

Align and Blend: A Unified Multi-Modal LiDAR Segmentation Network

Chuanxu Wang[✉], Jiajiong Li[✉], Xin Chen, Da Song[✉], and Binghui Wang

Abstract—The multimodal fusion of point cloud data is a challenging task in 3D computer vision, particularly in fine object segmentation. In this study, we propose an innovative unified multimodal LiDAR segmentation network, named Align and Blend (A2Blend for short), which skillfully integrates three representative forms of point clouds: point view, voxel view, and range view. The core innovation of A2Blend lies in addressing two primary tasks: Align and Blend. For the “Align” task, we have carefully designed a learnable cross-modal association module, whose core is a unique “Cross-Modal Triplet Alignment Loss” mechanism. To the best of our knowledge, this is the first application of such a mechanism in the field of LiDAR segmentation. This mechanism applies principles from contrastive learning, promoting the tight clustering of similar semantic groups both within and across modalities, while increasing the separation between dissimilar groups by expanding the distance between similar and dissimilar sample clusters in the feature space. This significantly enhances the discriminative power and representational capacity of the feature embeddings. For the “Blend” task, we propose a fusion strategy that integrates intragroup self-attention within modalities and intergroup self-attention across modalities. This approach combines key concepts from standard self-attention and cross-self-attention mechanisms to achieve more comprehensive multimodal fusion. The segmentation performance on two large-scale outdoor datasets and one indoor dataset surpasses that of most state-of-the-art algorithms.

Index Terms—Multi-modal LiDAR segmentation, align and blend, cross-modal triplet alignment loss, self attention.

I. INTRODUCTION

THE LiDAR-based semantic segmentation, which aims to assign semantic labels to each input point, is a crucial component in autonomous driving, digital cities, and service robots [1], [2]. With the advent of deep learning, numerous approaches [3], [4], [5], [6], [7], [8], [9] have been proposed, rapidly dominating various benchmarks such as SemanticKITTI [10] and nuScenes [11]. Point clouds [3], range images [13], voxel grids [4], and RGB images [12] are

four commonly used modalities, each with its own advantages and limitations. Point clouds provide reliable and accurate depth information, and can be processed in point view, voxel view, and range view. Specifically, the point view representation retains the full point information but struggles to capture neighboring point features efficiently due to the irregular (unordered) nature of the points. The voxel view discretizes the point cloud into voxel units, preserving a regular structure, however, voxelization can lead to significant information loss, especially when the voxel size is large. Range view representations are dense and compact, and can be processed using highly optimized 2D convolutions. However, spherical projection inherently distorts the original 3D geometric information, leading to inaccuracies in spatial data and issues such as occlusion and obscuration. Although RGB images provide rich color, boundary, texture, and semantic information, their usage in multimodal LiDAR segmentation tasks is limited by the lack of RGB data in many datasets. Furthermore, collecting RGB data usually requires additional processing, which restricts its extension as a supplementary modality.

Evidently, input data from multimodal and multiview point clouds are complementary to each other. Therefore, fully utilizing the synthesized information is beneficial for enhancing sensing capabilities. However, this cross-modal and cross-view fusion paradigm has not been thoroughly explored in the field of LiDAR segmentation. Current multimodal fusion methods primarily focus on fusing two views [5], [14] or more views [12]. In the former, the fusion of RGB and range images is commonly employed. Alternative representations, such as voxel views and point views of LiDAR point clouds, retain the original data structure and provide fine-grained spatial information, but have been overlooked by previous approaches. Moreover, when RGB modal data is introduced, RGB images and range images are often fused in a hard correlation manner using a calibration matrix, making them susceptible to calibration errors [12]. In the case of the latter, although many existing approaches achieve highly satisfactory performance, they typically focus on how to fuse data from different modalities [15], often overlooking the inherent differences between these modalities. This limits the integration of innovative insights from other research fields. Taking these factors into account, in our network, we discard RGB images and instead choose to fuse point-based views, voxel-based views, and range-based views. For the range-based approach, we use only the encoder structures to extract high-dimensional semantic information.

Received 10 June 2024; revised 14 October 2024 and 5 December 2024; accepted 21 January 2025. Date of publication 11 February 2025; date of current version 5 May 2025. This work was supported by the National Natural Science Foundation of China under Grant 61672305. The Associate Editor for this article was C. Wen. (Chuanxu Wang and Jiajiong Li contributed equally to this work.) (Corresponding author: Jiajiong Li.)

The authors are with the School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China (e-mail: wangchuanxu_qd@qust.edu.cn; 19937731387@163.com; chenxin13618485599@163.com; SongDa0319@163.com; 15154705338@163.com).

Digital Object Identifier 10.1109/TITS.2025.3533747

Contrastive Language-Image Pre-Training (CLIP) [16] introduces a novel approach by using the inherent differences between modalities (e.g., images and text) to establish connections through contrastive learning. This enhances the model's performance and generalization, facilitating cross-modal learning, which may offer a solution to the challenges mentioned earlier. MaskCLIP [17] extends CLIP to semantic segmentation, allowing for object segmentation with minimal modifications and without additional training. PointCLIP [18] demonstrates CLIP's zero-shot classification capabilities can transfer from 2D images to 3D point clouds by projecting point cloud data into 2D depth maps. These studies highlight CLIP's potential in 2D segmentation and 3D classification, though its application to 3D scene understanding remains underexplored.

Motivated by the insights from CLIP, we explore for the first time how to leverage its core principles to apply contrastive learning to the point cloud multiview fusion segmentation task. In point cloud multiview fusion segmentation, where there is a large number of points, applying contrastive learning presents two key challenges: (1) how to divide anchor, positive, and negative samples to ensure a relative balance between positive and negative samples, and (2) how to design a contrastive loss that ensures strong internal connections between data from different modalities. To address the first challenge, since the semantic structure of points in the network remains unordered throughout the training process, and the number of points can reach up to a million, selecting anchor, positive, and negative samples at the point-by-point level would be impractical. Instead, we propose a Semantic-guided Multimodal Feature Disentanglement module, which first soft-groups the point features of the three modalities based on their ground truth labels. Then, it divides the anchor, positive, and negative samples into groups (or clusters). On this basis, we design a novel Dual Contrast Loss to establish strong connections between the different modalities. This process is referred to as the "Align" task.

The "Blend" task is crucial for multimodal fusion. Simple fusion strategies are insufficient for handling complex downstream tasks. Therefore, for the "Blend" task, we employ self-attention to model interactions between different point cloud modalities. Specifically, for a single modality, we use standard self-attention to capture intra-cluster features of the same semantics. For intermodal interactions, we adopt cross-self-attention to enable cross-modal learning at the feature cluster level.

The key contributions of this paper are as follows:

- We propose an innovative unified multimodal LiDAR segmentation network that, for the first time, applies contrastive learning to multimodal fusion in 3D segmentation tasks. Specifically, we introduce a novel Cross-Modal Triplet Alignment Loss and a Semantic-Guided Multimodal Feature Disentanglement module, which efficiently groups point features.
- We design Dual Contrast Loss functions that, for a single modality, separate clusters of different semantics at a high semantic level, and for multimodal interactions, bring closer clusters of the same semantic features while

pushing apart clusters of different semantics between modalities.

