UniSeg: A Unified Multi-Modal LiDAR Segmentation Network and the OpenPCSeg Codebase

Youquan Liu[1,2,*] Runnan Chen[1,3,] Xin Li[1,4] Lingdong Kong[1,5] Yuchen Yang[1,6]
Zhaoyang Xia[1,6] Yeqi Bai[1,†] Xinge Zhu[7] Yuexin Ma[8] Yikang Li[1,†] Yu Qiao[1] Yuenan Hou[1,†]
[1]Shanghai AI Laboratory [2]Hochschule Bremerhaven [3]The University of Hong Kong [4]East China Normal University
[5]National University of Singapore [6]Fudan University [7]The Chinese University of Hong Kong [8]Shanghai Tech University

Abstract

*Point-, voxel-, and range-views are three representative forms of point clouds. All of them have accurate 3D measurements but lack color and texture information. RGB images are a natural complement to these point cloud views and fully utilizing the comprehensive information of them benefits more robust perceptions. In this paper, we present a unified multi-modal LiDAR segmentation network, termed UniSeg, which leverages the information of RGB images and three views of the point cloud, and accomplishes semantic segmentation and panoptic segmentation simultaneously. Specifically, we first design the **L**earnable cross-**M**odal **A**ssociation (LMA) module to automatically fuse voxel-view and range-view features with image features, which fully utilize the rich semantic information of images and are robust to calibration errors. Then, the enhanced voxel-view and range-view features are transformed to the point space, where three views of point cloud features are further fused adaptively by the **L**earnable cross-**V**iew **A**ssociation module (LVA). Notably, UniSeg achieves promising results in three public benchmarks, i.e., SemanticKITTI, nuScenes, and Waymo Open Dataset (WOD); it ranks **1**st on **two** challenges of two benchmarks, including the LiDAR semantic segmentation challenge of nuScenes and panoptic segmentation challenges of SemanticKITTI. Besides, we construct the OpenPCSeg codebase, which is the **largest** and **most comprehensive** outdoor LiDAR segmentation codebase. It contains most of the popular outdoor LiDAR segmentation algorithms and provides **reproducible** implementations. The OpenPCSeg codebase will be made publicly available at https://github.com/PJLab-ADG/PCSeg.*

1. Introduction

LiDAR-based semantic segmentation, whose objective is to assign a semantic label to each input point, acts as an essential component in autonomous driving, digital cities, and service robots [16, 18, 21, 37]. With the advent of deep learning, an enormous amount of methods [38, 62, 32, 31, 20, 48, 9, 52, 6, 5, 23, 22] have been proposed and quickly dominate various benchmarks, such as SemanticKITTI [1] and nuScenes [3, 15].

Point cloud and RGB images are two frequently used modalities. As depicted in Fig. 1 (a), different modalities have their own merits and drawbacks. Point cloud provides reliable and accurate depth information, and can be processed in different views, e.g., point-view, voxel-view, and range-view. Specifically, point-view representation maintains the complete point information but is inefficient in capturing the neighboring point features due to the unstructured point locations. Voxel-view methods rasterize the point cloud into voxel cells that retain regular structure but suffer from severe voxelization loss especially when the voxel size is large. Range-view representations are dense and compact, which can be efficiently processed by highly optimized 2D convolution. However, the spherical projection inevitably destroys the original 3D geometric information. As for the RGB image, it embraces rich color and texture information, but can not provide precise spatial information.

Apparently, the input data from multi-modality and multiple views of the point cloud are supplementary to each other. Therefore, fully utilizing the comprehensive information benefits a more robust perception. However, such a cross-modal and cross-view fusion paradigm is not fully explored in LiDAR segmentation [14, 28, 48, 52]. Current multi-modal fusion methods are concentrated on the fusion of RGB and range images [14, 28, 24]. Other representations such as voxel- and point-views of the LiDAR point cloud, which maintain original data structure and provide fine-grained spatial information, are ignored in prior methods. Besides, they typically fuse the image and point cloud in a hard association manner through calibration matrices, thus being vulnerable to calibration errors [25].

In this paper, to address the aforementioned problems, we make the first attempt to dynamically fuse four differ-

*Work performed during an internship at Shanghai AI Laboratory.
†Corresponding authors.

