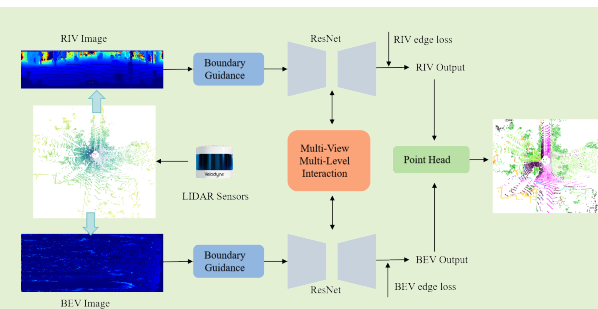


Accurate 3-D Semantic Segmentation of Point Clouds for Intelligent Vehicles Based on Multiview Edge Guidance and Fusion

Yan Liu¹, Lei Xu¹, Weiming Hu¹, Xiong Chen, Bo Yi², Qiu Mao¹, Dong Kong³, and Shengping Ruan

Abstract—Semantic segmentation based on LiDAR plays an important role in the environment perception, path planning, and decision control of unmanned ground systems. However, in challenging environments with complex background interleaving, the high proportion of drivable areas and significant target elements pose serious challenges to segmentation tasks. Existing studies attempt to address these challenges but still grapple with issues such as interclass semantic conflicts and inaccurate edge segmentation. To address these challenges effectively, a reliable point cloud segmentation network driven by multiview fusion and multiedge guidance is specially customized. Specifically, the network's point cloud multiview range image view (RIV) and bird's eye view (BEV) encoding backbones use lightweight pyramid architectures to extract specific details and semantic features of multiscale point clouds. Subsequently, a dedicated multiview edge guidance module focuses on enhancing interclass edge differentiation features from RIV and BEV, respectively. These features are fused with the encoding backbone hierarchy, and the extraction of edge information is supervised by a new hybrid edge loss to maximize the consistency of semantic segmentation and predicted edges. Furthermore, a multiview and multiscale feature fusion module based on a multihead attention mechanism is introduced to enhance the potential complementarity and interaction of the captured features. Through comprehensive experiments and ablation studies on the SemanticKITTI and RELLIS-3D datasets, recognized in the field of autonomous driving, the results demonstrate that our custom method is competitive in terms of intersection over union (IoU) accuracy and real-time indicators compared with typical baseline methods.

Index Terms—3-D semantic segmentation, complex environment understanding, multiedge guidance, multiview fusion.



I. INTRODUCTION

THE sophistication of advanced perception systems is crucial for maintaining robustness and achieving autonomy in ground unmanned systems, especially in complex and

dynamic road conditions. In the key aspect of supporting the all-dimensional perception of unmanned systems, accurate and reliable environmental understanding is of paramount importance [1]. As a central component in environmental perception, 3-D semantic segmentation plays a critical role in interpreting high-precision, all-dimensional perception data [2], [3]. It functions similar to the human eye, comprehensively interpreting essential information about roads, obstacles, pedestrians, and other key elements, which in turn fundamentally improves the path planning and decision-making control capabilities of ground unmanned systems [4], [5], [6].

Typically, comprehensive environmental understanding involves research on semantic segmentation based on visual and LiDAR sensors. Due to the susceptibility of visual images to weather and lighting changes, 3-D semantic segmentation based on LiDAR point clouds has gradually gained widespread attention in recent years [7]. In comparison to 2-D images, 3-D point cloud data not only provide accurate environmental geometric distance information but also distinctly scan and present

Manuscript received 24 May 2024; revised 18 June 2024; accepted 19 June 2024. Date of publication 27 June 2024; date of current version 15 August 2024. The associate editor coordinating the review of this article and approving it for publication was Dr. Guofa Li. (Corresponding author: Lei Xu.)

Yan Liu is with China Automotive Engineering Research Institute Company Ltd., Chongqing 401122, China, and also with the School of Electronics, Jilin University, Changchun 130012, China.

Lei Xu, Weiming Hu, Xiong Chen, and Qiu Mao are with China Automotive Engineering Research Institute Company Ltd., Chongqing 401122, China (e-mail: 00067924@caeri.com.cn).

Bo Yi is with Chinese Academy of Agricultural Mechanization Science Group Company Ltd., Beijing 100083, China.

Dong Kong is with the College of Transportation, Shandong University of Science and Technology, Qingdao 266590, China.

Shengping Ruan is with the School of Electronics, Jilin University, Changchun 130012, China.

Digital Object Identifier 10.1109/JSEN.2024.3417522

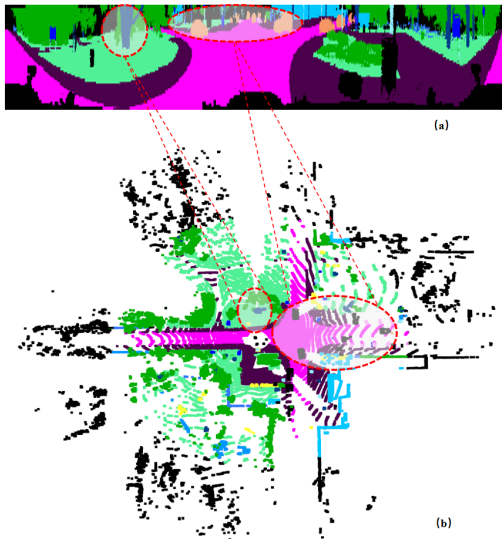


Fig. 1. BEV and RIV images generated by the point cloud projection produce different environmental structures and distributions. (a) Range image view of semantic point cloud projection. (b) Top view of semantic point cloud.

the topological relationships between elements in the scene. In addition, point cloud data exhibit stronger adaptability to real-world environmental changes and a wider perceptual field of view compared with visual images.

Currently, researchers have undertaken various efforts to improve the efficacy of LiDAR-based semantic segmentation in environmental applications. Depending on the format of the point cloud fed into the model, the existing studies can be broadly categorized into three groups: point-based methods [8], [9], [10], voxel-based methods [11], [12], [13], and projection-based methods [14], [15], [16], [17]. Point-based approaches have exhibited promising results; however, they face challenges in efficiently capturing relationships within a point neighborhood, making them susceptible to point cloud sparsity and resulting in blurred boundaries, ultimately leading to reduced accuracy and increased computational demands. Voxel-based methods involve transforming 3-D point clouds into regular 3-D voxels or cells using various downsampling techniques and encoding them in conjunction with 3-D sparse convolutional neural networks (CNNs). Nevertheless, voxelization operations can exacerbate point cloud sparsity and compromise crucial information. In addition, the encoding efficiency of 3-D CNNs tends to decrease as data volume increases.

The success of 2-D CNNs in visual image segmentation is remarkable, attributed to their outstanding segmentation performance and real-time capabilities within a lightweight structure. The 3-D point cloud-based segmentation methods draw inspiration from these achievements, using top-down or spherical projection techniques. These techniques project the original 3-D point cloud to create bird's-eye view (BEV) or range images, opening avenues for innovative point cloud processing approaches. As illustrated in Fig. 1, BEV and range images generated through different projection methods yield distinct original environment structures and distributions within the original point cloud. For example, range images

from point cloud projection effectively capture the shapes of objects such as vehicles and pedestrians in the Z-direction. However, this projection may map multiple points to the same pixel location, inevitably leading to the loss of information about farther scene elements. On the other hand, BEV images, generated using downsampling techniques, efficiently capture distribution and location information of scene elements, particularly in dense multielement regions. This perspective provides a clear delineation of the boundary contours of individual elements in the X-Y plane. Despite these advantages, downsampling methods used in generating BEV images may result in the loss of detailed information present in the original point cloud.

To compensate for the above research gaps, recent advancements in point cloud semantic segmentation methods have centered around multiprojection fusion [18], [19], [7], [20], emerging as a prominent and popular direction. However, despite acknowledging the benefits of multiview input representation fusion, most studies have primarily focused on superficially designed input-level or output-level fusion methods at the methodology level. They have not adequately emphasized the interplay between features from various projection viewpoints, consequently missing a thorough exploration of the intrinsic connections among diverse input representations. The most obvious difference with the above multiview methods is that we creatively embedded a multiview boundary attention guidance module to further focus on the interclass boundary information, enhance the interclass boundary distinguishing features, and improve the segmentation accuracy.

To effectively address the aforementioned research gap, we customize the multiview fusion and boundary guidance driven network for the point cloud semantic segmentation task. First, the proposed method maximizes the utilization of edge cues, incorporating an edge loss to effectively alleviate the boundary ambiguity issue in interclass segmentation elements. Second, a specifically designed multiview multilevel feature fusion module is capable of extracting complementary features and global contextual information from both RIV and BEV perspectives. Finally, complemented by the designed hybrid boundary loss, our method achieves end-to-end efficient and accurate semantic segmentation based on 3-D point clouds.