- Our cross-modal fusion strategy leverages standard self-attention to capture relationships between homogeneous semantic points within modalities and cross-self-attention to model connections between dissimilar semantics across modalities, enabling better reasoning across multimodal data.

II. RELATED WORK

A. Lidar-Based Semantic Scene Understanding

Semantic segmentation is a key task in LiDAR-based scene understanding, with most methods relying solely on point cloud data [3]. For instance, Cylinder3D [5] segments point clouds using cylindrical shapes and feeds these features into a Unet-based backbone, while SPVCNN [19] enhances segmentation by combining point and voxel-level features. LidarMultiNet [20] integrates LiDAR segmentation, panoramic segmentation, and 3D object detection into a single framework, and UniSeg [12] incorporates RGB modalities using a calibration matrix for improved performance. While these approaches focus on combining inter-modal data (e.g., via addition, concatenation, or self-attention), they often overlook the unique characteristics of each modality. In contrast, our network fully considers both the similarities and differences between modalities, allowing for a more effective fusion mechanism and broader application potential.

B. Revisiting Contrastive Learning

1) *Zero-Shot Learning in 3D*: The target of zero-shot learning (ZSL) is to recognize objects that have not been seen in the training set. Many efforts have worked on 2D recognition tasks [21], [22], while a few have focused on performing ZSL in the 3D domain [23], [24]. [23] explored transductive zero-shot segmentation for 3D scene understanding for the first time. More recently, several studies have introduced CLIP to zero-shot learning. MaskCLIP [17] investigated the use of CLIP to help with a 2D dense prediction task and demonstrated encouraging zero-shot semantic segmentation performance. PointCLIP [18] was a pioneering work in applying CLIP to 3D recognition, showing impressive performance on zero-shot and few-shot performance. However, the above work does not apply contrastive Learning to the segmentation of 3D point clouds. This still deserves further exploration in the field of 3D segmentation.

2) *Cross-Modal Knowledge Distillation*: Recently, more and more research has focused on transferring knowledge from 2D images to 3D point clouds for self-supervised representation learning [25], [26]. PPKT [25] utilizes InfoNCE loss to help 3D networks extract rich knowledge from 2D image backbones. SLidR [26] further introduces superpixels to facilitate cross-modal knowledge refinement. PVKD [4] is the first to apply knowledge distillation in both point and voxel modalities, thus enabling cross-modal interactions. But it only considered the simple two modalities. In our network, we utilize the knowledge of contrastive Learning to establish more comprehensive interactions among the three modalities for better interaction.

III. METHOD

A. Motivation and Overview

1) *Motivation*: Contrastive learning has been widely applied in self-supervised learning, transfer learning, and multimodal fusion to enhance model representation quality and performance. Building on this, we are the first to apply contrastive learning to 3D LiDAR multimodal fusion segmentation and explore its potential in 3D multimodal fusion.

2) *Overview*: Our proposed algorithm, A2Blend (Fig. 1), consists of four components: the Learnable Cross-Modal Association module (LCA), the Semantic-guided Multimodal Feature Disentanglement module, the lightweight self-attention mechanism (Blend), and a downstream network for cross-modal interactions. The LCA module features a three-branch network for multi-channel interaction, with range-based, point-based, and voxel-based encoders. The range and voxel branches use a semi-Unet structure to extract contextual information and perform four downsampling stages for high-level semantic representations. The point branch employs a simple PointNet architecture with multiple MLPs. After obtaining rich high-dimensional semantic information from all three modalities, the Semantic-guided Multimodal Feature Disentanglement module processes and fuses the features. The collaborative lightweight self-attention mechanism then passes these fused features to the segmentation-based network (point-to-voxel). The following sections will provide more detailed implementation.

B. Align

1) *Learnable Cross-Modal Association*: The original point cloud data can be transformed into different views, typically voxel view and range view. We treat the point view as an intermediate “transit station”, which establishes a connection between these views. Any view of the point cloud can be regarded as a mapping of the original point cloud. Subsequently, a multi-view representation index and multi-view feature propagation are established between the three views using a hash function.

We applied the same techniques of multi-view indexing, multi-view feature propagation, and multi-view interaction in the LCA module, following the ingenious ideas presented in [6]. Specifically, we note that there are two instances of multi-view interactions within the LCA modules, occurring at both the low-level semantic layer and the high-level semantic layer. In contrast to the previous work, we employ a lightweight self-attention mechanism to replace the earlier gate fusion operation. This design reduces model complexity and facilitates easier deployment across various hardware platforms. This strategy will be further detailed in Section III-C.

2) *Cross-Modal Triplet Alignment Loss*: Recent approaches, such as CLIP, using contrastive learning have demonstrated exceptional performance in tasks like intelligent Q&A, image-text matching, image segmentation, 2D object detection, and classification. However, there is still a gap in the development of 3D LiDAR point cloud segmentation. Inspired by this, we introduce a contrastive

Algorithm 1 Feature Clustering and Padding in Mini-Batches

Input: X^R, X^P, X^V, T , and the current mini-batch’s semantic class count cc

Output: Sequences after clustering and padding: X_n^R, X_n^P, X_n^V

```

1:  $X_n^R \leftarrow \emptyset, X_n^P \leftarrow \emptyset, X_n^V \leftarrow \emptyset$   $n \sim [1, 2, \dots, cc]$ 

2:  $i \sim [1, 2, \dots, n]$ 
3:  $F^{ri} \leftarrow$  Feature Clustering Generator ( $X^R, T$ )
4:  $F^{pi} \leftarrow$  Feature Clustering Generator ( $X^P, T$ )
5:  $F^{vi} \leftarrow$  Feature Clustering Generator ( $X^V, T$ )
6:  $m \leftarrow \text{NumMax}(F^{p1}, F^{p2}, \dots, F^{pn})$ 
7: for  $l$  in  $i$  do
8:   if  $\text{NumMax}(F^{pl}) < m$  then
9:      $h \leftarrow m - \text{NumMax}(F^{pl})$ 
10:     $F_{padding}^{rl} \leftarrow$  Feature Padding Generator( $F^{rl}, h$ )
11:     $F_{padding}^{pl} \leftarrow$  Feature Padding Generator( $F^{pl}, h$ )
12:     $F_{padding}^{vl} \leftarrow$  Feature Padding Generator( $F^{vl}, h$ )
13:  else
14:     $F^{rl} \leftarrow F^{rl}, F^{pl} \leftarrow F^{pl}, F^{vl} \leftarrow F^{vl}$ 
15:  end if
16:  if padding then
17:     $X_l^R \leftarrow F^{rl} \cup F_{padding}^{rl}, X_l^P \leftarrow F^{pl} \cup F_{padding}^{pl}, X_l^V \leftarrow$   

 $F^{vl} \cup F_{padding}^{vl}$ 
18:  else
19:     $X_l^R \leftarrow F^{rl}, X_l^P \leftarrow F^{pl}, X_l^V \leftarrow F^{vl}$ 
20:  end if
21:   $X_n^R \leftarrow X_n^R \cup \{X_l^R\}, X_n^P \leftarrow X_n^P \cup \{X_l^P\}, X_n^V \leftarrow X_n^V \cup \{X_l^V\}$ 
22: end for

