as initial features onto the BEV map, and utilizes a feature pyramid network for feature learning. Subsequently, Gerdzhev et al. [37] proposed TornadoNet. This approach proposes a Pillar Projection Learning (PPL) module designed to extract features from the bird's eye view and transfer them to the depth map. This process is performed using an encoder-decoder architecture equipped with rhombic context blocks, facilitating feature learning through multi-view projection. To incorporate geometric information, Qiu et al. [38] introduced the Geometry Flow Network (GFNet). This method explores geometric correspondences between different views in a pre-aligned fusion manner, achieving bidirectional alignment and propagation of complementary information between different views.

To fuse predictions from different view networks, Alnaggar et al. [39] introduced a post-fusion network called MPF. This method employs two separate branches to handle RV and BEV maps and combines the predictions from both branches using softmax probabilities to generate the final prediction. Additionally, Liong et al. [40] proposed an Advanced Multi-View Network (AMVNet) for further refinement. After obtaining predictions from the Range and BEV images, this method utilizes point sampling to feed each sampled point into a point-head to enhance the predictions of uncertain points in both branches.

While multi-view fusion methods have made progress in point cloud semantic segmentation, range images face challenges during the projection process such as many-to-one mapping, occluded pixels, and shape distortions, which affect the alignment of information between range and BEV images. Future research efforts should delve into addressing these issues to enhance the accuracy and robustness of point cloud semantic segmentation algorithms based on multi-view fusion.

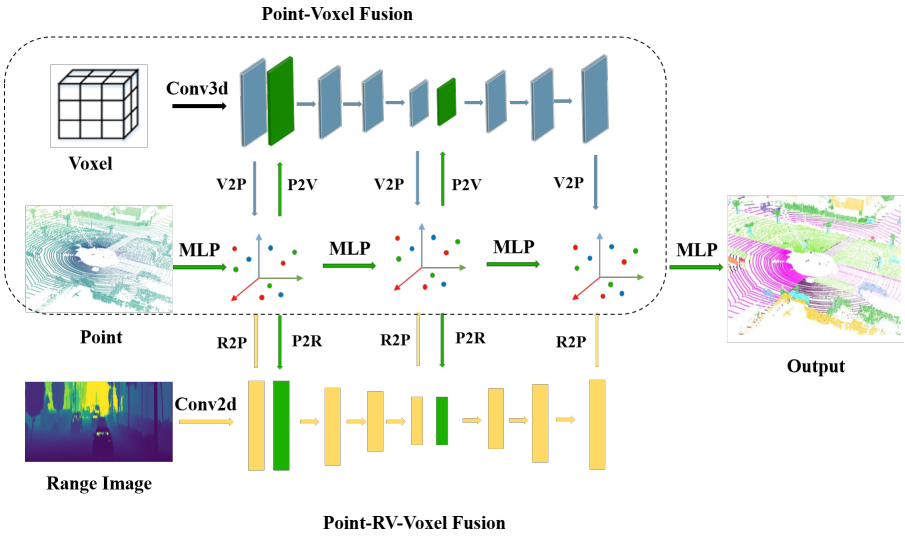


Fig. 5 The framework of multi-representation fusion: point-voxel fusion and point-rv-voxel fusion.

3.2 Multi-Representation Fusion

Point cloud semantic segmentation has three primary representations: points, projection images, and voxels. To exploit the information across these different representations, researchers have begun to explore multi-representation information fusion, leading to the development of a number of network architectures as shown in Fig 5. These architectures have been categorized as shown in Table 3.

To reduce the computational inefficiency associated with point-to-voxel based approaches, Liu et al. [41] introduced an efficient point-to-voxel network called PVCNN. This method exploits the coarse-grained local features provided by voxels and the fine-grained geometric features from the original point cloud. Efficient information fusion is achieved through simple point-wise convolutions. Building upon this, Zhao et al. [42] proposed a lightweight network, SPVNAS, suitable for segmenting small objects. This method introduces a lightweight 3D module, SPVConv, equipped with a high-resolution point-based branch for sparse voxel convolution, which ensures high-resolution performance in outdoor scenes. Additionally, the concept of 3D Neural Architecture Search (3D-NAS) was introduced to automatically search and optimize the optimal network architecture within a diverse design space. Similarly, to achieve more accurate point-wise convolutions, Zhang et al. [43] presented FusionNet, a network structure based on voxel-based mini-point networks. This method utilizes voxel-based mini-point network structures for point cloud representation and learning, facilitating neighborhood voxel-level feature aggregation and fine-grained point-wise feature learning. To explore information iteration between point and voxel features, Ye et al. [44] introduced DRINet, a

Table 3 Classification Of Multi-representation Fusion Methods .