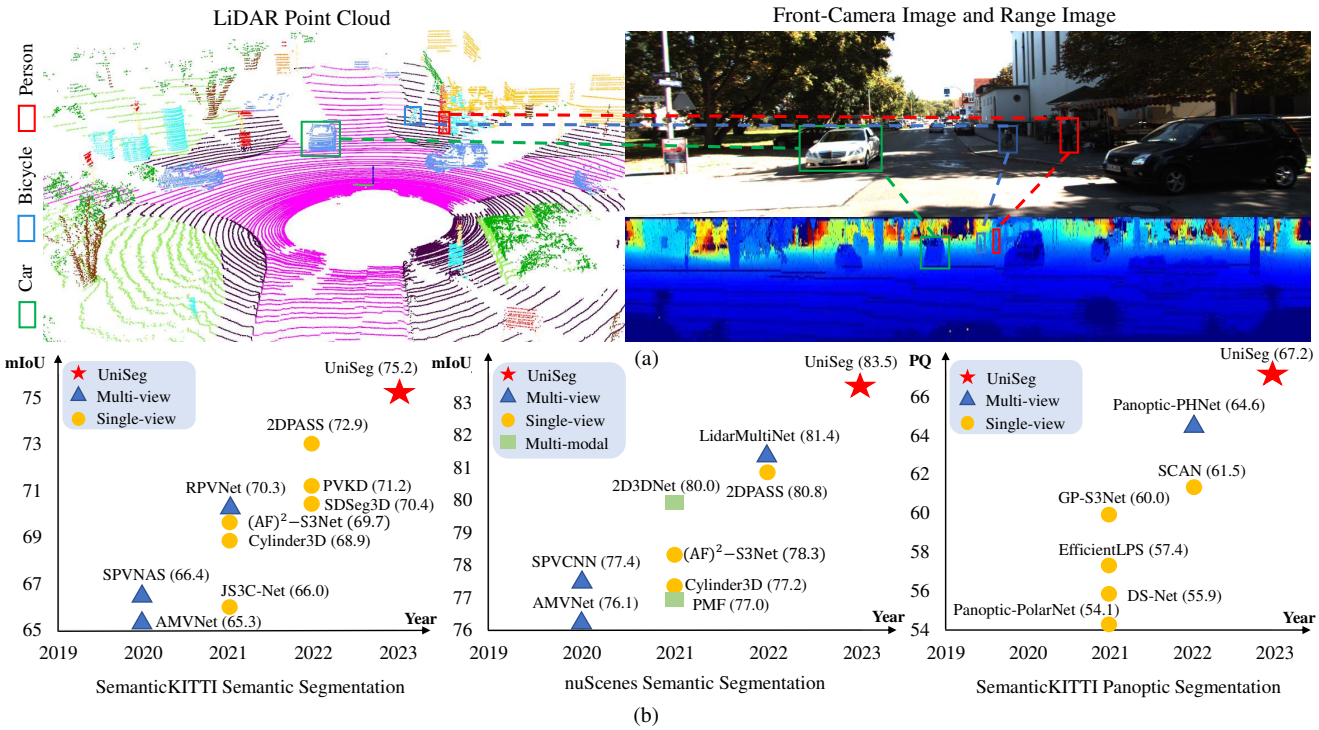

Figure 1: (a) Merits of different modalities. RGB images provide rich color, texture, and semantic information while point cloud embraces precise 3D positions of various objects. The pedestrian highlighted by the red rectangle is hard to find in the image but is visible in the point cloud. The combination of multi-modality and multi-views benefit a more robust and comprehensive perception. (b) Comparison of UniSeg with various competitive LiDAR segmentation algorithms on three challenges of SemanticKITTI and nuScenes benchmarks. The red pentagram, blue triangles, yellow circles, and green squares denote UniSeg, multi-view methods, uni-modal methods, and multi-modal ones, respectively. The selected baselines include state-of-the-art algorithms such as 2DPASS [52], RPVNet [48], Panoptic-PHNet [30], and LidarMultiNet [53].

ent modalities of data (voxel-, range-, and point-views of the point cloud and RGB images) for more robust and accurate perception. More formally, we propose a Learnable cross-Modal Association (LMA) and a Learnable cross-View Association module (LVA) to effectively fuse the different modalities inputs. Specifically, we first fuse the image features with range- and voxel-view point features through the LMA in a soft association schema with the deformable cross-attention [59] operation and alleviate calibration errors. Next, the image-enhanced range- and voxel-view features are transferred into the point-view feature, and all three views of point cloud features are fused adaptively by the LVA module.

Equipped with LMA and LVA, we design a unified network, dubbed UniSeg, for various semantic scene understanding tasks, *i.e.*, semantic, and panoptic segmentation. Extensive experimental results verify the generalizability of UniSeg across different tasks. As shown in Fig. 1 (b), UniSeg ranks 1st in **two** open challenges. It achieves 75.2 mIoU (semantic segmentation) and 67.2 PQ (panoptic segmentation) in SemanticKITTI; and 83.5 mIoU (semantic segmentation) and 78.4 PQ (panoptic segmentation) in nuScenes. The appealing performance strongly demonstrates the efficacy of our multi-modal fusion framework.

Besides, considering that many popular outdoor LiDAR segmentation methods [9, 48, 20, 30] either do not provide official implementations or the performance is difficult to reproduce, we construct the OpenPCSeg codebase which aims to provide reproducible and uniform implementations. We have benchmarked **14** competitive LiDAR segmentation algorithms and the reproduced performance of these algorithms all surpasses the reported value.

The contributions of our work are summarized as follows.

- We propose a unified multi-modal fusion network for LiDAR segmentation, leveraging the information of RGB images and three views of the point cloud for more accurate and robust perception.
- Our approach ranks 1st on two challenges of SemanticKITTI and nuScenes, strongly demonstrating the efficacy of the proposed multi-modal network.
- The largest and most comprehensive outdoor LiDAR

segmentation codebase dubbed OpenPCSeg will be released to facilitate related research.

2. Related Work

2.1. LiDAR-Based Semantic Scene Understanding

Semantic segmentation [48, 9, 20, 62, 61, 43, 52, 27, 26, 7, 8, 4, 34, 36, 50, 24] and panoptic segmentation [19, 30] are two basic tasks for LiDAR-based semantic scene understanding. LiDAR semantic segmentation aims to assign a class label to each point in the input point cloud sequence. LiDAR panoptic segmentation performs semantic segmentation and instance segmentation on the stuff class and thing class, respectively. The majority of the LiDAR segmentation approaches take the point cloud as the sole input signal. For instance, Cylinder3D [62, 61, 20] divides the point cloud with cylindrical partition and feeds these cylinder features into the UNet-based segmentation backbone. SPVCNN [43] introduces the point branch to complement the original voxel branch and performs pointwise segmentation based on the fused point-voxel features. LidarMultiNet [53] unifies LiDAR semantic segmentation, panoptic segmentation, and 3D object detection in one network and achieves impressive perception performance. The preceding methods ignore the rich information contained in RGB images, thus yielding sub-optimal performance. On the contrary, our UniSeg takes all modalities and all views of the point cloud into account and can benefit from the merits of all input signals.

2.2. Multi-Modal Sensor Fusion

Since the uni-modal signal has its own shortcomings, multi-modal fusion is gaining increasing attention in recent years [63, 14, 28]. Zhuang *et al*. [63] projects the point cloud into the perspective view and fuses the multi-modal features through the residual-based fusion module. El Madawi *et al*. [14] performs early fusion and middle fusion of the range images and re-projected RGB images. Krispel *et al*. [28] incorporates the image features into the range-image-based backbone via the calibration matrices. The above-mentioned approaches merely perform one-to-one multi-modal fusion and cannot fully utilize the rich semantic information of RGB images. And these methods yield inferior performance when the calibration matrices are inaccurate. By contrast, our method can achieve more adaptive multi-modal feature fusion and relieve point-pixel misalignment using the proposed learnable cross-modal association module.

3. The OpenPCSeg Codebase

In the outdoor LiDAR segmentation field, many popular semantic segmentation algorithms [9, 48, 30, 20] either do not release their official implementations or the released codes are difficult to reproduce the reported performance. Currently, only a few open-sourced projects have provided the implementations of LiDAR segmentation models such as the well-known mmdetection3d project [12]. However, it only includes some classical indoor LiDAR segmentation algorithms. A brief comparison between mmdetection3d and our OpenPCSeg is presented in Table 1. To facilitate the research in the outdoor LiDAR segmentation area, we construct the largest and most comprehensive OpenPCSeg codebase that contains the reproducible implementations of these competitive LiDAR segmentation models. OpenPCSeg is built upon the noted OpenPCDet [44] project. Considering the fact that many implementation details are missing in the original paper, constructing such a codebase is non-trivial. It takes us around one year to build the codebase through an enormous number of experiments to determine the optimal selection of hyperparameters, data augmentations, optimizers, learning rate schedules, data pre-processing, and post-processing strategies, *etc*. Till now, we have successfully reproduced more than ten competitive outdoor LiDAR segmentation algorithms, such as SalsaNext [13], Cylinder3D [62], RPVNet [48] and SPVCNN [35]. The reproduced performance of these algorithms all surpasses the reported value in their original publications. The chosen datasets include SemanticKITTI [1] and nuScenes [3, 15]. The selected tasks contain LiDAR semantic segmentation and panoptic segmentation. We provide a full suite of training and inference protocols for these algorithms to ensure reproducibility. The complete performance comparison and additional information on the OpenPCSeg codebase are in the Appendix.

Table 1: Comparisons between existing codebase.