In summary, the contributions of our work are itemized below.

To achieve efficient and accurate semantic segmentation of 3-D point clouds, especially among classes, a multiview fusion and boundary guidance driven segmentation network is proposed. It fully taps and uses the advantages of multiview complementarity and enhances the importance and context information of interclass boundary information by designing multiview boundary guidance module (BGM), so that our network can pay attention to and weigh the point cloud data of different distances and heights. To our knowledge, this is the first-time multiview boundary guidelines and cross-view encoding structures have been used to fuse multiview point cloud data for environmental understanding tasks.

We devised a multiview BGM, supplemented by boundary loss, with the aim of enhancing the contribution of boundary cues to the segmentation of boundary masks for different elemental classes. In addition, we designed a cross-view

multiscale feature interaction module that uses a multihead attention mechanism to explore the potential interaction and contextual information of features from RIV and BEV viewpoints. It learns the differences in features from each channel, enhancing the feature representation capability of sparse and distant point clouds, as well as the discriminative ability between interclass points.

Comprehensive comparative and ablation experiments are conducted on widely recognized autonomous driving urban datasets such as SemanticKITTI and the off-road dataset RELLIS-3D, involving both structured and unstructured scenario data. The experimental results demonstrate that our customized approach exhibits competitiveness in terms of IoU, real-time performance, and other metrics when compared with other representative baseline methods based on projection, voxelization, and similar techniques.

II. RELATED WORKS

Typically, based on the form of point cloud representations, the existing LiDAR-based semantic segmentation efforts can be broadly categorized into the following three groups: point-based methods, voxel-based methods, and projection-based methods.

A. Point-Based Methods

At present, point-based methods mainly rely on the spatial structure and topological relationship of the point cloud itself to encode features. Methods that act directly on the original point cloud can be categorized into multilayer perception (MLP)-based methods and CNN-based methods. The pioneer PointNet [21] uses a multilayer perceptron with a maximum pooling operation to merely encode the representation of each point. While PointNet efficiently extracts features from individual points, it lacks the capability to capture local structures, resulting in limited proficiency in detailed processing and generalization to complex scenes. PointNet++ [8] serves as a follow-up to PointNet, addressing some of its limitations to a certain extent. However, its uneven sampling design is prone to challenges in scenes with varying point cloud density, leading to blurred boundary segmentation between classes. In addition, kernel-point convolution (KPConv) [9] introduces a novel convolution technique that directly leverages neighboring points as inputs. This method enhances the extraction of local features using a small set of kernel points with spatially adaptive weights. RandLA-Net [22] uses random point sampling combined with a local feature aggregation design to incrementally increase the sensory field of each feature to efficiently preserve more geometric details. Generally, the aforementioned point-based methods face challenges such as capturing fine details, achieving accurate boundary segmentation, and exhibiting slow inference speeds. Specifically, their performance tends to be suboptimal when confronted with complex scenes characterized by sparse point clouds and diverse element classes.

B. Voxel-Based Methods

The voxel-based methods typically involve the transformation of raw point clouds into regular grids or cells, using 3-D

convolutions for feature extraction within each voxel. Representative works, exemplified by [11], [12], and [22], treat point clouds as uniformly distributed, partitioning them into regular cubes to obtain segmentation results. Nevertheless, these methods encounter challenges due to the inherent sparsity, disorder, and nonuniform distribution of point clouds, making them less effective in addressing variations in outdoor point cloud densities. Cylinder3D [12] introduces cylinder segmentation and asymmetric 3-D convolutions to investigate 3-D structures while preserving the intrinsic attributes of the original point cloud, including sparsity and density variations. Furthermore, a pointwise refinement module is devised to mitigate interference from voxel label encoding. SFD [23] uses only the original point cloud and is used for RPN to generate 3-D RoI, and then uses sparse convolution to extract voxel features to achieve good detection results. UVTR [24] displays the image and radar point cloud features in the voxel space and uses knowledge transfer for feature interaction, achieving excellent performance in both single mode and multimode. BEVFusion realizes multitask detection by representing the radar point cloud and the image as BEV spatial features.

In conclusion, voxel-based methods maintain relatively comprehensive geometric and topological information in 3-D point clouds. Nevertheless, the performance of these methods is contingent on the density of the partitioned 3-D grids, influenced by the inherent sparsity of point clouds. Consequently, given the extensive operations involved in 3-D convolutions, striking a balance between real-time processing and accuracy remains challenging.

C. Projection-Based Methods

The projection-based methods can project the 3-D point cloud into a 2-D image space, such as RV or BEV, and then extract features using an effective 2-D convolutional neural network, which enables better real-time performance. SqueezeSeg [14] introduced a spherical projection technique to transform 3-D point cloud data into a 2-D range image, and then used a combination of lightweight SqueezeNet and conditional random field (CRF) [25] to perform the segmentation task. Furthermore, SqueezeSegV2 [26] designed a context aggregation module aiming to aggregate contextual information from broader perceptual field of view. SqueezeSegV3 [15] introduced a spatially adaptive computation (SAC) structure to dynamically adjust the weights of inputs and locations, aiming to efficiently deal with the case of uneven spatial distribution of features in LiDAR images and to improve the computational efficiency. RangeNet++ [27] integrated Darknet in SqueezeSeg [14] and proposed an efficient K -nearest neighbors (KNNs) postprocessing method for point label prediction. Extending the encoder-decoder architecture from SalsaNet [16], SalsaNext [28] incorporates a context module and integrates Bayesian methods to compute both epistemic and aleatoric uncertainties for each point cloud.

Meta-RangeSeg [29] presents a novel representation known as the distance residual image, aimed at capturing spatiotemporal information. This representation helps mitigate the inconsistency between the 2-D distance image coordinate input and the 3-D Cartesian coordinate output. Lite-HDSeg [30]

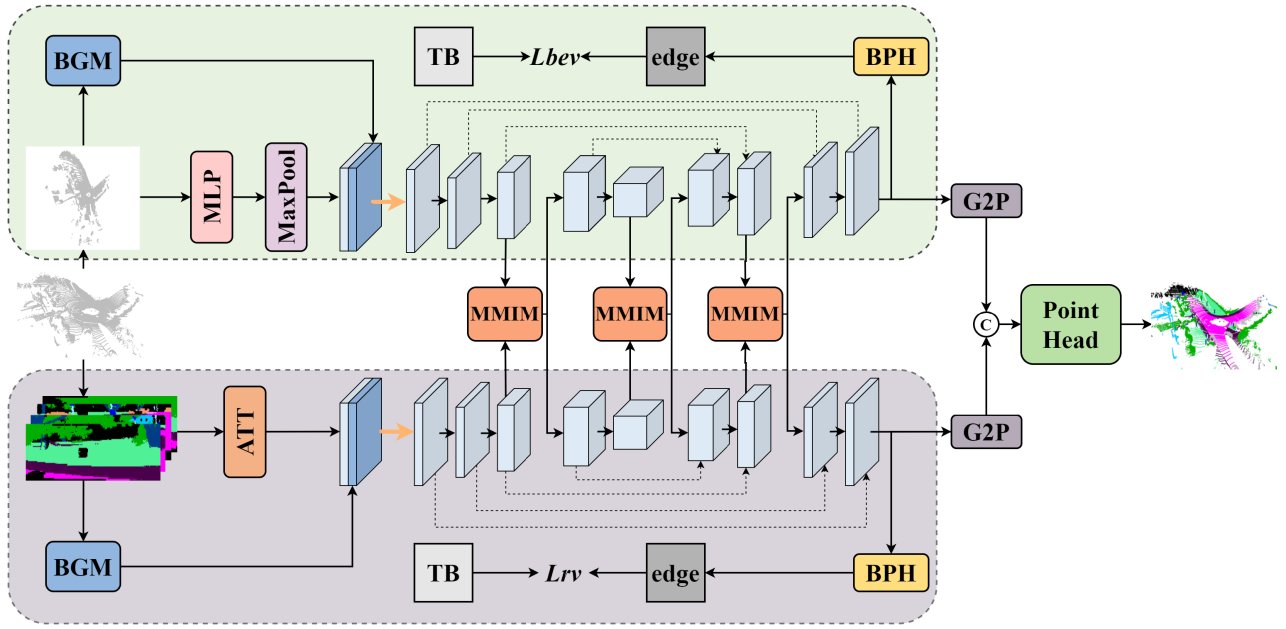


Fig. 2. Overall structure of the proposed network. One of the focuses is the multiview BGM, which guides the network to focus on interclass features. Another focus is the multiview feature-level fusion module, which improves the richness of features.

introduces three different modules, Inception-like context module, multiclass spatial propagation network, and boundary loss, through integration that claimed to achieve state-of-the-art performance. Addressing the unbalanced spatial distribution of the point cloud, PolarNet [8] allocates points to grid cells using quantized azimuths and radius, which in turn designs the circular convolution to implement grid point cloud encoding.