Representation	Methods	Network	Advantages	Disadvantages
Point-voxel	PVCNN[41], SPVNAS[42], FusionNet[43] DRINet(++)[44, 45], PCSCNet[47]	3D	Fusion of geometric information and spatial locality	Resolution depends on voxel grid size
Point-RV	3DminiNet[50] CPGNet[48]	2D+3D	Low Computational complexity Small storage overhead	Discretization error Shape distortion
Voxel-BEV	LiDARmultiNet[46]	2D+3D	Fusion of Local and Global information	High computational complexity Discretization error
point-RV-voxel	RPVNet[49]	2D+3D	Robustness in complex scenes High Accuracy	High computational complexity Large storage overhead

semantic segmentation network with dual representation iterative learning. This approach iteratively employs sparse point-to-voxel feature extraction and sparse voxel-to-point feature extraction, allowing both representations to propagate features to each other. Furthermore, Ye et al. [45] extended DRINet to DRINet++, aiming to strike a balance between performance, efficiency, and memory consumption. This method treats voxels as points and enhances point cloud sparsity and geometric characteristics through multi-scale sparse projection and multi-scale attention fusion. To mitigate the computational burden associated with high voxel resolutions, Jaehyun et al. [46] proposed PCSCNet. This approach combines point convolution for feature extraction and 3D sparse convolution for feature propagation, effectively improving performance at low voxel resolutions. Moreover, a position-aware (PA) loss is introduced to deal with discretization errors at large voxel sizes.

To explore the fusion of point-based and projection-based representations, Alonso et al. [47] introduced a semantic segmentation network called 3DMiniNet that combines 3D semantic information with 2D learning layers. This approach extracts local and global information from the original 3D point cloud to learn 2D projection representations, replacing the conventional spherical projection with point-based learning representations.

To strike a better balance between speed and accuracy, Li et al. [48] introduced CPGNet. This method incorporates a novel point-grid fusion module that organically integrates point views, bird's eye views, and depth maps within a cascaded framework to mitigate information loss in projection views. Additionally, to speed up the inference process, the approach introduces transformation consistency loss and narrows the gap between single model inference and test time augmentation (TTA) by reducing differences between the original point cloud and augmented points.

To enhance the global feature learning capability of 3D sparse voxel convolution, Ye et al. [49] proposed LidarMultiNet. This approach introduces a Global Context Pooling (GCP) mechanism, which aggregates rich contextual information into sparse voxels through feature mapping between sparse 3D convolution and dense bird's eye view representations. This enhancement significantly improves the network's capacity for global feature learning. To explore the fusion of three representation methods, Xu et al. [50] proposed RPVNet, a semantic segmentation network that combines range, point, and voxel representations. This method adopts a gated fusion module (GFM) to adaptively merge features between the three views, thereby enhancing the mutual information exchange between these representation modes.

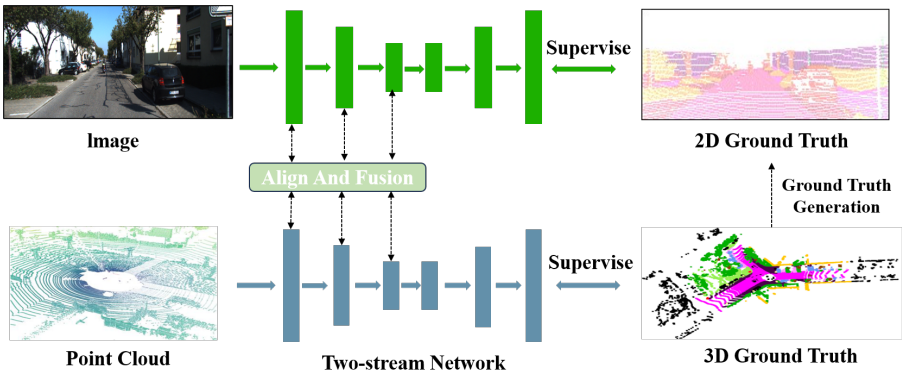


Fig. 6 The framework of multi-sensor fusion.