Codebase	Task	Task Difficulty	#Method
MMDetection3D	Indoor Seg	Relatively Easy	3
OpenPCSeg	Outdoor Seg	Difficult	**14**

4. Methodology

4.1. Framework Overview

UniSeg takes point cloud (voxel-, range- and point-views) and RGB images as input and performs semantic segmentation and panoptic segmentation in a single network. Specifically, the input point cloud is $\mathbf{X} \in \mathbb{R}^{N\times 3}$ and the input image is $I \in \mathbb{R}^{H\times W\times 3}$. N is the number of points, H and W are the height and width of the image, respectively. We obtain the range image representation by performing the spherical projection on the point cloud. The range image is fed to a range-view-based backbone to extract range image features $\mathbf{F}^R \in \mathbb{R}^{H_R\times W_R\times C_R}$. H_R, W_R, and C_R are the height, width, and number of channels of the range image feature, respectively. Then, we extract the point features $\mathbf{F}^P \in \mathbb{R}^{N\times C_p}$ via a series of Multi-Layer Perceptrons (MLPs), where C_p is the number of channels of the point

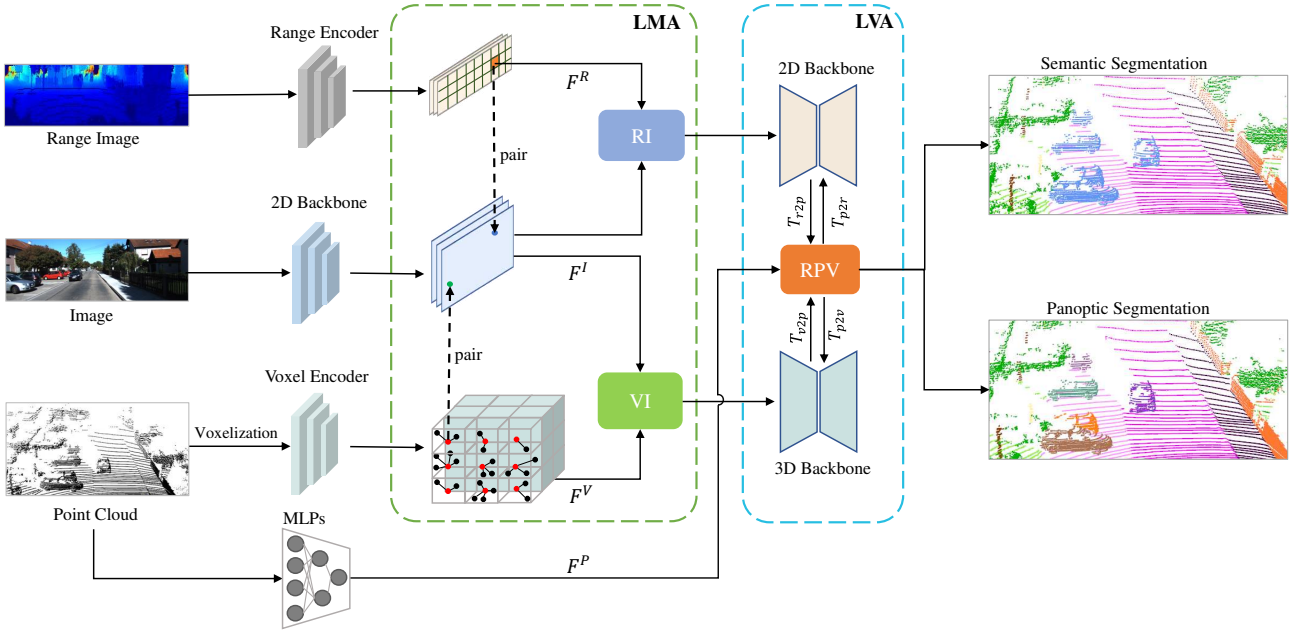

Figure 2: **Framework overview**. UniSeg takes four signals as input, *i.e.*, point cloud (voxel-, range- and point-views) and RGB images. Given the input point cloud, the range-, voxel-, and point-view features are produced by a 2D range encoder, a 3D voxel encoder, and MLPs, respectively. For the voxel features and range image features, we fuse them with the RGB image features (VI and RI) via the proposed learnable cross-modal association (LMA) module. Then, for the range image features and voxel features, we project them to the point space via the range-image-to-point transformation T_{r2p} and voxel-to-point transformation T_{v2p}. Features of these three views of the point cloud are fused (RPV) by the learnable cross-view association (LVA) module and we perform fusion at different layers to leverage both low-level and high-level information.

features. The voxel features $\mathbf{F}^V \in \mathbb{R}^{N_v \times C_p}$ are produced by the voxelization process that performs max pooling on the point features in one voxel. N_v is the number of non-empty voxels. The input image is fed to a ResNet-based architecture to extract the image features $\mathbf{F}^I \in \mathbb{R}^{H_I \times W_I \times C_I}$. H_I, W_I, and C_I are the height, width, and number of channels of the image feature, respectively.

Our method consists of two modules, *i.e.*, Learnable cross-Modal Association (LMA) and Learnable cross-View Association (LVA). The LMA module copes with the voxel-image fusion and range-image fusion, and the LVA module concentrates on range-point-voxel fusion. In what follows, we present LMA and LVA in detail.

4.2. Learnable Cross-Modal Association

Point-Image Calibration. We build the correspondence between the points and RGB image pixels via camera calibration matrices. Specifically, for each point coordinate (x_i, y_i, z_i), the corresponding pixel (u_i, v_i) is found by the following:

$$[u_i, v_i, 1]^{\mathrm{T}} = \frac{1}{z_i} \cdot S \cdot T \cdot [x_i, y_i, z_i, 1]^{\mathrm{T}}, \tag{1}$$

where $T \in \mathbb{R}^{4\times4}$ is the camera extrinsic matrix that consists of a rotation matrix and a translation matrix, and $S \in \mathbb{R}^{3\times4}$ is the camera intrinsic matrix. Here, we denote this pixel (u_i, v_i) as calibrated pixel p_i and the corresponding image feature as calibrated image feature F_i^I.

Voxel-Image Fusion. Previous multi-modal fusion approaches [14, 28] heavily rely on imperfect camera calibration matrices, which are vulnerable to calibration errors. Inspired by deformable detr [60], we adaptively fuse the voxel features with image features to alleviate the calibration errors. As shown in Fig. 3, the voxel coordinate is the voxel centre, and the calibrated image pixel is calculated by Equation 1. Next, we estimate the image pixel offsets from the calibrated image pixel, and then we fuse the selected image feature with the corresponding voxel feature as follows:

$$\begin{aligned} F_{i,l}^I &= F^I(\mathbf{p}_i + \Delta\mathbf{p}_{i,l}), \\ \hat{F}_i^V &= \sum_{m=1}^{M} W_m \left[\sum_{l=1}^{L} A_{i,l,m} \cdot (W_m^{'} F_{i,l}^I) \right], \end{aligned} \tag{2}$$

where F^I is the image feature, $F_{i,l}^I$ are the sampled image features and $\hat{F}_i^V$ is the image-enhanced voxel feature. W_m and $W_m^{'}$ are the learnable weights, m indexes the attention head, M is the number of self-attention heads and L is the total number of sampled image features. $\Delta\mathbf{p}_{i,l}$ and $A_{i,l,m}$ denote the sampling offset and attention weight of the l-th sampled image feature in the m-th attention head,

