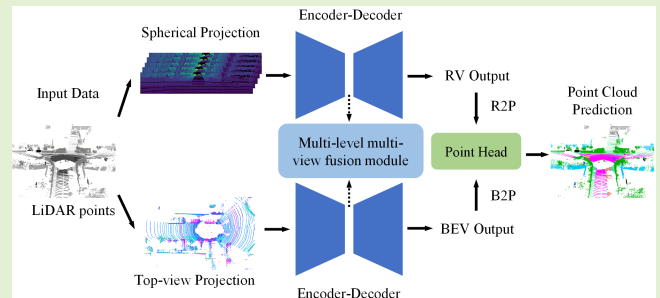


Multiview Fusion Driven 3-D Point Cloud Semantic Segmentation Based on Hierarchical Transformer

Wang Xu^{ID}, Xu Li^{ID}, Peizhou Ni^{ID}, Xingxing Guang, Hang Luo, and Xijun Zhao^{ID}

Abstract—Three-dimensional semantic segmentation is a key task of environment understanding in various outdoor scenes. Due to the sparsity and varying density of point clouds, it becomes challenging to obtain fine-grained segmentation results. Previous point-based and voxel-based methods suffer from the expensive computational cost. Recent 2-D projection-based methods, including range-view (RV), bird-eye-view (BEV), and multiview fusion methods, can run in real time, but the information loss during the projection leads to the low accuracy. Also, we find that the occlusion and interlacing problems exist in single projection-based methods and most multiview fusion networks only focus on the output-level fusion. Considering the above issues, we propose a multilevel multiview fusion network using attention modules and hierarchical transformer, which ensures the effectiveness and efficiency mainly by the following three aspects: 1) the spatial-channel attention module (SCAM) integrates contextual information between points and learn differences of each channel's features; 2) the proposed geometry-based multiprojection fusion module (GMFM) achieves the geometric feature alignment between RV and BEV and fuses the features of the two views at both feature level and output level; and 3) we introduce KPConv to replace KNN, which can reduce the information loss during the postprocessing. Experiments are conducted on both structured and unstructured datasets, including urban dataset SemanticKITTI and off-road dataset Rellis3D. Our results achieve a better performance compared to other projection-based methods and are comparable with the state-of-the-art Cylinder3D.

Index Terms—3-D point cloud, multihead attention, multiview fusion, semantic segmentation.



I. INTRODUCTION

SEMANtic segmentation plays a crucial role in the field of unmanned driving and environmental perception, which can provide us with more precise perception information. It increases the vehicle maneuvering and path planning ability of unmanned autonomous vehicles [1], [2]. In recent years, with the development of LiDAR sensors, it has become

easier to obtain environmental information. Compared to 2-D images, a point cloud can provide distance information and topological relationship of elements in different scenarios [3]. Also, it is not affected by changes in ambient light and has a strong anti-interference ability. By providing these capabilities, 3-D semantic segmentation is essential for achieving efficient autonomous driving and has been widely researched [4].

Many models have been proposed to improve the performance of LiDAR semantic segmentation. These methods can be divided into the following three categories: point-based method [9], [10], [11], [12], [40], [41], voxel-based method [13], [14], [15], [39], and projection-based method [16], [17], [18], [19], [20], [21].

Although point-based method has made impressive progress, it is hard to capture the associated features between adjacent points efficiently, making it have low accuracy and slow reasoning speed. Voxel-based method converts point cloud data into regular 3-D voxels and allows 3-D convolutional neural networks (CNNs) to operate on the voxels directly. However, it may lead to sparsity and granular information loss during the voxelization process and the efficiency of 3-D CNNs decreases with the increase in data size [14].

Manuscript received 27 September 2023; accepted 22 October 2023. Date of publication 6 November 2023; date of current version 14 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3904404, in part by the Primary Research and Development Plan of Jiangsu Province under Grant BE2022053-5, and in part by the Collective Intelligence and Collaboration Laboratory under Grant QXZ23012201. The associate editor coordinating the review of this article and approving it for publication was Dr. Guofa Li. (Corresponding author: Xu Li.)

Wang Xu, Xu Li, and Peizhou Ni are with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: 220223306@seu.edu.cn; lixu.mail@163.com; 230198300@seu.edu.cn).

Xingxing Guang, Hang Luo, and Xijun Zhao are with the China North Artificial Intelligence and Innovation Research Institute, and Collective Intelligence and Collaboration Laboratory (CIC), Beijing 100072, China (e-mail: guangxingxing@yeah.net; lhbuaa@hotmail.com; heejunzhao@163.com).

Digital Object Identifier 10.1109/JSEN.2023.3328603

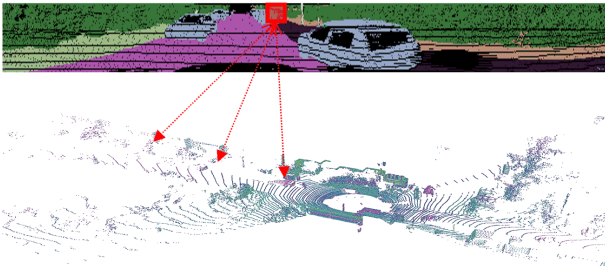


Fig. 1. Occlusion problem in RV compared with point clouds.