In recent years, semantic segmentation for point clouds has witnessed rapid progress in methods based on multi-representation fusion. However, there remains a significant need for extensive experimentation to determine how to efficiently combine three distinct representations: the geometric characteristics of point clouds, the real-time efficiency of projection views, and the sparsity of voxels.

3.3 Multi-Sensor Fusion

In complex autonomous driving scenarios, a single sensor often cannot provide sufficient information to classify the semantic category of each point accurately. Therefore, multi-sensor fusion can gather richer information from multiple sensors, thereby enhancing the accuracy and robustness of point cloud semantic segmentation, as shown in Fig 6.

To effectively integrate 2D image information into 3D point cloud semantic segmentation, Madawi et al. [51] proposed an early fusion segmentation algorithm called RGBAL. This method initially performs early fusion on the raw data, then converts RGB images into polar coordinate network mappings to extract modality features, and performs feature-level midterm fusion. To incorporate semantic information from RGB images, Vora et al. [52] introduced a sequence fusion algorithm called PointPainting. This approach first projects 3D points onto 2D images to obtain image segmentation results, then back-projects them into the LiDAR space using either bird's eye projection or spherical projection. Finally, it concatenates the projected segmentation scores with the original point cloud features to enhance the performance of the LiDAR network. To explore the cooperative fusion of RGB image and point cloud information, Zhuang et al. [53] introduced PMF. This method utilizes perceptual information from both modes: the appearance information from RGB images and the spatial depth information from point clouds are coherently fused. To introduce image information without increasing inference time, Yan et al. [54] proposed a 2Dpass network for image-assisted 3D semantic segmentation. This approach, without strict paired data constraints, distills rich

Table 4 Advantages And Disadvantages Of Information Fusion Method Classification.

Information Fusion Methods	Advantages	Disadvantages
Multi-view fusion	Alleviate occlusion and viewpoint restrictions in a single view enhance the robustness of point cloud data	High computational complexity noise from viewpoint alignment and fusion
Multi-Representation fusion	Rich point cloud features spatial consistency in point cloud data	Higher storage and computational complexity limited by voxel grid resolution
Multi-sensor fusion	Introduction of RGB texture information enhancing robustness in complex environments	High system complexity difficulty in modal fusion

semantic and structural information obtained from multimodal data into a pure 3D network through image-assisted modality fusion and multiscale fusion. It uses image assistance only during training, which significantly reduces the inference time cost.

To reduce the dependence on annotated 3D point cloud data, Genova et al. [55] introduced a method for training 3D semantic segmentation from 2D image supervision called 2D3DNet. This approach operates a pre-trained 2D model on each image, generating per-pixel semantic labels. It then back-projects 3D points onto the images to generate weighted voting for feature extraction and pseudo-labeling. This significantly alleviates the dependency on point cloud semantic labels. To further reduce reliance on point cloud annotation data and exploit complementary information from images, Sun et al. [56] proposed a weakly supervised semantic segmentation method based on cross-modal correlation learning, known as IUPC-AL. This method enhances the consistency between image superpixels (pixels within visually similar regions often contain similar semantic information) and 3D points, mining complementary supervisory information from images. Simultaneously, the introduction of a pseudo-label self-correction mechanism effectively filters out noisy labels.

To address the weak spatio-temporal synchronization between LiDAR point clouds and RGB images, Zhao et al. [57] proposed a coarse-to-fine semantic segmentation framework called Lif-seg. In this approach, point cloud and low-level image context information are fused in the coarse stage, followed by offset correction between coarse features. Finally, the aligned features are passed into a subnetwork to refine the segmentation results. To mitigate the modality heterogeneity between RGB images and point clouds, along with the limited cross-view field of view (FOV), Li et al. [58] introduced a multi-modal 3D semantic segmentation network called MSeg3D. This method promotes maximum correlation and complementarity between heterogeneous modalities through joint optimization of intra-modal feature extraction and inter-modal feature fusion. Additionally, for points outside the FOV cross-view, predicted pseudo-camera features are used to supplement missing camera features.