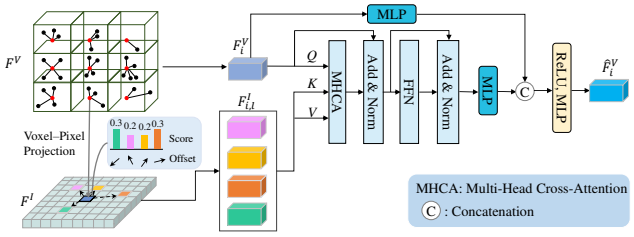

Figure 3: **Voxel-Image Fusion**. For each voxel feature F_i^V, We first calculate the calibrated image feature F_i^I based on the voxel center and calibration matrices. Then, we leverage learned offsets to sample L image features. The voxel feature is treated as *Query*, and the sampled image features are denoted as *Key* and *Value*. The voxel and sampled image features are fed to the multi-head cross-attention module to obtain image-enhanced voxel features. These features are concatenated with the original features to produce the final fused features.

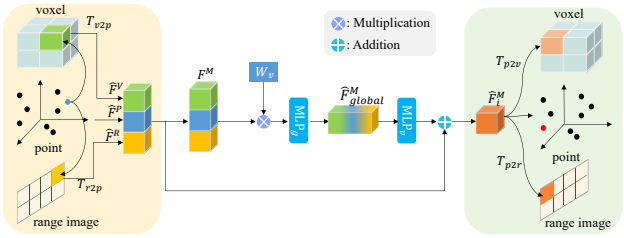

Figure 4: **Learnable cross-View Association (LVA)**. Voxel and range image features are first mapped to the point space where interpolations are employed to address the quantity mismatch problem through T_{v2p} and T_{r2p} transformations. Given voxel-, point- and range-view features, the LVA extracts its global representation and view-wised adapted features. Via a residual connection, the cross-view fused feature is obtained and projected back to the original voxel and range image space through T_{p2v} and T_{p2r} transformations.

respectively. Both are obtained by performing the linear projection on the voxel feature F_i^V. We concatenate the image-enhanced voxel feature $\hat{F}_i^V$ with the original voxel feature to obtain the final fused voxel feature $\hat{F}_i^V \in \mathbb{R}^{N\times 2C_f}$, where C_f is the number of channels of the voxel feature. Therefore, the voxel feature will automatically find the most relevant image features to fuse. Note that those voxel features that do not have the corresponding image features will be appended with zero vectors.

Range-Image Fusion. As to the range-image fusion, we follow the same process with voxel-image fusion (Equation 2), thus producing the final image-enhanced range-view features $\hat{F}^R \in \mathbb{R}^{H_R\times W_R\times C_f}$.

4.3. Learnable Cross-View Association

After the learnable cross-modal association module, we obtain the image-enhanced voxel- and range-view features. For the range-, point-, voxel-view features fusion, we first apply the range-to-point transformation T_{r2p} and voxel-to-point transformation T_{v2p} on the range-, voxel-view features to transfer them into the point-view respectively. And we propose a learnable cross-view association module to dynamically integrate these three modalities' features, as shown in Fig. 4.

Specifically, in the T_{r2p} and T_{v2p} transformations, since the number of voxel features and range image features is smaller than the number of points, directly appending all-zero vectors to voxel features and range image features yields sub-optimal performance. To address the aforementioned quantity mismatch problem, we resort to trilinear interpolation and bilinear interpolation to generate interpolated voxel features and pseudo range image features, respectively.

After these transformations, we obtain the point-wise voxel features $\hat{F}^V \in \mathbb{R}^{m\times C_f}$, point-wise range image features $\hat{F}^R \in \mathbb{R}^{m\times C_f}$ and point features $\hat{F}^P \in \mathbb{R}^{m\times C_f}$. And we concatenate them to produce the multi-view feature $F^M \in \mathbb{R}^{m\times 3C_f}$. Then F^M is weighted by the learnable parameters W_v and obtains the compact global point feature via the first two layers of LVA, *i.e.*, MLP_g, as follows:

$$\hat{F}^M_{global} = \mathrm{ReLU}(\mathrm{MLP}_g(W_v(\mathrm{concat}(\hat{F}^V, \hat{F}^R, \hat{F}^P)))), \tag{3}$$

where $\hat{F}^M_{global} \in \mathbb{R}^{m\times C_f}$.Through this cross-view aggregation, multi-view features fuse into a summative representation. After that, the view-wise adapted feature is generated from the globally enhanced features $\hat{F}^M_{global}$ and adds its original features of different views which are obtained by a residual connection as follows:

$$\hat{F}_i^M = \hat{F}_i + \mathrm{ReLU}(\mathrm{MLP}_v(\hat{F}^M_{global})), \tag{4}$$

where $\hat{F}_i \in \mathbb{R}^{m\times C_f}$ denotes the original feature in point space for view i. On the one hand, $\hat{F}_i^M$ provides global adapted features into $\hat{F}_i$ for a better representation of three different views. On the other hand, the residual style combines the benefits of multi-view knowledge with those of its advantages, which further encourages cross-view interaction. The final cross-view feature $\hat{F}_i^M$ is projected back to the original voxel and range image space by the T_{p2v} and T_{p2r} transformations respectively.

4.4. Task-Specific Heads

The fused features obtained by the LMA and LVA modules will be fed to the classifier to produce the semantic segmentation predictions. The semantic predictions are passed to the panoptic head to estimate instance centre positions and offsets of the thing class, producing the panoptic segmentation results. Detailed panoptic segmentation implementation is described in the supplementary material.

Table 2: Quantitative results of UniSeg and SoTA LiDAR semantic segmentation methods on the SemanticKITTI *test* set.