At present, there have been researches on semantic segmentation using multiview fusion or multimodes [31], [32], [33]. For instance, AMVNet [34] introduces nodding heads to reconcile discrepancies between the range view (RV) and BEV branches in predictions, thereby enhancing overall segmentation accuracy. In addition, certain studies [7], [20] independently process BEV and RV data during encoding and decoding, with postfusion occurring solely at the output stage. It is crucial to note that regardless of whether the point cloud projection operation generates RV or BEV, these perspectives inevitably lead to the loss of vital information from the original point cloud. Specifically, the RV generation process may map multiple original 3-D points to the same pixel point in the image, resulting in distortion of element contours or boundaries. Conversely, the BEV generation process compresses information about the height direction, leading to insensitivity to perceived Z-axis direction features. Therefore, an important consideration is how to effectively leverage the complementary advantages of RV and BEV viewpoints to thoroughly explore the correlation of potential features from multiple perspectives and attain the realization of a unified value across multiple viewpoints, representing a noteworthy direction for further exploration and research efforts.

III. METHODOLOGY

Considering the shortcomings that the point cloud range image view (RIV) may lose the information of the distant

target due to the occlusion of scene elements caused by projection, and the point cloud BEV image may lose the vertical information of the target, for this reason, this article purposefully designs a multiview fusion and multiboundary-guided semantic segmentation network of the point cloud, whose structure is shown in Fig. 2, to make up for the above information loss of the different projected views and enhance the complementarity. The proposed method consists of two main branches: the BEV encoding branch and the RIV encoding branch, and both are supplemented with a BGM aiming at highlighting interclass boundary information. In addition, a multiview multilevel interaction module (MMIM) is introduced to facilitate feature-level deep fusion between the two branches. Finally, a new hybrid loss function is custom-designed to direct the model's attention to interclass boundary segmentation.

A. Multiview Boundary Guidance Module

Interclass boundary information plays a crucial role in semantic segmentation tasks as it provides more accurate target localization and reduces confusion with neighboring classes. Excellent interclass boundary guidance can improve the segmentation results, especially when segmenting classes with comprehensive coverage, such as bushes and vegetation. Inspired by [35], to emphasize the importance of boundary information and reduce the interference of redundant information, we introduce a multiview BGM to demonstrably enhance the efficacy of boundary features in BEV and RIV branches. Specifically, the BGM segments the BEV and RIV into subgraphs based on the input channels and assigns different weights to each subgraph. Among them, the weights are assigned by calculating the similarity between the subgraphs and the guidance boundary, thus emphasizing the

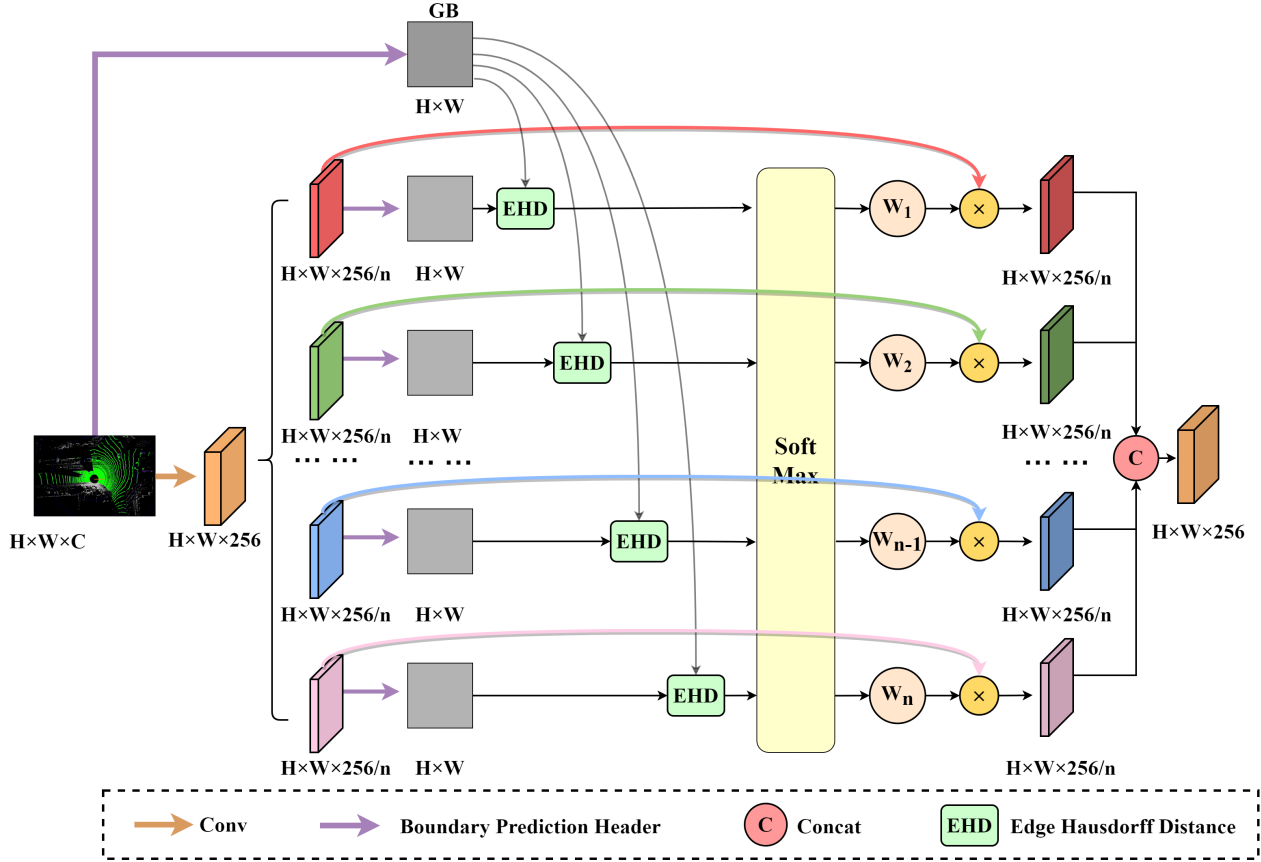


Fig. 3. Overall structure of the multiview BGM.

boundary effect. The overall structure of the designed BGM is shown in Fig. 3.

Specifically, first, the number of channels of the point cloud projection image is extended by convolution operation. Based on the number of channels, the expanded image is segmented into N subimages with the same shape. Then the boundary prediction header is used to generate the boundary between the original image and the N subimages. Considering that boundary images are binary, capable of distinguishing only between boundaries and nonboundaries, the Hausdorff distance is used. This distance metric is computed using a specific formula to assess the maximum dissimilarity between the ground-truth boundary and the predicted boundary, effectively quantifying their similarity. Subsequently, the similarity value is then determined using softmax, which assigns weights to each subgraph. Higher weights indicate that the subgraph boundaries are more similar to the actual boundaries. These weights are then multiplied with the corresponding subgraphs to obtain weighted subgraphs, which are concatenated and fed into the branch encoding backbone. This operation can effectively give greater weight to the subgraphs with high similarity, thus achieving a more prominent boundary while retaining the original image information

$$d(TB, FB) = \min_{tb \in TB} \|fb - tb\| \quad (1)$$

$$d(FB, TB) = \min_{fb \in FB} \|tb - fb\| \quad (2)$$

$$\text{Similarity} = \frac{1}{\max(d(TB, FB), d(FB, TB))} \quad (3)$$

where TB and FB represent the ground-truth boundary and the subgraph boundary, respectively, tb and fb represent the edge pixels in TB and FB , respectively. $\max(d(TB, FB), d(FB, TB))$ represents the maximum difference between TB and FB , and thus we use its inverse as the similarity.

B. BEV Encoding Branch

Given the characteristics of LiDAR sensor, where the point cloud exhibits higher density in the near range and sparsity in the far range, achieving a well-balanced treatment of such inhomogeneous point cloud distributions through traditional Cartesian partitioning may prove challenging. To address this issue, we chose a polar coordinate segmentation method that naturally captures the distribution of point clouds at different distances and angles to obtain a more balanced representation. Following the generation of the BEV image, it is input into the BGM to accentuate boundary information. Concurrently, we use an MLP with maximum pooling to capture feature information of each BEV mesh, thus eliminating the need for manual feature design. Ultimately, both the outcomes are concatenated at the channel level, and the channel dimensions are restored using a convolution operation.