Despite rapid advances in semantic segmentation methods based on multisensor fusion in recent years, alignment errors between LiDAR point clouds and RGB images, information loss during modality fusion, and more sophisticated fusion operations still require extensive exploration by researchers. Based on the review of the information fusion methods described above, a detailed comparison of fusion-based methods is provided in Table 4.

Table 5 Comparison Of MIoU, Time Complexity, And Space Complexity In Information Fusion-Based Point Cloud Semantic Segmentation Methods

Methods	Network	Time	mIoU/ % (cm)		Speed/ms		Parapeter/M
			SemanticKITTI	nuScenes	SemanticKITTI	nuScenes	
Multi-view fusion	MvlidarNet[36]	2020	52.5	——	10	——	4.5
	AmvNet[38]	2020	65.3	76.1	85	——	——
	TornadoNet[40]	2021	63.1	——	250	——	——
	MPF[39]	2021	55.5	——	50	——	3.18
	GFNet[37]	2022	65.4	76	100	——	87.6
Multi-Representation fusion	PVCNN[41]	2019	39.0	——	146	——	2.5
	SPVNAS[42]	2020	66.4	77	256	——	12.5
	3D-MiniNet[49]	2020	55.8	——	35	——	3.97
	FusionNet[43]	2020	61.3	——	256	——	——
	DRINet[44]	2021	67.5	——	62	——	——
	DRINet++[45]	2021	70.7	80	59	——	——
	RPVNet[46]	2021	70.3	77.6	168	——	24.8
	CPGNet[48]	2022	68.3	76.9	43	43	——
	PCSCNet[50]	2023	62.7	72.0	176	——	——
	LidarMultiNet[47]	2023	——	81.4	——	——	——
Multi-sensor fusion	RGBAL[51]	2019	56.2	——	——	——	——
	PointPainting[52]	2020	54.5	——	——	——	——
	PMF[53]	2021	63.9	76.9	125	——	36.4
	2D3Dnet[57]	2021	——	80.0	——	——	——
	Lif-seg[54]	2021	——	78.2	——	——	——
	IUPC-AL[56]	2022	64.9	77.8	——	——	——
	2DPASS[55]	2022	72.9	80.8	62	44	1.9
	Mseg3D[58]	2023	66.7	81.1	——	——	——

4 Comprehensive Comparative Analysis

In this section, we conducted a comprehensive comparative analysis of the previously proposed deep learning-based point cloud semantic segmentation algorithms. To facilitate the comparison of experimental results, Table 5 lists the comparative experimental results of mIoU, time complexity, and space complexity on representative 3D point cloud datasets, following the classification method of point cloud semantic segmentation algorithms mentioned earlier in the text. From Table 5, it can be seen that in autonomous driving, the SemanticKITTI dataset is the most widely used. Regarding the development of semantic segmentation methods based on information fusion, research on methods based on multi-representation fusion is more extensive compared to those based on multi-view fusion and multi-sensor fusion. In terms of the mIoU metric, methods based on multi-sensor fusion and multi-representation fusion outperform others in the category of information fusion methods. In terms of time and space complexity, methods based on multi-sensor fusion and multi-representation fusion have lower memory consumption and time consumption. When considering mIoU and time and space complexity together, methods based on multi-sensor fusion show superior performance. In recent years, methods based on multi-view fusion have shown continuous improvement in the mIoU performance metric, indicating significant potential for further improvement. The development of methods based on multi-representation fusion has also been rapid in recent years, with an increasing trend in mIoU performance and a reduction in time and space complexity. Methods based on multi-sensor fusion have seen rapid development in various tasks in the field of autonomous driving, and currently demonstrate the best performance in terms of mIoU, with time and space complexity comparable to methods based on multi-representation fusion.