Method	mIoU	car	bicy	moto	truc	o.veh	ped	b.list	m.list	road	park	walk	o.gro	build	fenc	veg	trun	terr	pole	sign
AMVNet [33]	65.3	96.2	59.9	54.2	48.8	45.7	71.0	65.7	11.0	90.1	71.0	75.8	32.4	92.4	69.1	85.6	71.7	69.6	62.7	67.2
JS3C-Net [51]	66.0	95.8	59.3	52.9	54.3	46.0	69.5	65.4	39.9	88.9	61.9	72.1	31.9	92.5	70.8	84.5	69.8	67.9	60.7	68.7
SPVNAS [43]	66.4	97.3	51.5	50.8	59.8	58.8	65.7	65.2	43.7	90.2	67.6	75.2	16.9	91.3	65.9	86.1	73.4	71.0	64.2	66.9
Cylinder3D [62]	68.9	97.1	67.6	63.8	50.8	58.5	73.7	69.2	48.0	92.2	65.0	77.0	32.3	90.7	66.5	85.6	72.5	69.8	62.4	66.2
AF2S3Net [9]	69.7	94.5	65.4	**86.8**	39.2	41.1	**80.7**	80.4	**74.3**	91.3	68.8	72.5	**53.5**	87.9	63.2	70.2	68.5	53.7	61.5	71.0
RPVNet [48]	70.3	97.6	68.4	68.7	44.2	61.1	75.9	74.4	73.4	**93.4**	70.3	**80.7**	33.3	**93.5**	72.1	86.5	75.1	71.7	64.8	61.4
SDSeg3D [29]	70.4	97.4	58.7	54.2	54.9	65.2	70.2	74.4	52.2	90.9	69.4	76.7	41.9	93.2	71.1	86.1	74.3	71.1	65.4	70.6
GASN [54]	70.7	96.9	65.8	58.0	59.3	61.0	80.4	**82.7**	46.3	89.8	66.2	74.6	30.1	92.3	69.6	87.3	73.0	72.5	66.1	**71.6**
PVKD [20]	71.2	97.0	67.9	69.3	53.5	60.2	75.1	73.5	50.5	91.8	70.9	77.5	41.0	92.4	69.4	86.5	73.8	71.9	64.9	65.8
2DPASS [52]	72.9	97.0	63.6	63.4	61.1	61.5	77.9	81.3	74.1	89.7	67.4	74.7	40.0	**93.5**	**72.9**	86.2	73.9	71.0	65.0	70.4
RangeFormer [24]	73.3	96.7	69.4	73.7	59.9	66.2	78.1	75.9	58.1	92.4	73.0	78.8	42.4	92.3	70.1	86.6	73.3	72.8	66.4	66.6
UniSeg (Ours)	**75.2**	**97.9**	**71.9**	75.2	**63.6**	**74.1**	78.9	74.8	60.6	92.6	**74.0**	79.5	46.1	93.4	72.7	**87.5**	**76.3**	**73.1**	**68.3**	68.5

Table 3: Quantitative results of UniSeg and SoTA LiDAR semantic segmentation methods on the nuScenes *test* set.

Method	mIoU	barr	bicy	bus	car	const	motor	ped	cone	trail	truck	driv	other	walk	terr	made	veg
PMF [63]	77.0	82.0	40.0	81.0	88.0	64.0	79.0	80.0	76.0	81.0	67.0	97.0	68.0	78.0	74.0	90.0	88.0
Cylinder3D [62]	77.2	82.8	29.8	84.3	89.4	63.0	79.3	77.2	73.4	84.6	69.1	97.7	70.2	80.3	75.5	90.4	87.6
AMVNet [33]	77.3	80.6	32.0	81.7	88.9	67.1	84.3	76.1	73.5	84.9	67.3	97.5	67.4	79.4	75.5	91.5	88.7
SPVCNN [43]	77.4	80.0	30.0	91.9	90.8	64.7	79.0	75.6	70.9	81.0	74.6	97.4	69.2	80.0	76.1	89.3	87.1
AF2S3Net [9]	78.3	78.9	52.2	89.9	84.2	77.4	74.3	77.3	72.0	83.9	73.8	97.1	66.5	77.5	74.0	87.7	86.8
2D3DNet [17]	80.0	83.0	59.4	88.0	85.1	63.7	84.4	82.0	76.0	84.8	71.9	96.9	67.4	79.8	76.0	**92.1**	89.2
GASN [54]	80.4	85.5	43.2	90.5	**92.1**	64.7	86.0	83.0	73.3	83.9	75.8	97.0	71.0	**81.0**	**77.7**	91.6	**90.2**
2DPASS [52]	80.8	81.7	55.3	92.0	91.8	73.3	86.5	78.5	72.5	84.7	75.5	97.6	69.1	79.9	75.5	90.2	88.0
LidarMultiNet [53]	81.4	80.4	48.4	**94.3**	90.0	71.5	87.2	**85.2**	**80.4**	**86.9**	74.8	**97.8**	67.3	80.7	76.5	**92.1**	89.6
UniSeg (Ours)	**83.5**	**85.9**	**71.2**	92.1	91.6	**80.5**	**88.0**	80.9	76.0	86.3	**76.7**	97.7	**71.8**	80.7	76.7	91.3	88.8

4.5. Overall Objective

The overall loss function is comprised of four terms, *i.e.*, the cross-entropy loss, the Lovasz-softmax loss [2], the heatmap regression via MSE loss, and the offset map regression by L1 loss, *i.e.*,

$$\mathcal{L} = \mathcal{L}_{\text{wce}} + \alpha\mathcal{L}_{\text{lovasz}} + \beta\mathcal{L}_{\text{heatmap}} + \gamma\mathcal{L}_{\text{offset}}, \quad (5)$$

where α, β, and γ are the loss coefficients to balance the effect of each loss term.

5. Experiments

Datasets. Following the practice of popular LiDAR segmentation models [62, 19, 20], we conduct experiments on three popular benchmarks, *i.e.*, nuScenes [3, 15], SemanticKITTI [1], and Waymo Open [41]. For nuScenes, it consists of 1000 driving scenes where 850 scenes are selected for training and validation, and the remaining 150 scenes are taken as the testing split. 16 classes are utilized for LiDAR semantic segmentation after merging similar classes and eliminating infrequent classes. As to SemanticKITTI, it has 22 point cloud sequences. Sequences 00 to 10, 08, and 11 to 21 are used for training, validation, and testing, respectively. 19 classes are chosen for training and evaluation after merging classes with distinct moving statuses and discarding classes with very few points. The Waymo Open Dataset (WOD) has 798, 202, and 150 sequences for training, validation, and testing, respectively. The duration of each sequence is 20 seconds and the frame rate is 10 Hz. However, for the 3D semantic segmentation task, not all frames are provided with 3D segmentation annotations. Specifically, only the last frame of a fixed number of frames is annotated. The number of annotated frames for training and validation is 23691, and 5976, respectively. The total number of classes is 23, including one ignored and 22 valid semantic categories. Note that both the first return and second return of the point cloud need to be segmented.

Evaluation Metrics. Following the practice of [20, 62], we adopt the Intersection-over-Union (IoU) of each class and mIoU of all classes as the evaluation metric. The IoU of class i is calculated via $IoU_i = \frac{TP_i}{TP_i+FP_i+FN_i}$, where TP_i, FP_i and FN_i denote the true positive, false positive and false negative of class i, respectively. For panoptic segmentation, we adopt the Panoptic Quality (PQ) as the main metric.

Implementation Details. For the point cloud branch, we first construct the point-voxel backbone based on the Minkowski-UNet34 [10]. Then, we add the range-image branch, *i.e.*, SalsaNext [13], to the point-voxel network and perform point-voxel-range fusion at four levels. The number of training epochs is set as 36 and the initial learning rate is set as 0.12. We use SGD as the optimizer. We use 1 epoch to warm up the network and adopt the cosine learning rate schedule for the remaining epochs. The momentum is set at 0.9 and weight decay is set at 0.0001. The voxel size is set as 0.05 for SemanticKITTI and WOD, and 0.1 for nuScenes. The gradient norm clip is set to 10 to stabilize the training process. α, β and γ are set as 1, 100, and 10, respectively. As

Table 4: Quantitative results of UniSeg and SoTA LiDAR panoptic segmentation methods on SemanticKITTI *test* set.