```

learning mechanism in the multi-view fusion of point clouds for the first time.

a) *Semantic-guided multimodal feature disentanglement*: Given three sequences containing high-level semantic features from range-based, voxel-based, and point-based encoders, denoted as $X^R \in \mathbb{R}^{N \times C}$, $X^P \in \mathbb{R}^{N \times C}$, and $X^V \in \mathbb{R}^{N \times C}$, where N is the number of points and C is the number of channels, we perform two key operations:

- **Clustering of features** for each modality based on the true label T of the input data, where $T \in \mathbb{R}^N$.
- **Consistent padding** of the number of features within each feature cluster for each modality.

These steps are implemented in the following pseudo-code for Algorithm 1.

Finally, we obtain a pseudo-grouped feature representation for the range view, point view, and voxel view: $X^R \rightarrow \{X_1^R, X_2^R, \dots, X_n^R\}$, $X^P \rightarrow \{X_1^P, X_2^P, \dots, X_n^P\}$, $X^V \rightarrow \{X_1^V, X_2^V, \dots, X_n^V\}$, where n represents the number of semantic categories.

b) *Sample cluster assembly*: The clustering of samples is a crucial part of the contrastive learning process. To strengthen the relationships between different modalities, rather than using the traditional point-level division, we propose a new approach for clustering samples at the group level,

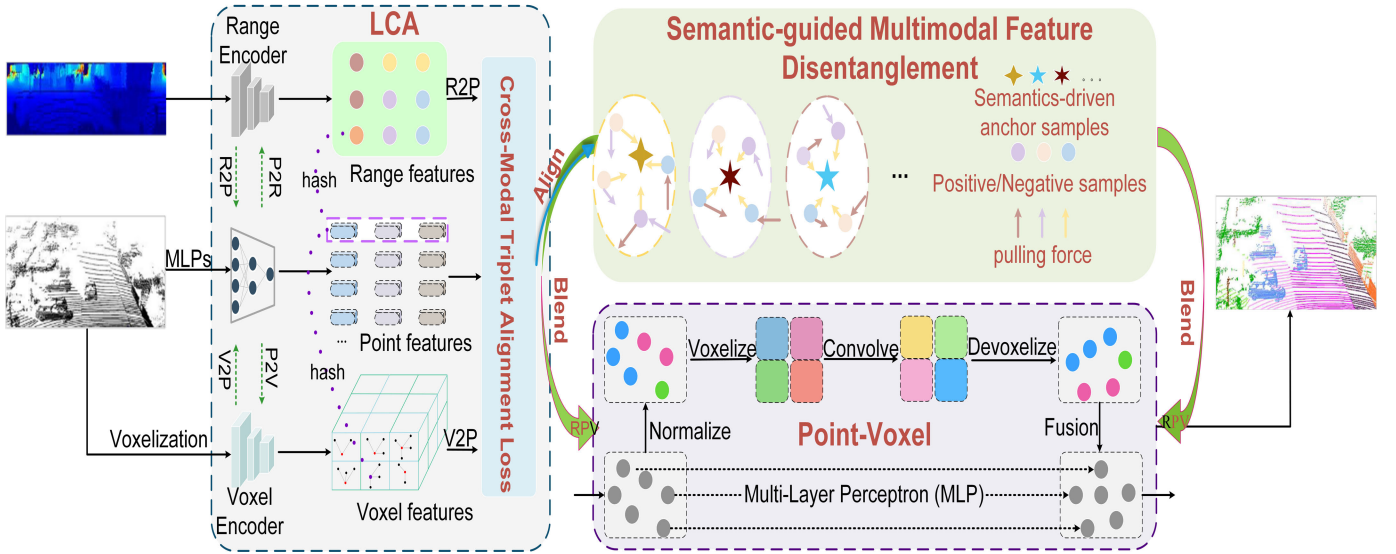


Fig. 1. The A2Blend method processes three input signals: voxel, range, and point views. At the core is the Learnable Cross-Modal Association (LCA) module, which facilitates two stages of cross-modal interactions between the 2D range encoder, 3D voxel encoder, and MLPs. Before the second interaction, our Semantic-guided Multimodal Feature Disentanglement mechanism is applied. Subsequently, the low-level semantic features from the point, range, and voxel views are fed into the RPV (Blend) module for cross-modal fusion of high-level semantic information. This fused information is then compressed and refined, allowing for further cross-modal interaction and fusion at both coarse and fine granularities between the point and voxel modalities. Best viewed in colors.

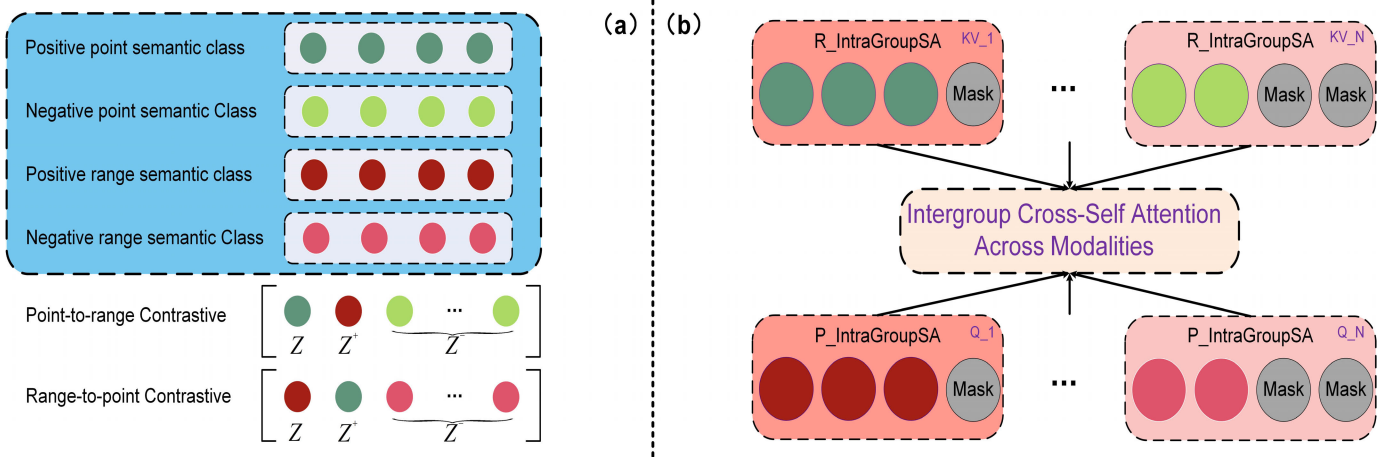


Fig. 2. Contrast pair selection procedure for selecting positive and negative semantic class features between range view and point view as shown in block (a). Our proposed cross-modal intergroup group attention mechanism is shown in block (b). The same procedure is applied to the point view and voxel view modes. Best viewed in colors.

as illustrated in Fig. 2. The same technique is applied between the point view and the voxel view. A more detailed method for dividing negative samples is outlined in Algorithm 2.

c) *Loss functions:* We employ three loss functions to train our model: classification loss, Lovász-Softmax loss, and the newly introduced Cross-Modal Triplet Alignment Loss.