5 Existing Challenges and Future Prospects

5.1 Multi-Sensor Fusion Optimization

Multi-sensor fusion methods also face challenges related to data misalignment [59, 60] and resolution differences [61, 62] between RGB images and point clouds. These issues pose significant challenges for point cloud semantic segmentation methods that rely on multi-sensor fusion. Due to the weak spatio-temporal synchronization between LiDAR and cameras, some points in the three-dimensional point cloud may project onto areas outside of object instances in the images. Additionally, RGB images provide dense texture information, whereas the original point cloud data only corresponds to a subset of points in the images.

Multi-sensor fusion techniques are crucial for enhancing the performance of point cloud semantic segmentation, especially for distant or small objects [63]. However, the field of point cloud semantic segmentation based on multi-sensor fusion is still in its infancy. Migrating point cloud object detection methods to point cloud semantic segmentation to achieve weak spatio-temporal alignment, information preservation, and efficient fusion is a vital research direction for the future. With advancements in sensor technology and multi-modal algorithms, multi-sensor fusion for point cloud semantic segmentation holds great promise for making significant breakthroughs in fields such as autonomous driving, ultimately improving the reliability of environmental perception and decision making.

5.2 Fine-grained Annotation Of Point Cloud Data

Due to the three-dimensional nature of point cloud data, achieving fine-grained annotation requires classifying each point within the point cloud. This process is more complex and time-consuming than annotating images at the pixel level. The lack of fine-grained annotation can lead to inadequate model learning of models in specific scenes or categories, which affects the generalization and accuracy of point cloud semantic segmentation.

Weakly supervised learning methods extract valuable information from limited annotated data [64, 65], such as label inconsistency, spatial relationships, and contextual cues. They improve model training when labeled data is scarce, resulting in more robust models. Unsupervised learning methods, on the other hand, uncover latent patterns and associations in unlabeled data, providing a broader perspective to improve generalization and adaptability in diverse and complex scenarios [66]. Therefore, exploring weakly supervised/unsupervised learning approaches to reduce dependence on annotated data and improve model scalability, generalization, and domain adaptation is a critical factor in moving three-dimensional point cloud semantic segmentation from theory to practice.

5.3 Complex Weather

In challenging weather conditions [67, 68] such as fog, rain, and snow, the phenomena of light scattering and absorption significantly affect the signals received by sensors, thereby reducing the accuracy and visibility of depth information in point clouds. This leads to the presence of significant noise and distortion in point cloud data, resulting in increased segmentation error rates and blurred object boundaries. Additionally, the noise and deformation under complex weather conditions pose challenges for feature extraction, causing a decrease in the performance of traditional segmentation algorithms.

Several approaches [69, 70] can be considered to address these issues: 1) Sensor Technology Advancements: Improving sensor technology by using more advanced LiDAR and cameras can improve sensor performance in adverse weather conditions. This, in turn, improves the quality and reliability of point cloud data. 2) Foggy Image Enhancement: Foggy weather image enhancement techniques can be applied to image data to reduce blur and distortion caused by adverse weather conditions, providing clearer visual information. 3) Multi-Sensor Fusion: Employing a multi-sensor fusion strategy that combines information from various sensors can compensate for the limitations of different sensors under complex weather conditions and enhance the robustness of segmentation algorithms.

By implementing these strategies, researchers can mitigate the challenges posed by complex weather conditions, improving the overall performance of point cloud semantic segmentation algorithms.

5.4 3D segmentation universal large models

In recent years, large-scale universal AI models have made tremendous progress in various domains, especially in the area of language, exemplified by models such as GPT-4. The scaling up of these models has received widespread attention due to their enhanced capabilities. In the realm of the image segmentation domain, models such as Segment Anything [71] have demonstrated excellent zero-shot capabilities, allowing for interactive object segmentation in single images through methods such as mouse-clicking, drawing bounding boxes, text input, and more. However, when dealing with more complex three-dimensional spaces, particularly in high-dimensional data scenarios, there remains a lack of universal large models capable of addressing data-related challenges [72]. Looking ahead, as technology continues to evolve, we can foresee that universal large-scale models have the potential to play a more significant role in the three-dimensional domain. They are poised to provide smarter and more efficient solutions for segmenting and labeling three-dimensional data, thereby driving further advances in three-dimensional spatial applications.