Following the above steps, to achieve multiscale target feature extraction and segmentation while taking into account the segmentation accuracy and computational cost, we used ResNet with 34 layers as the encoder for the BEV trunk branch. In addition, the decoder consists of upsampling using

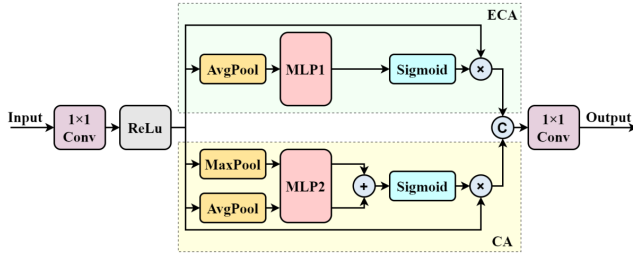


Fig. 4. CA and ECA collaborative combination structure. It allows the network to focus more on the features that are most useful for the task, helps reduce redundant information, and improves the representation of the model.

pixel shuffle [36] and merging with the output characteristics of the previous transformer block via skip-connection.

C. RIV Encoding Branch

For the input of the RIV encoding branch, it is initially obtained by spherical projection of the point cloud using the following projection equations:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2} [1 - \tan^{-1}(y, x) \pi^{-1}] W \\ [1 - (\sin^{-1}(z, r^{-1}) + f_{up}) f^{-1}] H \end{pmatrix}. \quad (4)$$

First, the projected RIV image is fed into the BGM to obtain the boundary-enhanced RIV image and the ground-truth boundary. Meanwhile, the encoder-decoder architecture is consistent with the BEV branch, which ensures the balance of computation time and segmentation accuracy and facilitates the subsequent multiview feature-level deep fusion. However, RIV images are insensitive to long-range targets due to the occlusion effect of close-range targets. To address this limitation, we introduce channel attention (CA) to obtain more accurate position information over the entire range, thus improving the model's attention to targets at different locations. In addition, considering that different channels in the distance image contain different information, efficient CA (ECA) is also introduced to enhance cross-channel information exchange between feature maps. The synergistic combination structure of CA and ECA is shown in Fig. 4, which helps reduce redundant information and improve the representation of the model by introducing attentional weights between the channels so that the network is more focused on the features that are most useful for the task.

D. Multiview Multilevel Interaction Module

Since the point cloud will result in the occlusion of distant targets and thus information loss when projected to RIV, and the height information of the object will be lost when projected to BEV, this article designs the MMIM to compensate for the loss of information in multiple projections and make the advantages of the different viewpoints complementary to each other. As shown in Fig. 5, in the MMIM, a multihead attention (MHA) mechanism is introduced to achieve effective feature-level depth fusion between the RIV encoding branch and the BEV encoding branch. Specifically, we use the feature map of BEV as queries and the feature maps of RIV as keys and values; conversely, we use the feature map of RIV as

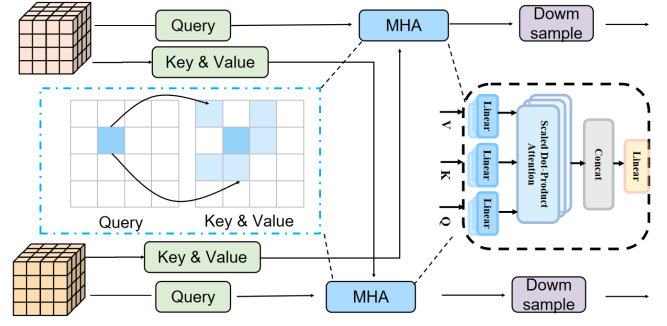


Fig. 5. Framework of multiview MMIM. It not only facilitates bidirectional information transfer between BEV and RIV feature maps at the feature level but also dynamically adjusts the weights of the features through the attention mechanism.

queries and the feature maps of BEV as keys and values. The aforementioned operations result in the generation of fused BEV and fused RIV through multihead attention. Aimed at striking a balance between effectiveness and complexity, we embed three attention heads, mathematically described as in 5. The inclusion of multiple attention heads enables the model to focus on information at different scales and better capture long-range dependencies in the multiview features, compensating for the information scale differences between BEV and RIV images. This, in turn, enhances the model's adaptability to objects of different scales in various scenes.

In addition, the parallel structure of multiple attention heads helps capture complementary and contextual information of local and global features from both RIV and BEV perspectives. This inclusive feature mining and integration helps improve the overall feature richness and diversity. After completing these steps, the fused feature map is downsampled to match the input size of the next layer in the model

$$\begin{aligned} \text{head}_i &= \text{Attention}(Q_i, K_i, V_i) \\ &= \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V \end{aligned} \quad (5)$$

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_0, \text{head}_1, \dots, \text{head}_H) \quad (6)$$

where Q, K , and V are obtained from the input via the fully connected layer, B is the location information, d_k is the dimension size of K , i is the number of attention heads, and head_i represents the output of a single attention.

The advantage of the above purposely designed fusion mechanism is that it helps realize cross-sectional information transfer at multilevel features, thus prompting effective interaction between BEV and RIV feature maps. In addition, by introducing an attention mechanism to adaptively and dynamically adjust the weights of the features, the model's ability to understand and capture the potential correlation relationship between the BEV and the RIV is further enhanced, which in turn helps improve the expressiveness and generalization of the model, and enhances the effectiveness of point cloud segmentation.

E. Hybrid Edge Metric Loss

Specifically, in this article, ground-truth semantic segmentation is used to extract the ground truth edges of multiview

point cloud projection images. First, the neighboring pixels of different labels are determined as the boundary. Second, since projection causes information loss and affects the edge extraction effect, this article adopts morphological closure operation, i.e., confirming the labels of each pixel by the neighboring pixel categories in eight directions, so as to improve the clarity of the edges and obtain the ground truth boundary TB . Given the significant role of boundary information in the segmentation task, we have devised a lightweight boundary prediction header (BPH) to predict the boundary information for both BEV and RV images. The BPH comprises two fully connected layers, each followed by the rectified linear unit (RELU) activation function to learn nonlinear features. In addition, dropout is used to prevent overfitting, and the output of the convolution with one channel is used to predict the boundaries. The BPH output, akin to Canny, is a binary image. Consequently, to effectively capture the distinction between the predicted boundary and the actual boundary in the BGM output, we adopt the Dice loss as the boundary loss.

L_{et} and P denote the boundary prediction results for the BEV and RV branches, respectively, and Tb and Tr represent the true boundaries of BEV and RV, respectively. The calculation formula for Dice loss is as follows:

$$L_{bev} = 1 - \frac{2|T_b \cap P_b|}{|T_b| + |P_b|} \quad (7)$$

$$L_{rv} = 1 - \frac{2|T_r \cap P_r|}{|T_r| + |P_r|}. \quad (8)$$

Simultaneously, to ensure that the model considers the segmentation effect of the point cloud, Lovasz-softmax and cross-entropy are used to characterize the dissimilarity between the actual point cloud and the predicted point cloud. The formula is as follows, and ultimately, the sum of the three losses is used as the overall loss:

$$L_{rb} = \zeta_{lovasz}(T_p, P_p) + \zeta_{CE}(T_p, P_p) \quad (9)$$

$$\text{Final loss} = L_{bev} + L_{rv} + L_{rb}. \quad (10)$$

IV. EXPERIMENTS

A. Dataset

The SemanticKITTI dataset [37], a benchmark dataset for large-scale point cloud semantic segmentation and scene understanding widely used in the field of automated driving, uses the Velodyne HDL-64E lidar to capture a variety of rural areas and road scenes in the medium-sized city of Karlsruhe. Note that we have selected a number of scenes with complex scene elements as a test set to demonstrate the segmentation results. Experiments on the SemanticKITTI dataset validate the adaptability of the proposed method in different environments.

The RELIS-3D dataset [38] is designed for semantic segmentation in off-road environments for autonomous driving. This dataset was collected at Texas A&M University using a 64-channel Ouster OS1 lidar. It consists of 13 556 raw 3-D point cloud data, 6235 images, and their corresponding semantic labels. The point cloud data are split into 7800/2413/3343 for the training, validation, and test sets, respectively.

Note that our experiments were trained and tested using only 14 annotated categories.

To validate the model proposed in this article more comprehensively, this article also uses a self-collected dataset of real vehicles. The data consist of a vehicle equipped with a Velodyne 32 line LiDAR while considering an outdoor scenario.