The weighted cross-entropy loss [27] is given by:

$$L_{wce}(y, \tilde{y}) = - \sum_{i=1} a_i P(y_i) \log p(\tilde{y}), a_i = 1/\sqrt{v_i} \quad (1)$$

where v_i is the class frequency, $P(y_i)$ and $p(\tilde{y})$ are the true and predicted values, and $P(y_i)$ is an indicator of the correct class. This loss is suitable for multi-class tasks like semantic segmentation as it minimizes the difference between predicted and true distributions. The Lovász-Softmax loss [28]

is defined as:

$$L_{ls} = \frac{1}{|C|} \sum_{c \in C} J(e(c)) \quad (2)$$

where J extends Lovász's IoU, and $e(c) \in [0, 1]^p$ is the error vector for category c , with p being the point under consideration. This loss has shown effectiveness in various tasks like object detection and segmentation [29], and will be combined with other losses to improve accuracy.

Inspired by the success of contrastive learning in image-language pre-training tasks [30], we extend its use to point clouds across multiple modalities. We focus on interactions both within and across modalities. Using the two algorithms described earlier, we obtain clusters of anchor, positive, and negative samples for the point, range, and voxel views, represented as aps,pps,pns,ars,rps,rns,avs,vps,vns (see Fig. 2). The

Algorithm 2 Generate Negative Sample Clusters

Input: Sequences after clustering and padding: X_n^R, X_n^P, X_n^V
Output: Negative sample clusters: $X_n^{Rneg}, X_n^{Pneg}, X_n^{Vneg}$

```

1:  $l \sim [1, 2, \dots, cc]$ 
2: for  $j$  in  $l$  do
3:   for  $k$  in  $l$  do
4:     if  $X_j^R$  is ans and  $X_k^P$  is pps and  $j \neq k$  then
5:       remove  $X_j^R$  from  $X_n^R$ 
6:        $X_n^{Rneg} \leftarrow X_n^R$ 
7:     end if
8:   end for
9: end for
10: Similarly,  $X_n^{Pneg}$  and  $X_n^{Vneg}$  can be derived

```

Cross-Modal Triplet Alignment Loss is defined as:

$$L_{Dclpr} = E_{Z \sim pps_j, Z^+ \sim rps_j, Z^- \sim pns_{j \neq k}} \mathcal{L}(Z, Z^+, Z^-) + E_{Z \sim rps_j, Z^+ \sim pps_j, Z^- \sim rns_{j \neq k}} \mathcal{L}(Z, Z^+, Z^-) \quad (3)$$

$$L_{Dclpv} = E_{Z \sim pps_j, Z^+ \sim vps_j, Z^- \sim pns_{j \neq k}} \mathcal{L}(Z, Z^+, Z^-) + E_{Z \sim vps_j, Z^+ \sim pps_j, Z^- \sim vns_{j \neq k}} \mathcal{L}(Z, Z^+, Z^-) \quad (4)$$

$$L_{cmtal} = L_{Dclpr} + L_{Dclpv} \quad (5)$$

where j and k represent pseudo-labels or semantic clusters, and $\mathcal{L}(Z, Z^+, Z^-)$ is the standard contrastive loss:

$$\mathcal{L}(Z, Z^+, Z^-) = -\log\left(\frac{\exp(Z^T * Z^+ / \tau)}{\exp(Z^T * Z^+ / \tau) + \sum_k \exp(Z^T * Z^- / \tau)}\right) \quad (6)$$

where τ is a learnable temperature factor. Finally, the total loss is a combination of the three loss functions:

$$L_{total} = \beta_{wce} L_{wce} + \beta_{ls} L_{ls} + \beta_{cmtal} L_{cmtal} \quad (7)$$

where β_{wce} , β_{ls} , and β_{cmtal} are hyperparameters that balance the three loss components.

C. Blend

The ‘‘Align’’ operation in the feature space is followed by the ‘‘Blend’’ operation, as shown in Fig. 2(b). We use standard self-attention within each feature cluster to explore associations between homogeneous semantics within modalities. This process is defined as follows:

$$\hat{X}_n^R, \hat{X}_n^P, \hat{X}_n^V = \mathbb{SA}(X_n^R, Mask_n^R), \mathbb{SA}(X_n^P, Mask_n^P), \mathbb{SA}(X_n^V, Mask_n^V) \quad (8)$$

where $Mask_n^R, Mask_n^P, Mask_n^V$ indicate whether a feature point in the n -th semantic cluster is a complementary feature for each modality. The function $\mathbb{SA}$ refers to the standard self-attention mechanism. To further explore semantic relationships across modalities, we introduce a group-level cross-modal self-attention mechanism, described in Eqs 9-11.

$$\hat{X}_n^{PR} = \text{MCSA}(\hat{X}_n^R, \hat{X}_n^P, Mask_n^R, Mask_n^P) \quad (9)$$

$$\hat{X}_n^{PV} = \text{MCSA}(\hat{X}_n^P, \hat{X}_n^V, Mask_n^P, Mask_n^V) \quad (10)$$

$$Blended_{features} = \text{Cat}(\text{Merg}(\hat{X}_n^{PR}, Mask_n^P), \text{Merg}(\hat{X}_n^{PV}, Mask_n^P)) \quad (11)$$

where MCSA refers to multi-head cross-modal attention, *Merg* denotes the merging of feature points in the cluster, and *Cat* represents concatenation.

The fused feature set *Blended_{features}* is then passed to the subsequent network for further processing.

IV. EXPERIMENTS**A. Datasets**

We conducted experiments on three widely used large-scale public datasets: nuScenes, SemanticKITTI, and the indoor dataset S3DIS.

1) *nuScenes*: The nuScenes dataset contains 1,000 driving scenarios, with 850 scenarios used for training and validation (700 for training and 150 for validation) and 150 for testing. After merging similar classes and removing rare ones, 16 classes are used for LiDAR-based semantic segmentation.

2) *SemanticKITTI*: The SemanticKITTI dataset consists of 22 point cloud sequences. Sequences 00 to 10 are used for training, with sequence 08 for validation, while sequences 11 to 21 are used for testing. After merging moving classes and removing underrepresented ones, 19 classes are selected for training and evaluation.

3) *S3DIS*: The S3DIS dataset comprises 272 rooms across six large-scale areas. One area was designated as the validation set, while the remaining areas were utilized for training.