6 Conclusion

In autonomous driving, deep learning-based point cloud semantic segmentation has emerged as a prominent research focus in recent years. This review provides a comprehensive overview of the current state of research in deep learning-based point cloud semantic segmentation, categorizing it into three main approaches: 1) methods based on multi-view fusion, 2) methods based on multi-modal fusion, and 3) methods based on multi-sensor fusion. Furthermore, this review summarizes commonly used datasets and evaluation metrics in the autonomous driving domain. Moreover, this review conducts a thorough comparative analysis of the overall performance of different methods, highlighting the existing challenges in point cloud semantic segmentation and outlining future research directions. As a rapidly advancing research field, point cloud semantic segmentation offers substantial research opportunities in areas such as multi-sensor fusion optimization, weak and self-supervised learning, enhancement of performance in complex weather conditions, and the application of three-dimensional AI universal large models.

Funding

Supported by National Natural Science Foundation of China U1913602 and 61991412

Acknowledgments

The authors thank the developers and maintainers of all open-source software, languages, and systems used in the article for their contributions. The authors also thank the instructor Jie Ma and the reviewers for their valuable comments, which greatly improved the content and readability of this article.

References

[1] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361 (2012). IEEE

[2] Gan, C., Zhao, H., Chen, P., Cox, D., Torralba, A.: Self-supervised moving vehicle tracking with stereo sound. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7053–7062 (2019)

[3] Liu, Z., Zhou, S., Suo, C., Yin, P., Chen, W., Wang, H., Li, H., Liu, Y.-H.: Lpd-net: 3d point cloud learning for large-scale place recognition and environment analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2831–2840 (2019)

[4] Rusu, R.B., Marton, Z.C., Blodow, N., Dolha, M., Beetz, M.: Towards 3d point cloud based object maps for household environments. Robotics and Autonomous Systems **56**(11), 927–941 (2008)

[5] Shan, T., Englot, B.: Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4758–4765 (2018). IEEE

[6] Panev, S., Vicente, F., De la Torre, F., Prinet, V.: Road curb detection and localization with monocular forward-view vehicle camera. IEEE Transactions on Intelligent Transportation Systems **20**(9), 3568–3584 (2018)

[7] Spielberg, N.A., Brown, M., Kapania, N.R., Kegelman, J.C., Gerdes, J.C.: Neural network vehicle models for high-performance automated driving. Science robotics **4**(28), 1975 (2019)

[8] Johnson, B., Havlak, F., Kress-Gazit, H., Campbell, M.: Experimental evaluation and formal analysis of high-level tasks with dynamic obstacle anticipation on a full-sized autonomous vehicle. Journal of Field Robotics **34**(5), 897–911 (2017)

[9] Tian, Y., Dong, H.-h., Jia, L.-m., Li, S.-y.: A vehicle re-identification algorithm based on multi-sensor correlation. Journal of Zhejiang University SCIENCE C **15**(5), 372–382 (2014)

[10] 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, pp. 652–660 (2017)

[11] Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)

[12] 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, pp. 11108–11117 (2020)

[13] 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, pp. 6411–6420 (2019)

[14] Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K.: Squeezesegv2: Improved

model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 4376–4382 (2019). IEEE

[15] 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: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16, pp. 1–19 (2020). Springer

[16] Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: Rangenet++: Fast and accurate lidar semantic segmentation. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4213–4220 (2019). IEEE

[17] Cortinhal, T., Tzelepis, G., Erdal Aksoy, E.: Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In: Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15, pp. 207–222 (2020). Springer

[18] Zhou, H., Zhu, X., Song, X., Ma, Y., Wang, Z., Li, H., Lin, D.: Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. arXiv preprint arXiv:2008.01550 (2020)

[19] Cheng, R., Razani, R., Taghavi, E., Li, E., Liu, B.: 2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12547–12556 (2021)

[20] Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9224–9232 (2018)

[21] Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3075–3084 (2019)