B. Evaluation Metric

The mean intersection over union (mIoU) metric is commonly used to evaluate point cloud semantic segmentation tasks. It gauges the accuracy of the predicted segmentation mask by examining the overlap between the predicted labels and the ground-truth labels. The mIoU value is derived by calculating the intersection over union (IoU) score for each category and then averaging these scores across all the categories. The real-time performance evaluation metric is frames per second (FPS). The mIoU is calculated using the following formula:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (11)$$

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^C \text{IoU}_c \quad (12)$$

where TP , FP , and FN represent the number of true-positive points, false-positive points, and false-negative points, respectively, IoU denotes the IoU of class c , and C indicates the number of classes.

C. Implementation Details

Our proposed method is implemented in PyTorch and executed on a desktop running the Ubuntu 20.04 operating system, equipped with a single GeForce RTX 3090 GPU with 24-GB RAM. We conducted thorough experiments on SemanticKITTI, RELIS-3D, and self-collected real-vehicle datasets. The input dimensions for the RV branch are set to $W = 2048$ and $H = 64$, and for the BEV branch, they are set to $W = 320$ and $H = 480$. The model is optimized using a stochastic gradient descent (SGD) optimizer, with momentum and weight decay parameters set to 0.9–0.001, respectively. The initial learning rate is set to 0.01, and the training period extends to 120 epochs.

D. Comparative Experimental Validation

This section provides a comprehensive comparison of our proposed method with some highly competitive baseline methods, such as point-based, sparse voxel-based, single projection-based, and multiview-based.

1) *Comparison on the SemanticKITTI Dataset*: Based on the SemanticKITTI dataset, we conducted a comprehensive comparison between our proposed method and several representative baseline methods, presenting both the quantitative and qualitative results in Table I and Fig. 6, respectively. The metrics in Table I mainly include the IoU, mIoU, and single-frame prediction time (inclusive of data preprocessing time) for each category. The rows in the table represent

TABLE I
COMPARISON WITH DIFFERENT REPRESENTATIVE BASELINE METHODS ON THE SEMANTICKITTI DATASET

Methods	FPS	mIoU	Road	Parking	Sidewalk	Vegetation	Terrain	Truck	Car	Bicycle	Motor	Other-vehicle	Person	Building	Fence	Trunk	Pole	Traffic-sign
Cylinder3D [12]	5.6	69.5	93.2	41.3	79.6	83.9	63.0	75.4	95.4	75.2	53.0	44.9	73.3	90.6	60.9	67.5	65.4	49.7
SPVNAS [19]	3.9	69.0	91.7	43.5	74.9	85.6	63.7	82.4	96.0	81.6	40.3	62.8	65.2	88.9	56.7	61.5	60.8	48.4
PolarNet [17]	16	54.3	90.8	61.7	70.4	84.0	67.8	22.9	93.8	40.3	5.6	28.5	43.2	40.3	61.3	65.5	51.8	57.5
SalsaNext [28]	24	65.6	93.2	38.9	79.2	83.8	69.6	38.9	91.9	79.6	37.4	45.5	66.7	85.2	48.4	60.3	56.6	47.2
LENet [39]	26	64.2	92.4	68.6	78.2	84.5	69.1	40.0	93.9	55.7	48.6	44.0	67.4	55.7	68.8	69.4	59.7	62.4
GFNet [40]	10	65.4	93.6	72.5	80.8	85.2	69.3	31.7	96.0	53.2	48.3	47.3	62.8	53.2	73.9	71.1	61.8	68.0
AMVNet [34]	-	65.3	90.1	71.0	75.8	85.6	69.6	48.8	96.2	59.9	54.2	45.7	71.0	59.9	69.1	71.7	62.7	67.2
Our Method	11	68.6	94.2	72.3	78.4	86.2	71.2	43.4	96.2	58.4	52.6	49.1	65.3	59.3	69.2	71.4	64.8	68.6

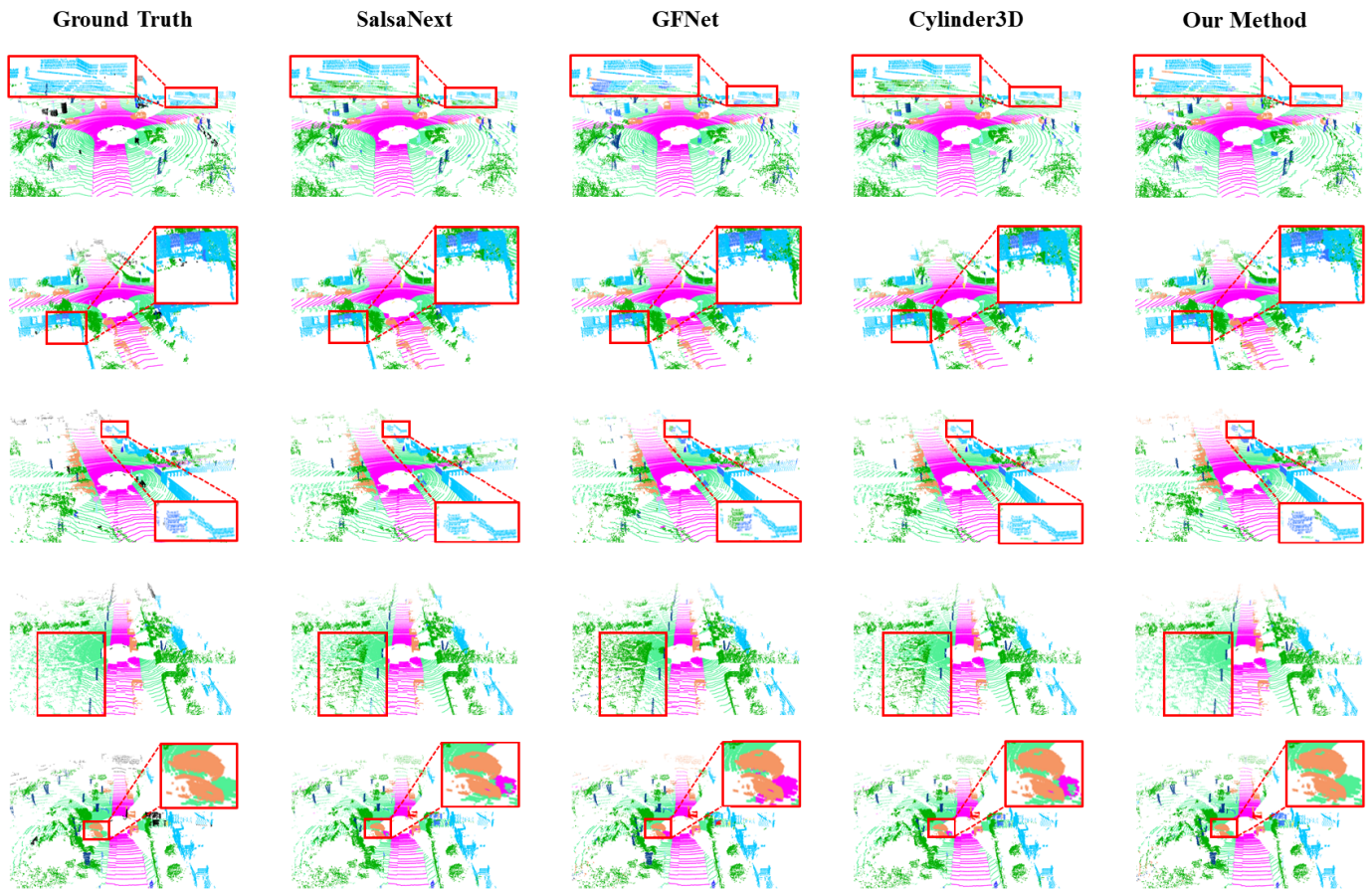


Fig. 6. Qualitative comparison with different representative baseline methods on the SemantickITTI dataset.

various baseline methods, including the point-based method voxel-based method Cylinder3D [12], SPVNAS [19], single-view projection-based methods PolarNet [17], SalsaNext [28], LENet [39], and multiview projection fusion-based methods GFNet [40], AMVNet [34], alongside the quantitative results of our proposed method.