B. Evaluation Indicators

Following [3], [15], and [27], we adopt the inter-class intersection-over-union (IoU) for each class and mean IoU (mIoU) across all classes as evaluation metrics. The IoU for the i -th class is calculated as:

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i} \quad (12)$$

where TP_i , FP_i , and FN_i represent true positives, false positives, and false negatives, respectively.

C. Deployment Details

Our network consists of two primary components: the LCA module and the PV module. The LCA network adopts a semi-UNet structure, similar to SalsaNext [29] and SPVCNN [19], without any pre-trained extractor. Both the range and voxel branches undergo four downsampling stages, with feature dimensions at each stage of 32, 64, 128, and 256. For the range branch, the input size is 64×2048 for SemanticKITTI and 32×2048 for nuScenes, later adjusted to match SemanticKITTI settings. The voxel resolution remains constant at 0.05m throughout the experiment. For the point branch, we use two point-wise MLPs with dimensions 32 and 256. The PV structure employs the complete SPVNAS [19] architecture.

TABLE I

QUANTITATIVE RESULTS OF A2BLEND AND THE STATE-OF-THE-ART LiDAR SEMANTIC SEGMENTATION METHODS ON THE NUSCENES TEST SET. THE METHODS IN GRAY REPRESENT THE TOP THREE SOTA MODELS. ALL RESULTS CAN BE FOUND IN THE ONLINE LEADERBOARD

Methods	mIoU	barr	bicy	bus	car	const	motor	ped	cone	trail	truck	driv	other	walk	terr	made	veg
RPVNet [6]	77.6	78.2	43.4	92.7	93.2	49.0	85.7	80.5	66.0	66.9	84.0	96.9	73.5	75.9	76.0	90.6	88.9
AF2S3 [15]	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
GASN [34]	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
SpFormer [37]	81.9	83.3	39.2	94.7	92.5	77.5	84.2	84.4	79.1	88.4	78.3	97.9	69.0	81.5	77.2	93.4	90.2
Uniseg [12]	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
A2Blend(Our)	80.8	85.7	55.7	93.1	93.8	51.2	86.4	81.9	72.8	81.8	84.9	97.3	74.1	76.7	76.5	91.2	89.8

TABLE II

QUANTITATIVE RESULTS OF A2BLEND AND THE STATE-OF-THE-ART LiDAR SEMANTIC SEGMENTATION METHODS ON THE SEMANTICKITTI TEST SET. THE METHODS IN GRAY REPRESENT THE TOP THREE SOTA MODELS. ALL RESULTS CAN BE FOUND IN THE ONLINE LEADERBOARD

Methods	mIoU	car	bicy	moto	truc	o.veh	ped	b.list	m.list	road	par	walk	o.gro	build	fenc	veg	trun	terr	pole	sign
RPVNet [6]	70.3	97.6	68.4	68.7	44.2	61.6	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 [33]	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 [34]	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 [4]	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
SpFormer [37]	74.8	97.5	70.1	70.5	59.6	67.7	79.0	80.4	75.3	91.8	69.7	78.2	41.3	93.8	72.8	86.7	75.1	72.4	66.8	72.9
Uniseg [12]	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
A2Blend(Our)	71.9	97.7	57.6	68.9	44.4	62.8	76.1	75.1	74.7	93.7	71.1	80.9	34.3	93.7	72.4	87.0	76.2	72.1	66.4	61.0

D. Training and Reasoning Details

We trained the network in an end-to-end manner using the loss function defined in Eq. 7. The model was trained on the SemanticKITTI and nuScenes datasets using the SGD optimizer for 50 and 70 epochs, respectively, starting with a learning rate of 1.0 that decayed from 0.125 to 1.0, featuring one warm-up epoch followed by a cosine learning rate schedule. The training took about a week on the Nvidia GeForce RTX 3090. Momentum was set to 0.9, and weight decay to 0.0001. The hyperparameters β_{wce} , β_{ls} , and β_{cmtal} were set to 1, 1, 2.5 and 1, 1, 7.5, respectively. In each mini-batch, the hyperparameter m was set based on the semantic class with the most points. For data augmentation in the point branch, we applied random flipping, scaling, and transformations like LaserMix [31] and PolarMix [32] to increase the diversity of training samples.

E. Quantitative Comparisons

1) *Quantitative Results:* We summarize the performance of A2Blend and state-of-the-art LiDAR segmentation methods in Tables I and II. On the SemanticKITTI dataset, A2Blend outperforms most state-of-the-art algorithms in terms of mIoU, exceeding RPVNet [6] by 1.6 mIoU, SDSeg3D [33] by 1.5 mIoU, GASN [34] by 1.2 mIoU, and PVKD [4] by 0.7 mIoU. Among them, A2Blend excels in segmentation accuracy across 9 categories, including cars, motorcyclists, roads, parking lots, sidewalks, buildings, fences, tree trunks, and poles. Compared to the current state-of-the-art method, Uniseg [12], A2Blend shows improvements of 14.1 mIoU for motorcyclists, 1.1 mIoU for roads, 1.4 mIoU for sidewalks, and 0.3 mIoU for buildings. Even in challenging cases like cars, bicyclists, fences, vegetation, and tree trunks, the performance gap is minimal, only 0.2 to 0.3 mIoU. Similarly, on the nuScenes dataset, A2Blend outperforms RPVNet [6] by 3.2 mIoU, AF2S3 [15] by 2.5 mIoU, and GASN [34] by 0.4 mIoU, showing superior segmentation in 7 out of 16 classes. Compared to Uniseg [12], A2Blend delivers better

TABLE III

QUANTITATIVE RESULTS OF A2BLEND AND STATE-OF-THE-ART LiDAR SEMANTIC SEGMENTATION METHODS ON THE S3DIS AREA 5 DATASET. THE METHOD IN GRAY REPRESENTS THE TOP SOTA MODEL. ALL RESULTS CAN BE FOUND IN THE ONLINE LEADERBOARD

Method	mIoU
ST [35]	72.0
Swin3D [36]	72.5
PTv3 [3]	74.7
A2Blend(Our)	74.1

performance in segmenting buses, cars, pedestrians, trucks, other vehicles, and vegetation, with improvements of 1.0, 2.2, 1.0, 8.2, 2.3, and 1.0 mIoU, respectively. In the barrier, drivable, terrain, and manmade categories, the results are nearly identical. The segmentation gaps in other challenging cases may be due to Uniseg's use of an additional RGB modality, whereas A2Blend operates with only three modalities.

To further explore the scalability and robustness of model A2Blend, and considering the impact of occlusions caused by the dense distribution of indoor datasets, we opted not to use the entire framework directly. Instead, we fed the adjusted features from the A2Blend network into the state-of-the-art downstream network, Swin3D [36], to verify the generalization ability of our proposed method. Notably, only a single instance of dual contrast loss was utilized. As shown in Table III, our method achieves a 1.6 mIoU improvement over Swin3D [36], while trailing the state-of-the-art algorithm PTv3 [3] by just 0.6 mIoU.

F. Ablation Experiment

We conducted ablation experiments to examine the impact of different encoders, loss function combinations, and cross-view fusion methods.