Method	PQ
Panoptic-PolarNet [57]	54.1
DS-Net [19]	55.9
EfficientLPS [40]	57.4
GP-S3Net [39]	60.0
SCAN [49]	61.5
Panoptic-PHNet [30]	64.6
UniSeg (Ours)	**67.2**

Table 5: Quantitative results of UniSeg and SoTA LiDAR panoptic segmentation methods on nuScenes *test* set.

Method	PQ
EfficientLPS [40]	62.4
Panoptic-PolarNet [57]	63.6
SPVNAS [43] + CenterPoint [55]	72.2
Cylinder3D++ [62] + CenterPoint [55]	76.5
AF2S3Net [9] + CenterPoint [55]	76.8
SPVCNN++ [43]	79.1
LidarMultiNet [53]	81.4
Panoptic-PHNet [30]	**81.5**
UniSeg (Ours)	78.4

to data augmentation of the point cloud branch, we employ random flip, random scaling, random translation as well as LaserMix [27] and PolarMix [47] to increase the diversity of training samples. For the RGB image branch, we use ImageNet-pretrained ResNet-34 as the feature extractor. The parameters in the image branch are trainable. More details are put in the supplementary.

Multi-Modal Fusion Baselines. We take classical early fusion, PointPainting [45] and PointAugmenting [46] as multi-modal fusion baselines. Early fusion conducts input-level fusion and we select two early fusion variants, *i.e.*, addition and concatenation of input signals. PointPainting appends the point cloud with the semantic segmentation scores while PointAugmenting fuses the point cloud with the image features of the segmentation branch.

5.1. Comparative Study

Quantitative Results. We summarize the performance of UniSeg and state-of-the-art LiDAR segmentation methods in Table 2-6. For LiDAR semantic segmentation, our UniSeg outperforms the competitive 2DPASS [52] by **2.3** mIoU. For classes of bicycle, motorcycle, and other vehicles, UniSeg is at least 8 IoU higher than 2DPASS [52]. As to panoptic segmentation, UniSeg achieves 67.2 PQ, surpassing the rival Panoptic-PHNet [30] by **2.6** PQ. On the nuScenes benchmark, UniSeg obtains **83.5** mIoU on the LiDAR semantic segmentation task and outperforms the second place, *i.e.*, LidarMultiNet [53], by **2.1** mIoU. As for panoptic segmentation, our UniSeg achieves 78.4 PQ and is on par with competitive panoptic segmentation algorithms such as SPVCNN++. Encouraging results are also observed in the WOD val set. UniSeg obtains 69.6 mIoU and is **2.2** mIoU higher than SPVCNN[43]. The impressive experimental results strongly prove the effectiveness of the presented multi-modal fusion network.

Table 6: Quantitative results of UniSeg and SoTA LiDAR semantic segmentation methods on the WOD *val* set. Methods with * denote our implementations.

Method	mIoU
Point Transformer* [56]	63.3
Cylinder3D* [62]	66.0
SPVCNN* [43]	67.4
UniSeg (Ours)	**69.6**

Table 7: The comparisons between efficiency (run-time) and accuracy (mIoU) on the SemanticKITTI *val* set.

Method	#Param	Latency	mIoU
Cylinder3D [62]	56.3M	75.1ms	65.9
MinkowskiNet [11]	21.7M	48.4ms	61.1
SPVCNN [43]	21.8M	52.4ms	63.8
UniSeg 0.2× **(Ours)**	28.8M	84.6ms	**67.0**
UniSeg 1.0× **(Ours)**	147.6M	145.0ms	**71.3**

Table 8: Comparison with different multi-modal feature fusion strategies on the SemanticKITTI *val* set. Methods with * denote our implementations.

Method	mIoU	Δ
Early Fusion Add (Baseline)	70.1	+0.0
Early Fusion Concat	69.4	−0.7
PointPainting* [45]	70.4	+0.3
PointAugmenting* [46]	70.5	+0.4
LMA (Ours)	**71.3**	+1.2

Comparisons of Efficiency and Accuracy. We provide comparisons of efficiency and accuracy as shown in Table 7, our UniSeg_0.2× achieves the best accuracy when the parameters and latency are comparable to other methods. Note that UniSeg_0.2× is produced from the original UniSeg model by pruning 80% channels for each layer. Besides, when increasing the parameters, the accuracy is further improved (UniSeg). All models are tested at NVIDIA A100 GPU.

Is the Implementation Optimal? We would like to show that the implementation achieves the best performance after trials and errors. Specifically, **For the LMA module:** considering the calibration errors caused by the imperfect

Table 9: Influence of different modalities and views.

Voxel	Point	Range image	RGB Image	mIoU
✓				68.4
	✓			13.7
		✓		55.8
✓			✓	69.7
		✓	✓	58.1
✓	✓			68.5
✓	✓	✓		69.7
✓	✓	✓	✓	**71.3**

Table 10: Robustness on the SemanticKITTI *val* set. The symbol * denotes calibration matrices with noises.

Method	Add	LMA	Add*	LMA*
mIoU	70.1	**71.3**	68.5	71.0

calibration matrices between the LiDAR and the camera. We have made several attempts to alleviate this issue (Table 8). Firstly, we directly added or concatenated the image-point feature, and achieved +0.4 mIoU and -0.3 mIoU, respectively. Secondly, we adopt PointPainting [45] and PointAugmenting [46] to fuse feature, the improvement is 0.7 mIoU and 0.8 mIoU, respectively, but these fusion methods are sensitive to calibration errors. Thirdly, We tried the Self-attention operation. However, it suffers from the high computational cost introduced by the global-wise attention calculation. Lastly, we adopt the Deformable cross-attention in our method due to its efficiency and effectiveness. As shown in Table 8, the LMA module improved **1.6** mIoU and outperformed add, concatenate, PointPainting, and PointAugmenting by 1.2, 1.9, 0.9, and 0.8 in mIoU, respectively.

For LVA module: We explore how to leverage the advantages of different modality data. Firstly, we conduct the baseline method, i.e., it transfers all modality data into the point-view and then directly adds or concatenates them, the performance is 70.4 mIoU and 70.5 mIoU, respectively. Secondly, we tried self-attention for feature fusion but could not achieve improvement. Lastly, we design the LVA module to adaptively fuse the different modality data based on the learned attention weights. As shown in Table 11, the improvement is **0.9** mIoU compared to the direct addition and concatenation.

5.2. Ablation Study

We perform an ablation study to verify the effect of each modality/view and different cross-view fusion variants on the final performance. The following experiments are conducted in the SemanticKITTI validation set.

Effect of Each Modality. We summarize the influence of each modality as well as their combinations on the final performance in Table 9. From the first three rows, we can see

Table 11: Comparisons among cross-view fusion strategies.