Based on the data in Table I, our proposed method achieves an mIoU of 68.6%, marking a 4.4% improvement compared with LENet [39], which is the top performer among the single-view methods. We hypothesize that this improvement is mainly due to the fact that the multiview boundary attention guidance and feature interaction fusion module significantly enhance

TABLE II
COMPARISONS WITH DIFFERENT REPRESENTATIVE BASELINE METHODS ON THE RELLIS-3D DATASET

Methods	FPS	mIoU	Grass	Tree	Bush	Mud	Rubble	Concrete	Log	Pole	Water	Vehicle	Person	Fence	Barrier	Puddle
Cylinder3D [12]	6.7	45.6	64.9	76.5	71.7	10.6	2.2	80.4	5.4	63.7	0	50.3	87.0	11.8	80.5	33.0
SPVNAS [19]	6.0	42.6	64.1	76.0	71.5	8.5	5.2	70.0	15.0	54.4	0	48.5	85.7	10.9	64.8	22.4
SalsaNext [28]	40	42.5	65.3	79.6	73.5	11.5	6.6	77.8	22.1	44.2	0	26.6	84.4	13.2	63.6	26.0
GFNet [40]	12	43.6	64.3	76.4	72.4	10.1	7.2	73.5	21.5	49.8	0	34.4	83.1	12.6	79.3	24.2
Baseline [18]	20	44.6	65.8	79.0	73.9	12.9	9.3	77.2	23.8	50.4	0	27.9	85.7	12.9	78.4	23.8
Baseline [41]	14	44.5	65.6	79.3	72.4	10.4	9.2	73.8	24.6	50.4	0	32.4	85.7	13.2	78.3	24.6
Our Method	14	44.8	65.9	79.7	74.1	11.2	6.5	77.4	22.3	52.3	0	28.2	85.6	13.3	78.7	23.1

the segmentation performance among most element classes, drivable areas, and larger sized target classes such as vehicle targets and trees.

Compared with SalsaNext [28], our proposed method improves the mIoU metric by about 3%. This enhancement is attributed to the multiview boundary guidance with complementary fusion, leading to improved segmentation performance across various elemental classes, drivable regions, and larger sized targets, such as vehicles. In the multiview encoding branch, the BEV of the point cloud retains the contour space information of larger objects, which is helpful to improve the clarity of interclass boundary segmentation. Notably, Cylinder3D [12] and SPVNAS [19] also demonstrate satisfactory segmentation accuracy, benefiting from their ability to preserve spatial information through voxel and point fusion. It is important to note that the higher mIoU metrics achieved by these methods are primarily influenced by uncommon element categories in the environment, which constitute a small percentage of scene elements, rather than the category-important and rich targets such as cars, bicycles, motors, utility poles, and traffic signs.

Compared with the representative GFNet [40], the proposed method shows an improvement of approximately 3.2% in terms of mIoU. Objective analysis indicates that this is primarily attributed to our designed multiscale feature interaction fusion module, which enhances the complementary nature and context-awareness of multiview features. In addition, the multiview BGM contributes to a better focus on interclass boundary features, thereby achieving clearer boundary segmentation.

Although our proposed method slightly reduces mIoU by about 0.9% compared with the optimal representative method Cylinder3D [12], this is mainly due to the projection operation, which inevitably leads to the loss of information about the point cloud geometry. In addition, compared with the more computationally expensive methods, our proposed method effectively achieves a competitive balance between segmentation accuracy and real-time performance. This makes it

ideal for downstream tasks that rely on point clouds. Overall, the proposed approach benefits from a multiview boundary attention guide module, which can better encode interclass boundary features. In addition, the cross-fusion of multiview and multiscale features is designed to enhance contextual information for scene categories, thereby improving segmentation performance for key targets such as drivable areas, sidewalks, and vegetation, which are critical for unmanned ground system perception.

2) Comparison on the RELLIS-3D Dataset: To further evaluate the effectiveness of our proposed method on unstructured datasets, the challenging RELLIS-3D dataset is selected for comparative validation, and the baseline methods for comparison mainly include representative methods with outstanding performance such as Cylinder3D [12], SPVNAS [19], SalsaNext [28], etc.

Unlike SemanticKITTI, the RELLIS-3D dataset mainly collects unstructured scene point clouds. The quantitative and qualitative results of the comparison are shown in Table II and Fig. 7, respectively. Overall, due to the complexity of the unstructured environmental background, the performance of the methods involved in the comparison experiences varying degrees of decline.

Specifically, our proposed method achieves competitive performance in both mIoU and real-time metrics (FPS). Compared with the baseline SalsaNext [28], which relies on the range approach and exhibits the best performance, our method improves by about 2.3% in the mIoU metric, reaching 44.8%. Furthermore, the proposed method demonstrates satisfactory performance in terms of segmentation accuracy and real-time efficiency when compared with other baselines, such as SPVNAS [19]. Notably, Cylinder3D [12] achieves the best mIoU metrics in unstructured, challenging, off-road scenarios, primarily due to its effective preservation of the inherent spatial properties of the point cloud. However, despite the superior mIoU performance of Cylinder3D, its inference time is approximately four times that of our proposed method.

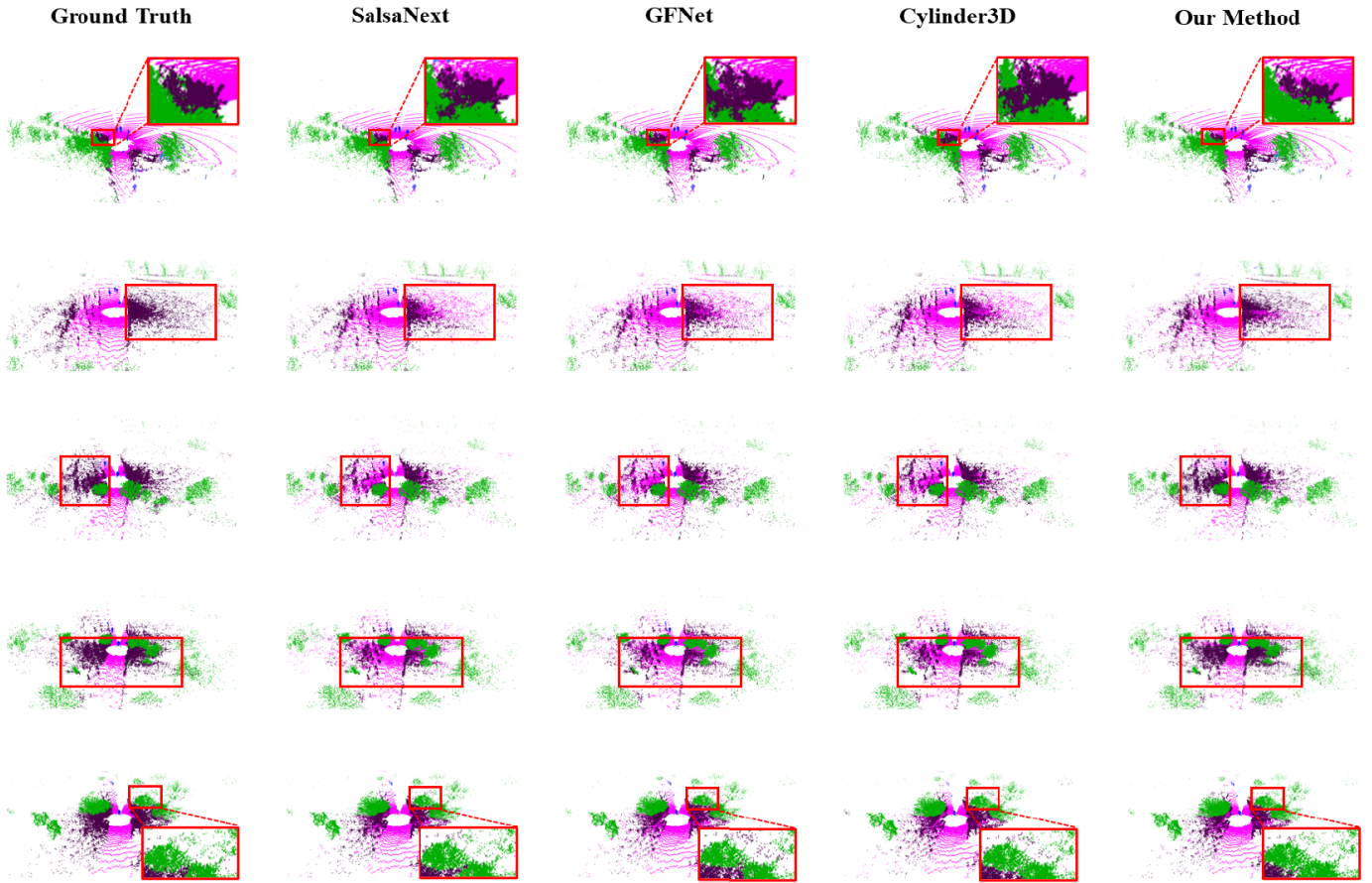


Fig. 7. Qualitative comparison with different representative baseline methods on the RELLIS-3D dataset.