1) *Effects of Each Modality:* To assess the effectiveness of each encoder, we isolated the LCA module.

TABLE IV
QUANTITATIVE EVALUATION OF EACH MODAL ENCODER
ON SEMANTICKITTI AND NUSCENES DATASETS

Range Encoder	Point Encoder	Voxel Encoder	mIoU	
			SemanticKITTI	nuScenes
✓	✓		43.0	50.9
✓	✓	✓	56.3	63.6
✓	✓	✓	58.0	66.5

TABLE V
QUANTITATIVE ASSESSMENT OF THE EFFECTIVENESS OF EACH MODULE
ON THE SEMANTICKITTI AND NUSCENES DATASETS

Baseline	Align(L_{cmtal})	Blend	mIoU	
			SemanticKITTI	nuScenes
✓			68.4	76.9
✓	✓		70.6	79.4
✓	✓	✓	71.9	80.8

TABLE VI
QUANTITATIVE EVALUATION OF L_{cmtal} ON THE SEMANTICKITTI
VALIDATION SET. WE COMPARE WITH THE BASELINE B1
USING L_{wce} AND L_{ls}

Method	L_{wce}	L_{ls}	L_{cmtal}	mIoU
B1	✓	✓		68.4
B2	✓	✓	✓	70.6

TABLE VII
QUANTITATIVE EVALUATION OF BLEND ON THE
SEMANTICKITTI VALIDATION SET

Method	mIoU	Δ
Add(Baseline)	70.5	+0.0
Concat	70.7	+0.2
Blend	71.9	+ 1.4

From Table IV, we observe that the voxel encoder contributes the most to the LCA module, while the range encoder contributes the least. Compared to the range encoder, the voxel encoder improves performance by 13.3 mIoU on SemanticKITTI and 12.7 mIoU on nuScenes. The best results were achieved by combining the point, range, and voxel encoders, resulting in 58.0 mIoU and 66.5 mIoU, respectively.

2) *Effects of Each Module*: We evaluated the effectiveness of the Align and Blend modules on SemanticKITTI and nuScenes datasets. As shown in Table V, the Align module (L_{cmtal}) improves mIoU by 2.2 and 2.5 on SemanticKITTI and nuScenes, respectively, while the Blend module boosts mIoU by 1.3 and 1.4, respectively. This highlights the Align operation's critical role in facilitating cross-modal fusion.

3) *Effect of Each Loss Function*: We evaluate the impact of our proposed L_{cmtal} loss in Table VI. Compared to the baseline using L_{wce} and L_{ls} , our loss function improves mIoU by 2.2.

4) *Mechanisms for Fusion*: Table VII compares different fusion strategies. Our Blend method outperforms the baseline by 1.4 mIoU, while the Concat method provides only a slight improvement.

5) *Efficiency and Accuracy Comparison*: Table VIII compares parameters, latency, and mIoU of A2Blend with

TABLE VIII
QUANTITATIVE EVALUATION OF PARAMETERS AND LATENCY ON THE
SEMANTICKITTI TEST SET. THE METHODS IN GRAY REPRESENT THE
TOP THREE SOTA MODELS

Method	Param(M)	Latency(ms)	mIoU
Cylinder3D [5]	56.3	75.1	65.9
SPVCNN [19]	21.8	150	63.8
RPVNet [6]	24.8	168	70.3
PTv3 [3]	46.2	-	75.5
Uniseg [12]	147.6	-	75.2
A2Blend(Our)	24.3	132.2	71.9

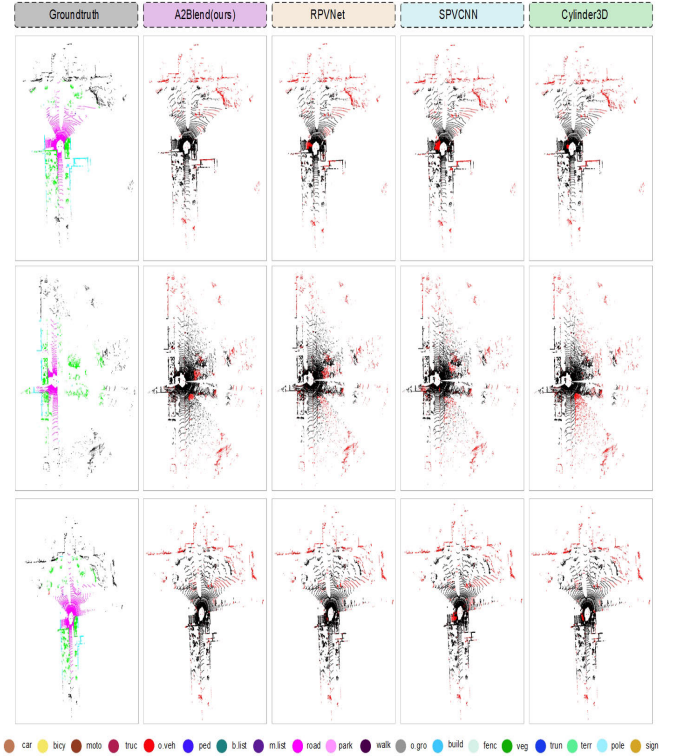


Fig. 3. Quantitative evaluation of A2Blend and state of the art algorithms on the SemanticKITTI validation set. Segmentation errors are highlighted in red. Best viewed in colors.

state-of-the-art algorithms. Our network outperforms Cylinder3D [5], SPVCNN [19], and RPVNet [6], with a competitive balance of latency and accuracy, despite the overhead from clustering and padding.

6) *Visualization and Analysis*: Fig. 3 shows a comparison of A2Blend and state-of-the-art algorithms on the SemanticKITTI validation set. A2Blend exhibits superior overall performance. To assess boundary segmentation accuracy, we visualized A2Blend and competing methods on categories like fence, road, sidewalk, and parking. As seen in Fig. 4, A2Blend significantly improves boundary segmentation accuracy over other methods. Furthermore, Fig. 5 presents a visual comparison after applying our proposed dual contrastive loss on top of method Swin3D. It can be seen that our approach outperforms the baseline in both missegmentation and boundary segmentation areas.