[22] Liu, W., Sun, J., Li, W., Hu, T., Wang, P.: Deep learning on point clouds and its application: A survey. Sensors **19**(19), 4188 (2019)

[23] Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep learning for 3d point clouds: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(12), 4338–4364 (2020)

[24] Cui, Y., Chen, R., Chu, W., Chen, L., Tian, D., Li, Y., Cao, D.: Deep learning for image and point cloud fusion in autonomous driving: A review. IEEE Transactions on Intelligent Transportation Systems **23**(2), 722–739 (2021)

[25] Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K.: Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems **22**(3), 1341–1360 (2020)

[26] Huang, K., Shi, B., Li, X., Li, X., Huang, S., Li, Y.: Multi-modal sensor fusion for auto driving perception: A survey. arXiv preprint arXiv:2202.02703 (2022)

[27] Li, Y., Ma, L., Zhong, Z., Liu, F., Chapman, M.A., Cao, D., Li, J.: Deep learning for lidar point clouds in autonomous driving: A review. IEEE Transactions on Neural Networks and Learning Systems **32**(8), 3412–3432 (2020)

[28] Wang, Y., Mao, Q., Zhu, H., Deng, J., Zhang, Y., Ji, J., Li, H., Zhang, Y.: Multi-modal 3d object detection in autonomous driving: a survey. International Journal of Computer Vision, 1–31 (2023)

[29] Li, Y., Ma, L., Zhong, Z., Liu, F., Chapman, M.A., Cao, D., Li, J.: Deep learning for lidar point clouds in autonomous driving: A review. IEEE Transactions on Neural Networks and Learning Systems **32**(8), 3412–3432 (2020)

[30] Kim, H., Kim, J., Nam, H., Park, J., Lee, S.: Spatiotemporal texture reconstruction for dynamic objects using a single rgb-d camera. In: Computer Graphics Forum, vol. 40, pp. 523–535 (2021). Wiley Online Library

[31] Wang, Z., Zhan, W., Tomizuka, M.: Fusing bird's eye view lidar point cloud and front view camera image for 3d object detection. In: 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1–6 (2018). IEEE

[32] Graham, B., Van der Maaten, L.: Submanifold sparse convolutional networks. arXiv preprint arXiv:1706.01307 (2017)

[33] 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 228–240 (2023)

[34] Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss,

C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9297–9307 (2019)

[35] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11621–11631 (2020)

[36] Chen, K., Oldja, R., Smolyanskiy, N., Birchfield, S., Popov, A., Wehr, D., Eden, I., Pehserl, J.: Mvlidarnet: Real-time multi-class scene understanding for autonomous driving using multiple views. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2288–2294 (2020). IEEE

[37] Gerdzhev, M., Razani, R., Taghavi, E., Bingbing, L.: Tornado-net: multiview total variation semantic segmentation with diamond inception module. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 9543–9549 (2021). IEEE

[38] Qiu, H., Yu, B., Tao, D.: Gfnet: Geometric flow network for 3d point cloud semantic segmentation. arXiv preprint arXiv:2207.02605 (2022)

[39] Alnaggar, Y.A., Afifi, M., Amer, K., ElHelw, M.: Multi projection fusion for real-time semantic segmentation of 3d lidar point clouds. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1800–1809 (2021)

[40] 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 preprint arXiv:2012.04934 (2020)

[41] Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. Advances in Neural Information Processing Systems **32** (2019)

[42] 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: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII, pp. 685–702 (2020). Springer

[43] Zhang, F., Fang, J., Wah, B., Torr, P.: Deep fusionnet for point cloud semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16, pp. 644–663 (2020). Springer

[44] Ye, M., Xu, S., Cao, T., Chen, Q.: Drinet: A dual-representation iterative learning network for point cloud segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7447–7456 (2021)

[45] Ye, M., Wan, R., Xu, S., Cao, T., Chen, Q.: Drinet++: Efficient voxel-as-point point cloud segmentation. arXiv preprint arXiv:2111.08318 (2021)

[46] Park, J., Kim, C., Kim, S., Jo, K.: Pcscnet: Fast 3d semantic segmentation of lidar point cloud for autonomous car using point convolution and sparse convolution network. Expert Systems with Applications **212**, 118815 (2023)