Notably, compared with Cylinder3D, our proposed method improves the mIoU by about 1%, 3.2%, and 2.4% in segmenting some complex convex scene elements such as grass, trees, and bushes. This improvement is primarily attributed to the designed BGM and the combined hybrid boundary loss, enhancing the focus on interclass information. Moreover, the segmentation performance is enhanced for targets such as pole, person, and fence, which possess richer vertical information. The primary reason is the two-way fusion between BEV and RV at the feature level, mitigating the loss of vertical information by BEV. The superior interclass segmentation performance of our method is evident in the local zoomed-in image shown in Fig. 7.

3) Comparison on the Real Vehicle Dataset: To demonstrate the generalization and robustness of our method, we perform extra comparisons in real vehicle experiments using VLP-32C LiDAR sensor, which have sparser point clouds and different data distributions. Considering the segmentation results, we choose Cylinder3D and GFNet as the representativeness of semantic segmentation benchmarks.

In Fig. 8, we visualize the comparison results between these methods on three different scenes. The results illustrate that our method can segment objects and drivable areas more precisely, to be specific, more complete obstacle edges (see orange class in red circles) and more continuous road surfaces (see purple class in red circles).

4) Ablation Experiment: To further validate the contributions of each module designed in our proposed method,

we conducted comprehensive ablation experiments on the SemanticKITTI and RELLIS-3D datasets. The validation results presented below are based on models trained on the aforementioned datasets, with identical training and testing environments for all the ablation experiments.

a) Effects of fusion strategy: To demonstrate the effectiveness of our proposed method, we first compared it with the RV single-branch, BEV single-branch, and dual-branch methods, and the comparison results are shown in Fig. 9. It is evident that compared with the methods using a single branch, our approach improves mIoU by 5.3%, while the single-branch method improves mIoU by 5.9%. Furthermore, when compared with the integrated method, our approach enhances mIoU by 3.4%. This confirms that the combination of the BGM and the specialized multiview feature level fusion module effectively improves the performance of our method.

b) Effects of module fusion: To ensure the comprehensiveness of the ablation experiments, we similarly validated them on the RELLIS-3D dataset. Specifically, the ablation experiments were validated by replacing or deleting one of the inputs to the network and the key modules of the design, and the comparative results are shown in Table III: 1) a network with only RIV inputs was selected; 2) a network with only BEV inputs was selected; 3) a network with both RIV and BEV inputs was selected; 4) a network with the boundary bootstrap module was deleted; and 5) a network with the cross-transformer module deleted was selected.

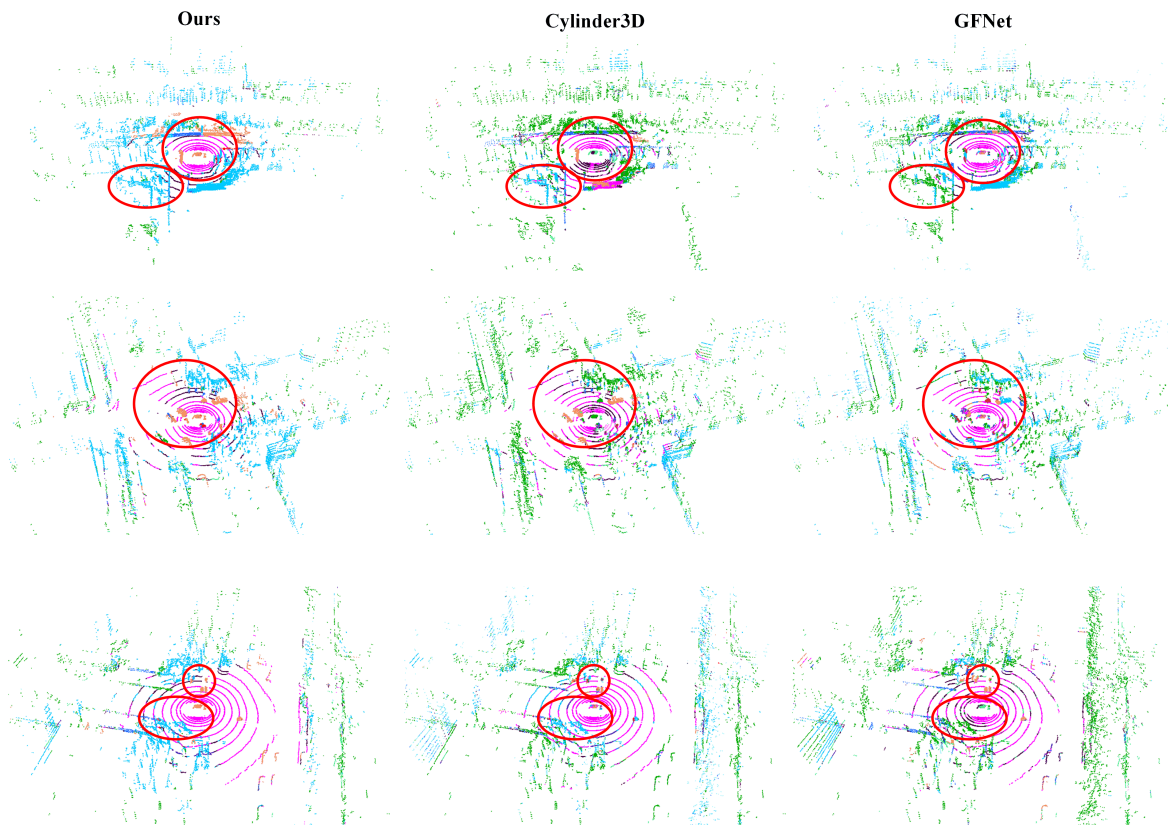


Fig. 8. Qualitative comparison with different representative baseline methods on the real vehicle dataset.

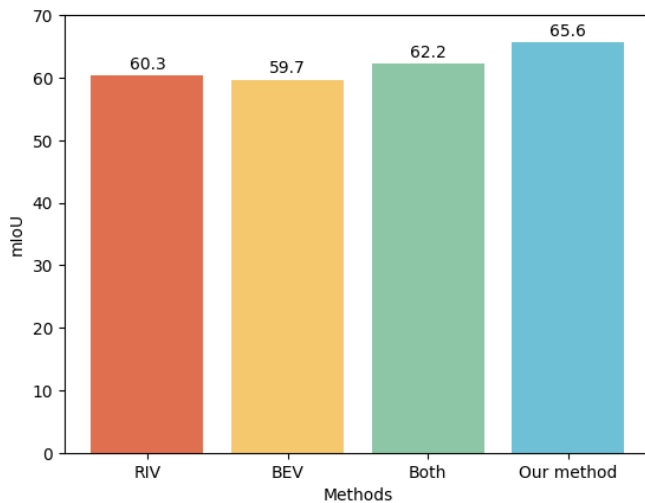


Fig. 9. Results of ablation studies on SemanticKITTI for different input types (in %).

As presented in Table III, our proposed method exhibits an improvement of approximately 2.4% in mIoU compared with the single-view input RIV. In addition, the introduction of BGM leads to an increase of about 0.6% in mIoU, accompanied by an additional 10.29 GFLOPs. This enhancement is primarily attributed to the synergistic effect of BGM and the hybrid loss function, reinforcing the model's attention to segmentation boundaries. Upon incorporating the MMIM module, the mIoU sees an improvement of approximately 0.5%, with an additional computational cost of 12.2 GFLOPs

TABLE III
QUANTITATIVE EVALUATION OF ABLATION VALIDATION BASED ON THE RELIS-3D DATASET

RV	BEV	BGM	MMIM	mIoU	FLOPS(G)	Time(ms)
✓	×	×	×	42.4	121.05	25
✓	✓	×	×	43.3	129.96	29
✓	✓	✓	×	43.9	140.25	33
✓	✓	×	✓	43.8	142.16	35
✓	✓	✓	✓	44.8	145.34	39

compared with the baseline. This improvement is mainly due to MMIM facilitating feature-level fusion between RIV and BEV, enabling information exchange and compensating for information deficits. Overall, compared with the baseline, our proposed method achieves an increase of around 2.4% in mIoU while adding an extra computational cost of 24.29 GFLOPs.

c) *Effects of remote segmentation:* It is necessary to emphasize that the segmentation performance of the LiDAR-scanned long-range scene region is particularly important considering that the LiDAR point cloud density decreases with distance. The statistical results depicted in Fig. 10 illustrate that our proposed method outperforms approaches relying solely on a single viewpoint (RIV-single, BEV-single). Notably, MMIM leverages the multiattention mechanism to effectively explore and bidirectionally fuse potential features from RIV and BEV at the feature level, making a significant contribution