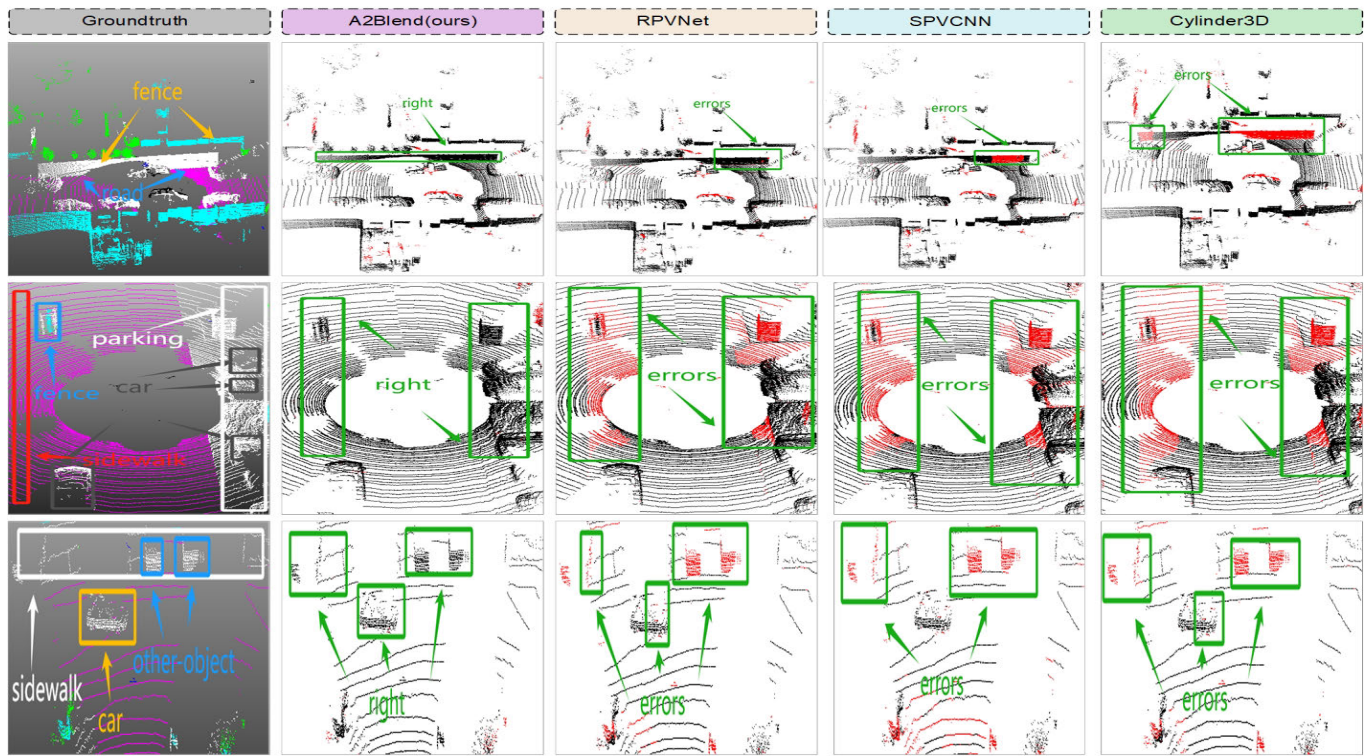


Fig. 4. Quantitative evaluation of A2Blend and state-of-the-art algorithms on boundary segmentation on the SemanticKITTI validation set. Segmentation errors are highlighted in red. Best viewed in colors.

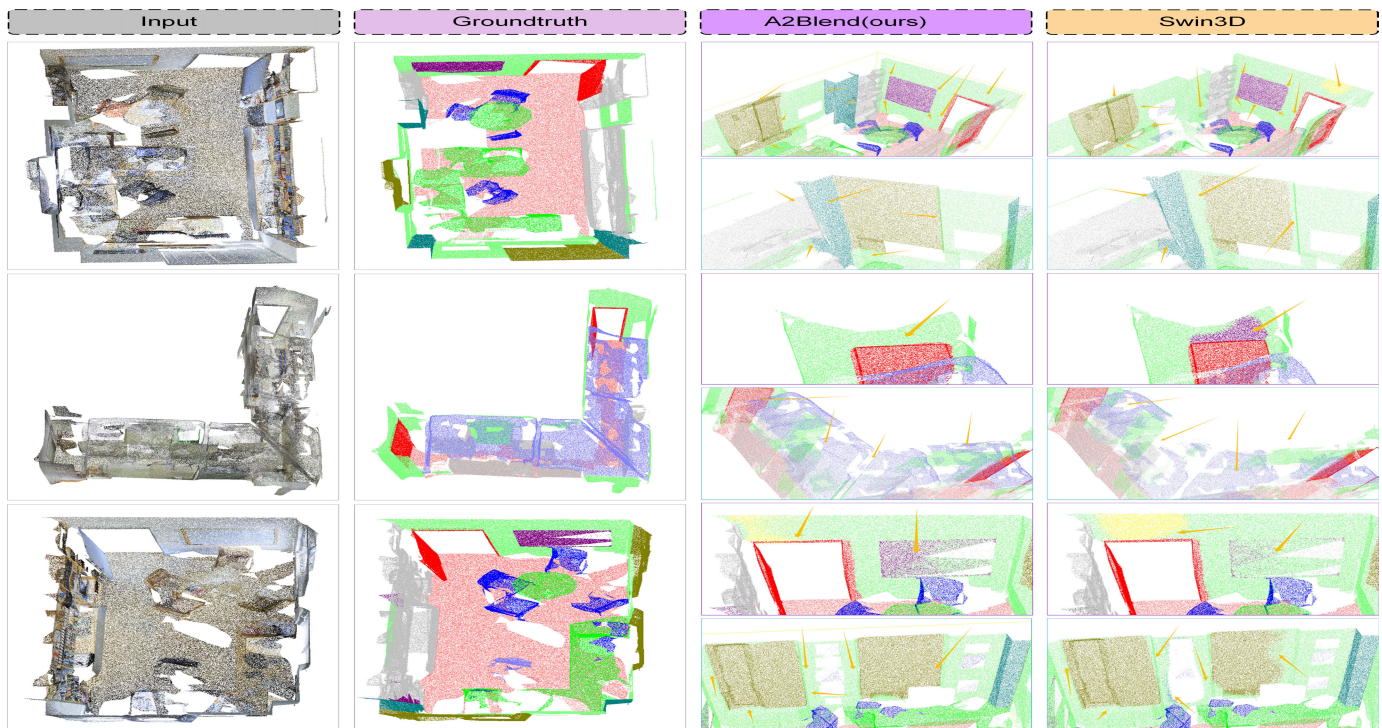


Fig. 5. Qualitative comparison of A2Blend and algorithm Swin3D for boundary segmentation on the S3DIS AREA 5 DATASET. Light yellow arrows highlight the segmentation differences. Best viewed in colors.

V. CONCLUSION

In this paper, we present A2Blend, a novel unified multi-modal LiDAR segmentation network that achieves outstanding

performance. A2Blend demonstrates exceptional proficiency in fine-grained segmentation tasks by synergistically integrating three distinct point cloud representations: point view, voxel

view, and range view. The introduction of key innovations in the align and blend operations significantly enhances the model's ability to represent and distinguish features. Specifically, the Learnable Cross-Modal Association (LCA) module in the align task enables the aggregation of semantically similar clusters across and within modalities while separating dissimilar clusters, aided by the "Cross-Modal Triplet Alignment Loss" mechanism. This substantially improves model performance. In the blend task, we leverage standard self-attention and cross self-attention mechanisms to capture relationships both within and across modalities, boosting the model's generalization ability. Experimental results on the SemanticKITTI, nuScenes and S3DIS datasets demonstrate a significant improvement in mIoU, positioning A2Blend as a pioneering approach in applying contrastive learning to 3D LiDAR multimodal fusion. In the future, we will focus on optimizing Algorithm 1 to further enhance the model's performance.