Method	**mIoU**	Δ
Add (Baseline)	70.4	+0.0
Concat	70.5	+0.1
Self-Attention	70.4	+0.0
LVA	**71.3**	+0.9

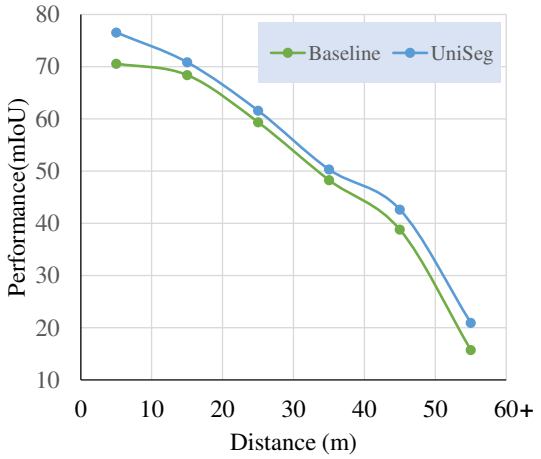

Figure 5: Comparison between the single-modal baseline and UniSeg with different distances on SemanticKITTI.

that the voxel branch exhibits much better performance than the other two representations, showing the indispensable role of the voxel representation. Fusing three views of the point cloud with images yield the best performance, demonstrating the value of every single modality on the segmentation results. Besides, our UniSeg also outperforms the single-modal baseline in different distances (Fig. 5). Obviously, the baseline degrades at a long distance due to more sparsity. And UniSeg consistently outperforms the uni-modal baseline, strongly demonstrating the value of the multi-modal representation.

Fusion Strategies. We compare our proposed LMA module with other fusion strategies as shown in Table 8, it brings a larger improvement than other methods and outperforms 1.2 mIoU than baseline. Notably, when we used UniSeg_0.2× to compare the LMA module with PointPainting, the LMA module was **1.5** mIoU higher than PointPainting, which directly demonstrates the benefits of the LMA module. With the help of the LVA module, the point-, voxel-, and range-view features are more effectively fused compared with other fusion methods as shown in Table 11.

Robustness to calibration error. We add Gaussian noise to the calibration matrices to evaluate the robustness. As shown in Table 10, UniSeg drops **0.3** mIoU while the addition operation drops **1.6** mIoU, indicating the LMA module is more tolerant to calibration noise.

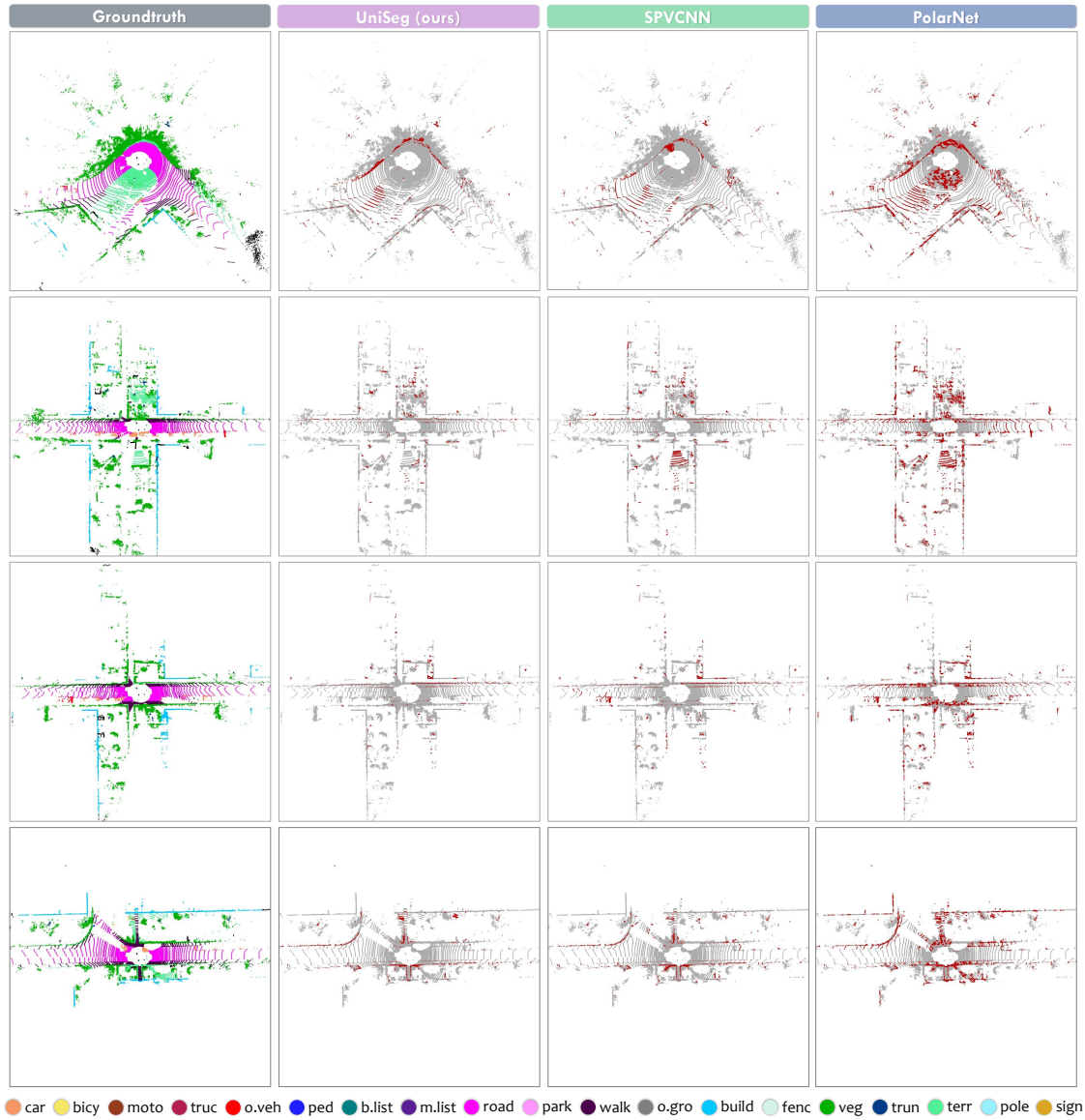

Figure 6: **Qualitative comparisons** with SPVCNN [42] and PolarNet [58] through **error maps**. To highlight the differences, the **correct** / **incorrect** predictions are painted in **gray** / **red**, respectively. Each scene is visualized from the LiDAR bird's eye view and covers a region of size 50m by 50m, centered around the ego-vehicle. Best viewed in colors.

5.3. Qualitative Results

We provide qualitative comparisons with SPVCNN [42] and PolarNet [58] through error maps in Fig. 6. Upon examining the results, it becomes evident that our approach demonstrates superior performance while maintaining minimal segmentation errors across each sampled frame.

6. Conclusion

We propose a unified multi-modal LiDAR segmentation network, dubbed UniSeg, that makes the first attempt to take RGB images and three views of the point cloud as input, and performs semantic and panoptic segmentation simultaneously. To fully leverage the information of different modalities data, we present the cross-Modal Association module (LMA) and the Learnable cross-View Association module (LVA). Equipped with LMA and LVA, UniSeg achieves compelling performance in three popular LiDAR segmentation benchmarks and ranks 1st in two open challenges.