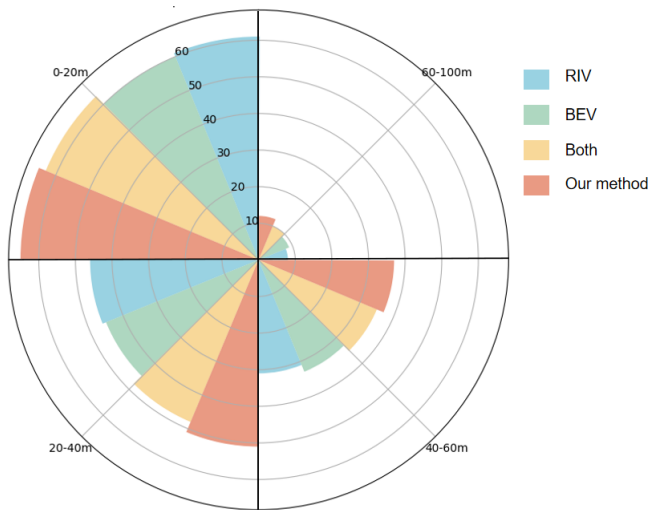


Fig. 10. Statistics of segmentation performance presented by the method are presented based on the LiDAR scanning distance.

to performance enhancement. In conclusion, comprehensive ablation validation results confirm that our proposed method ultimately improves the performance of point-cloud-based semantic segmentation by fully using the complementary features and unifying potential values between RIVs and BEVs, complemented by multiview fusion and multiboundary attention guidance modules.

V. CONCLUSION

In this study, our primary objective was to enhance the perceptual understanding of unmanned systems in complex background environments. To achieve this, we tailored a multiview, multiboundary, multilevel feature fusion approach for semantic segmentation of LiDAR point clouds. First, we introduced a multiboundary attention guidance module to augment focus on edge information among different categories. Second, a multiview multilevel feature interaction module is designed to effectively centralize information of different feature scales and capture remote dependencies between multiview features, which also helps bridge the information scale difference between BEV and RIV images. Finally, a hybrid boundary loss function was formulated to guide the training and optimization of the model. Through comprehensive validation experiments conducted on the SemanticKITTI urban dataset and the RELLIS-3D off-road dataset, our proposed method exhibited promising results in both the structured and unstructured scenarios. Compared with representative baseline methods with various point cloud inputs, our method has achieved a well-balanced performance between efficiency and accuracy, showcasing its competitiveness. For future research endeavors, our focus will be directed toward enhancing segmentation accuracy for small targets and addressing challenges associated with the loss of point cloud information due to RIV and BEV projections.

REFERENCES

- [1] D. Omeiza, H. Webb, M. Jirotko, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10142–10162, Aug. 2022.
- [2] D. Kong et al., "Explicit points-of-interest driven Siamese transformer for 3D LiDAR place recognition in outdoor challenging environments," *IEEE Trans. Ind. Informat.*, vol. 19, no. 10, pp. 10564–10577, Oct. 2023, doi: 10.1109/TII.2023.3240578.
- [3] Q. Hu et al., "Learning semantic segmentation of large-scale point clouds with random sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8338–8354, Jul. 2021.
- [4] H. Zhang, C. Liu, and W. Zhao, "Segmented trajectory planning strategy for active collision avoidance system," *Green Energy Intell. Transp.*, vol. 1, no. 1, Jun. 2022, Art. no. 100002.
- [5] J. Ruan, H. Cui, Y. Huang, T. Li, C. Wu, and K. Zhang, "A review of occluded objects detection in real complex scenarios for autonomous driving," *Green Energy Intell. Transp.*, vol. 2, no. 3, Jun. 2023, Art. no. 100092.
- [6] S. Xu, Q. Liu, Y. Hu, M. Xu, and J. Hao, "Decision-making models on perceptual uncertainty with distributional reinforcement learning," *Green Energy Intell. Transp.*, vol. 2, no. 2, Apr. 2023, Art. no. 100062.
- [7] Y. A. Alnaggar, M. Afifi, K. Amer, and M. ElHelw, "Multi projection fusion for real-time semantic segmentation of 3D LiDAR point clouds," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1799–1808.
- [8] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.
- [9] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2019, pp. 6411–6420.
- [10] Q. Hu et al., "RandLA-Net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11108–11117.
- [11] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.
- [12] X. Zhu et al., "Cylindrical and asymmetrical 3D convolution networks for LiDAR-based perception," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6807–6822, Oct. 2021.
- [13] M. Ibrahim, N. Akhtar, S. Anwar, and A. Mian, "SAT3D: Slot attention transformer for 3D point cloud semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 5456–5466, May 2023.
- [14] B. Wu, A. Wan, X. Yue, and K. Keutzer, "SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1887–1893.
- [15] C. Xu et al., "SqueezeSegV3: Spatially-adaptive convolution for efficient point-cloud segmentation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 1–19.
- [16] E. E. Aksoy, S. Baci, and S. Cavdar, "SalsaNet: Fast road and vehicle segmentation in LiDAR point clouds for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 926–932.
- [17] Y. Zhang et al., "PolarNet: An improved grid representation for online LiDAR point clouds semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9601–9610.
- [18] W. Xu, X. Li, P. Ni, X. Guang, H. Luo, and X. Zhao, "Multiview fusion driven 3-D point cloud semantic segmentation based on hierarchical transformer," *IEEE Sensors J.*, vol. 23, no. 24, pp. 31461–31470, Dec. 2023.
- [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] X. Li, G. Zhang, H. Pan, and Z. Wang, "CPGNet: Cascade point-grid fusion network for real-time LiDAR semantic segmentation," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 11117–11123.
- [21] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.
- [22] 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.
- [23] X. Wu et al., "Sparse fuse dense: Towards high quality 3D detection with depth completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5418–5427.

- [24] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3D object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 18442–18455.
- [25] J. Zhu, H. Huang, B. Li, and L. Wang, "E-CRF: Embedded conditional random field for boundary-caused class weights confusion in semantic segmentation," 2021, *arXiv:2112.07106*.
- [26] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 4376–4382.
- [27] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "RangeNet++: Fast and accurate LiDAR semantic segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4213–4220.
- [28] T. Cortinhal, G. Tzelepis, and E. E. Aksoy, "SalsaNext: Fast, uncertainty-aware semantic segmentation of LiDAR point clouds," in *Proc. 15th Int. Symp. Visual Comput.*, San Diego, CA, USA. Cham, Switzerland: Springer, Oct. 2020, pp. 207–222.
- [29] S. Wang, J. Zhu, and R. Zhang, "Meta-RangeSeg: LiDAR sequence semantic segmentation using multiple feature aggregation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 9739–9746, Oct. 2022.
- [30] R. Razani, R. Cheng, E. Taghavi, and L. Bingbing, "Lite-HDSeg: LiDAR semantic segmentation using lite harmonic dense convolutions," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 9550–9556.
- [31] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images," *IEEE Trans. Multimedia*, vol. 24, pp. 2192–2204, 2021.
- [32] W. Zhou, S. Dong, J. Lei, and L. Yu, "MTANet: Multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 48–58, Jan. 2023.
- [33] X. Liu et al., "Deep transfer learning for intelligent vehicle perception: A survey," *Green Energy Intell. Transp.*, vol. 2, no. 5, Oct. 2023, Art. no. 100125.
- [34] V. E. Liong, T. N. T. Nguyen, S. Widjaja, D. Sharma, and Z. J. Chong, "AMVNet: Assertion-based multi-view fusion network for LiDAR semantic segmentation," 2020, *arXiv:2012.04934*.
- [35] H. Ma, H. Yang, and D. Huang, "Boundary guided context aggregation for semantic segmentation," 2021, *arXiv:2110.14587*.
- [36] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 1874–1883.
- [37] J. Behley et al., "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Aug. 2019, pp. 9297–9307.
- [38] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "RELLIS-3D dataset: Data, benchmarks and analysis," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 1110–1116.
- [39] B. Ding, "LENet: Lightweight and efficient LiDAR semantic segmentation using multi-scale convolution attention," 2023, *arXiv:2301.04275*.
- [40] H. Qiu, B. Yu, and D. Tao, "GFNet: Geometric flow network for 3D point cloud semantic segmentation," 2022, *arXiv:2207.02605*.
- [41] P. Ni, X. Li, D. Kong, K. Wei, and Y. Hu, "Scene-adaptive 3-D semantic segmentation method based on multiphase edge enhancement for intelligent vehicles," *IEEE Sensors J.*, vol. 23, no. 24, pp. 31471–31482, Dec. 2023.