ACKNOWLEDGMENT

The authors extend their sincere gratitude for this support, which was instrumental in advancing their work.

REFERENCES

- [1] Y. Zhang, Q. Zhang, Z. Zhu, J. Hou, and Y. Yuan, "GLENNet: Boosting 3D object detectors with generative label uncertainty estimation," *Int. J. Comput. Vis.*, vol. 131, no. 12, pp. 3332–3352, Dec. 2023.
- [2] Q. Xu, Y. Zhou, W. Wang, C. R. Qi, and D. Anguelov, "SPG: Unsupervised domain adaptation for 3D object detection via semantic point generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15426–15436.
- [3] X. Wu et al., "Point transformer v3: Simpler, faster, stronger," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 4840–4851.
- [4] Y. Hou, X. Zhu, Y. Ma, C. C. Loy, and Y. Li, "Point-to-Voxel knowledge distillation for LiDAR semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8469–8478.
- [5] X. Zhu et al., "Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9934–9943.
- [6] J. Xu, R. Zhang, J. Dou, Y. Zhu, J. Sun, and S. Pu, "RPVNet: A deep and efficient range-point-voxel fusion network for LiDAR point cloud segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16004–16013.
- [7] Q. Xu, Y. Zhong, and U. Neumann, "Behind the curtain: Learning occluded shapes for 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 1–9.
- [8] W. Zheng, W. Tang, L. Jiang, and C.-W. Fu, "SE-SSD: Self-ensembling single-stage object detector from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14489–14498.
- [9] Y. Kim, K. Park, M. Kim, D. Kum, and J. Won Choi, "3D dual-fusion: Dual-domain dual-query camera-LiDAR fusion for 3D object detection," 2022, *arXiv:2211.13529*.
- [10] J. Behley et al., "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9296–9306.
- [11] H. Caesar et al., "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11618–11628.
- [12] Y. Liu et al., "UniSeg: A unified multi-modal LiDAR segmentation network and the OpenPCSeg codebase," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21605–21616.
- [13] L. Kong et al., "Rethinking range view representation for LiDAR segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 228–240.
- [14] L. Kong et al., "Robo3D: Towards robust and reliable 3D perception against corruptions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19937–19949.
- [15] R. Cheng, R. Razani, E. Taghavi, E. Li, and B. Liu, "(AF)²-S3Net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12542–12551.
- [16] Y. Rao et al., "DenseCLIP: Language-guided dense prediction with context-aware prompting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18061–18070.
- [17] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 696–712.
- [18] R. Zhang et al., "PointCLIP: Point cloud understanding by CLIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8542–8552.
- [19] H. Tang et al., "Searching efficient 3D architectures with sparse point-voxel convolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 685–702.
- [20] Y. Dongqiangzi et al., "LiDARMultiNet: Towards a unified multi-task network for LiDAR perception," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 3, pp. 1–10.
- [21] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany, "Point-Contrast: Unsupervised pre-training for 3D point cloud understanding," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 574–591.
- [22] S. Huang, Y. Xie, S.-C. Zhu, and Y. Zhu, "Spatio-temporal self-supervised representation learning for 3D point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6515–6525.
- [23] R. Chen et al., "Zero-shot point cloud segmentation by transferring geometric primitives," 2022, *arXiv:2210.09923*.
- [24] A. Cheraghian, S. Rahman, T. F. Chowdhury, D. Campbell, and L. Petersson, "Zero-shot learning on 3D point cloud objects and beyond," *Int. J. Comput. Vis.*, vol. 130, no. 10, pp. 2364–2384, Oct. 2022.
- [25] C. Sautier, G. Puy, S. Gidaris, A. Boulch, A. Bursuc, and R. Marlet, "Image-to-LiDAR self-supervised distillation for autonomous driving data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9881–9891.
- [26] Y.-C. Liu et al., "Learning from 2D: Contrastive pixel-to-point knowledge transfer for 3D pretraining," 2021, *arXiv:2104.04687*.
- [27] J. Li, C. Wang, C. Wang, M. Zhao, and Z. Jiang, "RPV-CASNet: Range-point-voxel integration with channel self-attention network for LiDAR point cloud segmentation," *Int. J. Speech Technol.*, vol. 54, nos. 17–18, pp. 7829–7848, Sep. 2024.
- [28] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4413–4421.
- [29] T. Cortinhal, G. Tzelepis, and E. E. Aksoy, "Salsanext: Fast, uncertainty-aware semantic segmentation of LiDAR point clouds," in *Proc. 15th Int. Symp. Vis. Comput.*, San Diego, CA, USA. Cham, Switzerland: Springer, Oct. 2020, pp. 207–222.
- [30] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9620–9629.
- [31] L. Kong, J. Ren, L. Pan, and Z. Liu, "LaserMix for semi-supervised LiDAR semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21706–21716.
- [32] A. Xiao, J. Huang, D. Guan, K. Cui, S. Lu, and L. Shao, "Polarmix: A general data augmentation technique for LiDAR point clouds," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 11035–11048.
- [33] J. Li, H. Dai, and Y. Ding, "Self-distillation for robust LiDAR semantic segmentation in autonomous driving," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 659–676.
- [34] M. Ye, R. Wan, S. Xu, T. Cao, and Q. Chen, "Efficient point cloud segmentation with geometry-aware sparse networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 196–212.
- [35] X. Lai et al., "Stratified transformer for 3D point cloud segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8490–8499.
- [36] Y.-Q. Yang et al., "Swin3D: A pretrained transformer backbone for 3D indoor scene understanding," 2023, *arXiv:2304.06906*.
- [37] X. Lai, Y. Chen, F. Lu, J. Liu, and J. Jia, "Spherical transformer for LiDAR-based 3D recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17545–17555.



Chuanxu Wang received the Ph.D. degree from the Ocean University of China. He is currently a Professor and a Master's Tutor with the School of Qingdao University of Science and Technology, mainly engaged in research on computer vision and pattern recognition.



Da Song received the bachelor's degree from the College of Information Science and Technology, Qingdao University of Science and Technology, in 2023, where he is currently pursuing the master's degree, where his current research interests include 3D point cloud and computer vision.



Jiajiong Li received the bachelor's degree in computer science and technology from Shangqiu Institute of Technology in 2021. He is currently pursuing the master's degree with the School of Information Science and Technology, Qingdao University of Science and Technology. His research interests focus on the segmentation of outdoor LiDAR data.



Xin Chen received the B.Eng. degree in data science and big data technology from Hunan University of Technology and Business, China, in 2022. Currently, he is pursuing the master's degree with the Department of Computer Technology, Qingdao University of Science and Technology, China.



Binghui Wang received the bachelor's degree from the College of Information Science and Technology, Qingdao University of Science and Technology, in 2023, where she is currently pursuing the master's degree. She conducting research in computer vision.