Acknowledgements. This work is supported by the Science and Technology Commission of Shanghai Municipality (grant No. 22DZ1100102).

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019.

[2] Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018.

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[4] Nenglun Chen, Lingjie Liu, Zhiming Cui, Runnan Chen, Duygu Ceylan, Changhe Tu, and Wenping Wang. Unsupervised learning of intrinsic structural representation points. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9121–9130, 2020.

[5] Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, Xinge Zhu, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models. *arXiv preprint arXiv:2306.03899*, 2023.

[6] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[7] Runnan Chen, Xinge Zhu, Nenglun Chen, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. Zero-shot point cloud segmentation by transferring geometric primitives. *arXiv preprint arXiv:2210.09923*, 2022.

[8] Runnan Chen, Xinge Zhu, Nenglun Chen, Dawei Wang, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. Towards 3d scene understanding by referring synthetic models. *arXiv preprint arXiv:2203.10546*, 2022.

[9] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. (af)2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12547–12556, 2021.

[10] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[11] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[12] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3d object detection. https://github.com/open-mmlab/mmdetection3d, 2020.

[13] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *International Symposium on Visual Computing*, pages 207–222, 2020.

[14] Khaled El Madawi, Hazem Rashed, Ahmad El Sallab, Omar Nasr, Hanan Kamel, and Senthil Yogamani. Rgb and lidar fusion based 3d semantic segmentation for autonomous driving. In *IEEE Intelligent Transportation Systems Conference*, pages 7–12, 2019.

[15] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022.

[16] Biao Gao, Yancheng Pan, Chengkun Li, Sibo Geng, and Huijing Zhao. Are we hungry for 3d lidar data for semantic segmentation? a survey of datasets and methods. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):6063–6081, 2021.

[17] Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. Learning 3d semantic segmentation with only 2d image supervision. In *International Conference on 3D Vision*, pages 361–372, 2021.

[18] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4338–4364, 2020.

[19] Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Lidar-based panoptic segmentation via dynamic shifting network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13090–13099, 2021.

[20] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8479–8488, 2022.

[21] Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022.

[22] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022.

[23] Rui Jiang, Ruixiang Zhu, Hu Su, Yinlin Li, Yuan Xie, and Wei Zou. Deep learning-based moving object segmentation: Recent progress and research prospects. *Machine Intelligence Research*, pages 1–35, 2023.

[24] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. *arXiv preprint arXiv:2303.05367*, 2023.

[25] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. *arXiv preprint arXiv:2303.17597*, 2023.

[26] Lingdong Kong, Niamul Quader, and Venice Erin Liong. Conda: Unsupervised domain adaptation for lidar segmentation via regularized domain concatenation. In *IEEE International Conference on Robotics and Automation*, pages 9338–9345, 2023.

[27] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023.

[28] Georg Krispel, Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Fuseseg: Lidar point cloud segmentation fusing multi-modal data. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1874–1883, 2020.

[29] Jiale Li, Hang Dai, and Yong Ding. Self-distillation for robust lidar semantic segmentation in autonomous driving. In *European Conference on Computer Vision*, pages 659–676, 2022.

[30] Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, and Dan Zhang. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11809–11818, 2022.

[31] Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Qin Chen, Yikang Li, Yu Qiao, and Liang He. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17524–17534, 2023.

[32] Xin Li, Botian Shi, Yuenan Hou, Xingjiao Wu, Tianlong Ma, Yikang Li, and Liang He. Homogeneous multi-modal feature fusion and interaction for 3d object detection. In *European Conference on Computer Vision*, pages 691–707, 2022.

[33] Venice Erin Liong, Thi Ngoc Tho Nguyen, Sergi Widjaja, Dhananjai Sharma, and Zhuang Jie Chong. Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. *arXiv preprint arXiv:2012.04934*, 2020.

[34] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *arXiv preprint arXiv:2306.09347*, 2023.

[35] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *arXiv preprint arXiv:1907.03739*, 2019.

[36] Yuhang Lu, Qi Jiang, Runnan Chen, Yuenan Hou, Xinge Zhu, and Yuexin Ma. See more and know more: Zero-shot point cloud segmentation via multi-modal visual data. *arXiv preprint arXiv:2307.10782*, 2023.

[37] Tao Ma, Xuemeng Yang, Hongbin Zhou, Xin Li, Botian Shi, Junjie Liu, Yuchen Yang, Zhizheng Liu, Liang He, Yu Qiao, et al. Detzero: Rethinking offboard 3d object detection with long-term sequential point clouds. *arXiv preprint arXiv:2306.06023*, 2023.

[38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[39] Ryan Razani, Ran Cheng, Enxu Li, Ehsan Taghavi, Yuan Ren, and Liu Bingbing. Gp-s3net: Graph-based panoptic sparse semantic segmentation network. In *IEEE/CVF International Conference on Computer Vision*, pages 16076–16085, 2021.

[40] Kshitij Sirohi, Rohit Mohan, Daniel Büscher, Wolfram Burgard, and Abhinav Valada. Efficientlps: Efficient lidar panoptic segmentation. *IEEE Transactions on Robotics*, 38(3):1894–1914, 2021.

[41] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[42] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, pages 6105–6114, 2019.

[43] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, pages 685–702, 2020.

[44] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020.

[45] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4604–4612, 2020.

[46] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021.

[47] Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. Polarmix: A general data augmentation technique for lidar point clouds. *arXiv preprint arXiv:2208.00223*, 2022.

[48] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 16024–16033, 2021.

[49] Shuangjie Xu, Rui Wan, Maosheng Ye, Xiaoyi Zou, and Tongyi Cao. Sparse cross-scale attention network for efficient lidar panoptic segmentation. *arXiv preprint arXiv:2201.05972*, 2022.

[50] Yiteng Xu, Peishan Cong, Yichen Yao, Runnan Chen, Yuenan Hou, Xinge Zhu, Xuming He, Jingyi Yu, and Yuexin Ma. Human-centric scene understanding for 3d large-scale scenarios. *arXiv preprint arXiv:2307.14392*, 2023.

[51] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 3101–3109, 2021.

[52] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted

semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, 2022.

[53] Dongqiangzi Ye, Zixiang Zhou, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarmultinet: Towards a unified multi-task network for lidar perception. *arXiv preprint arXiv:2209.09385*, 2022.

[54] Maosheng Ye, Rui Wan, Shuangjie Xu, Tongyi Cao, and Qifeng Chen. Efficient point cloud segmentation with geometry-aware sparse networks. In *European Conference on Computer Vision*, pages 196–212, 2022.

[55] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11784–11793, 2021.

[56] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021.

[57] Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13194–13203, 2021.

[58] Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13194–13203, 2021.

[59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[60] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020.

[61] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Wei Li, Yuexin Ma, Hongsheng Li, Ruigang Yang, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6807–6822, 2022.

[62] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9939–9948, 2021.

[63] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 16280–16290, 2021.