[47] 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 Robotics and Automation Letters **5**(4), 5432–5439 (2020)

[48] Li, X., Zhang, G., Pan, H., Wang, Z.: Cpgnet: Cascade point-grid fusion network for real-time lidar semantic segmentation. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 11117–11123 (2022). IEEE

[49] Ye, D., Zhou, Z., Chen, W., Xie, Y., Wang, Y., Wang, P., Foroosh, H.: Lidarmultinet: Towards a unified multi-task network for lidar perception. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 3231–3240 (2023)

[50] 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16024–16033 (2021)

[51] El Madawi, K., Rashed, H., El Sallab, A., Nasr, O., Kamel, H., Yogamani, S.: Rgb and lidar fusion based 3d semantic segmentation for autonomous driving. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 7–12 (2019). IEEE

[52] Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4604–4612 (2020)

[53] Zhuang, Z., Li, R., Jia, K., Wang, Q., Li, Y., Tan, M.: Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16280–16290 (2021)

[54] 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: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII, pp. 677–695 (2022). Springer

[55] Genova, K., Yin, X., Kundu, A., Pantofaru, C., Cole, F., Sud, A., Brewington, B., Shucker, B., Funkhouser, T.: Learning 3d semantic segmentation with only 2d image supervision. In: 2021 International Conference on 3D Vision (3DV), pp. 361–372 (2021). IEEE

[56] Sun, T., Zhang, Z., Tan, X., Qu, Y., Xie, Y., Ma, L.: Image understands point cloud: Weakly supervised 3d semantic segmentation via association learning. arXiv preprint arXiv:2209.07774 (2022)

[57] Zhao, L., Zhou, H., Zhu, X., Song, X., Li, H., Tao, W.: Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation. arXiv preprint arXiv:2108.07511 (2021)

[58] Li, J., Dai, H., Han, H., Ding, Y.: Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21694–21704 (2023)

[59] Chen, Y., Liu, F., Pei, K.: Cross-modal matching cnn for autonomous driving sensor data monitoring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3110–3119 (2021)

[60] Shin, K., Kwon, Y.P., Tomizuka, M.: Roarnet: A robust 3d object detection based on region approximation refinement. In: 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 2510–2515 (2019). IEEE

[61] Langer, F., Milioto, A., Haag, A., Behley, J., Stachniss, C.: Domain transfer for semantic segmentation of lidar data using deep neural networks. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 8263–8270 (2020). IEEE

[62] Yi, L., Gong, B., Funkhouser, T.: Complete & label: A domain adaptation approach to semantic segmentation of lidar point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15363–15373 (2021)

[63] Lai, X., Chen, Y., Lu, F., Liu, J., Jia, J.: Spherical transformer for lidar-based 3d recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17545–17555 (2023)

[64] 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, pp. 1726–1736 (2021)

[65] Zhang, Y., Li, Z., Xie, Y., Qu, Y., Li, C., Mei, T.: Weakly supervised semantic segmentation for large-scale point cloud. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 3421–3429 (2021)

[66] Sautier, C., Puy, G., Gidaris, S., Boulch, A., Bursuc, A., Marlet, R.: Image-to-lidar self-supervised distillation for autonomous driving data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9891–9901 (2022)

[67] Rasshofer, R.H., Spies, M., Spies, H.: Influences of weather phenomena on automotive laser radar systems. Advances in radio science **9**, 49–60 (2011)

[68] Hahner, M., Sakaridis, C., Dai, D., Van Gool, L.: Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15283–15292 (2021)

[69] Yu, M.-Y., Vasudevan, R., Johnson-Roberson, M.: Lisnownet: Real-time snow removal for lidar point clouds. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6820–6826 (2022). IEEE

[70] Seppänen, A., Ojala, R., Tammi, K.: 4denoisenet: Adverse weather denoising from adjacent point clouds. IEEE Robotics and Automation Letters **8**(1), 456–463 (2022)

[71] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

[72] Shen, Q., Yang, X., Wang, X.: Anything-3d: Towards single-view anything reconstruction in the wild. arXiv preprint arXiv:2304.10261 (2023)