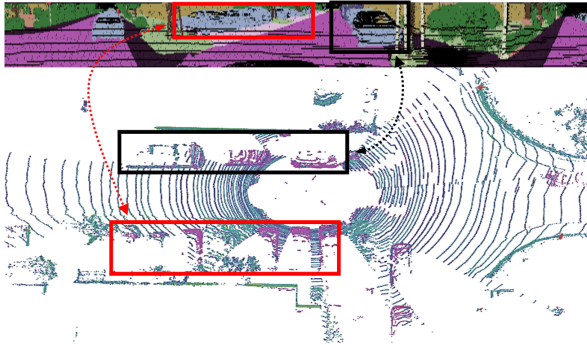


Fig. 2. Difference between RV and BEV.

Due to the compactness of images and the lightness of 2-D CNNs, projection-based methods can achieve superior segmentation effect while ensuring real-time performance. Projection-based method usually uses top-down projection or spherical projection to deal with the raw point cloud, thus forming bird's-eye view (BEV) and range view (RV). Obviously, RV and BEV contain different distributed information of raw point cloud due to the different projection methods. A point cloud projecting into an RV image inevitably leads to the information loss of distant elements for several points may be projected into one pixel, as shown in Fig. 1. However, BEV images preserve those features, as it can project points from the top view, making it more efficient to discern objects in different areas especially at the farther distance, as seen in Fig. 2, but it may lose the detailed information of point clouds while projecting to BEV grids. Additionally, RV demonstrates superior performance at vertically extended objects, such as motorcycle and person, while BEV is sensitive to those objects with large spatial on the xy plane, such as car and vegetation. Thus, multiprojection-fusion methods [24], [25], [26], [27] have become a trend in point cloud semantic segmentation based on projection. However, most networks based on multiprojection fusion only focus on the input-level fusion or output-level fusion, ignoring the correspondence and alignment between different views in the feature level.

In order to solve the above challenges, we propose a multilevel multiprojection fusion network, which can aggregate the features deeply both in feature-level and output-level. First, we design a spatial-channel-attention module (SCAM) to learn the association of global information and the correlation between different channels. Second, the proposed geometry-based multiprojection fusion module (GMFM) can accomplish the feature-level fusion of the two views' features employing multihead cross attention (MHCA) mechanism.

Finally, inspired by Kochanov et al. [28], we use KPConv [12] to replace KNN to back project two branches' output to pointwise output, which can implement the output-level fusion more efficiently. Our article's contributions are as follows.

- 1) To increase the segmentation accuracy of unevenly distributed and distant points, a multilevel multiprojection fusion network is proposed to aggregate the benefits of the RV and BEV. It fuses RV and BEV both at the feature level and output level, which greatly reduces the information loss brought by RV projecting 3-D points into a common pixel and BEV projecting points to 2-D grids. It achieves the state-of-art performance among projection-based methods.
- 2) The proposed SCAM can integrate contextual information between points and learn differences in each channel's features, which improves the feature expressiveness of remote and sparse points.
- 3) We design an accurate and efficient GMFM for the feature-level fusion between RV and BEV. It aligns the features based on the geometric association between RV and BEV. Then, the aligned features are fed into multihead cross-attention to complete the deep fusion.

II. RELATED WORKS

The existing LiDAR semantic segmentation methods can be divided into the following aspects according to the input and way of feature extraction: point-based methods, voxel-based methods, and projection-based methods.

A. Point-Based Methods

Methods that directly operate on raw point clouds can be categorized into three types: multilayer perception (MLP)-based and CNN-based. PointNet [9] is a pioneering method that processes point clouds by applying MLP to each point and using max pooling to integrate all point information. While PointNet effectively extracts features from individual points, it tends to perform poorly when dealing with complex scenes due to its lack of consideration for neighboring point information. RandLa-Net [11] presents a method to reduce the number of points through random sampling while preserving the geometric features of the raw point clouds as much as possible. KPConv [12] introduces a novel convolution operator called kernel point convolution, which directly takes neighboring points as an input and applies spatially located weights during processing. This approach enhances the extraction of local features and improves performance. In general, point-based methods suffer from slow processing speeds, low accuracy, and limited robustness, particularly when dealing with complex scenes and small objects.

B. Voxel-Based Methods

Voxel-based methods are commonly employed to process 3-D point clouds by converting them into regular cubes and applying 3-D convolutions within each voxel. In the cases of [13], [14], [15], and [31], the point cloud data are partitioned into 3-D voxel units. However, the accuracy of these methods is heavily influenced by the sparsity and irregularity of the

point cloud. Recognizing the sparsity and varying density of point clouds, Cylinder3D [13] proposed a voxel division approach that transforms points from Cartesian coordinates to Cylinder coordinates.

Additionally, Cylinder3D designed an asymmetrical residual block, a CNN module, to enhance the extraction of cuboid-style features. AF2S3Net proposed a novel multi-branch attention feature fusion module used in the encoder and a unique adaptive feature selection module used in the decoder. Ibrahim et al. [39] introduced the first-ever slot attention transformer-based technique to effectively model object-centric features in point cloud data. They use cylindrical splits of space for voxelization and computes channelwise positional embeddings before repetitively encoding the point cloud with slot attentions. While these methods offer improvements, they still suffer from the computational cost associated with 3-D convolutions. This limitation poses challenges for meeting the real-time requirements for LiDAR semantic segmentation.

C. Projection-Based Methods

Projection-based methods can achieve better real-time performance as they project 3-D point clouds into either RV [32] or BEV. This conversion allows for the adoption of efficient fully convolutional network (FCN) on 2-D pseudo images. By leveraging the 2-D nature of the projected data, Wu et al. [18], [19], Xu et al. [20], Aksoy et al. [21], Ding [22], and Cheng et al. [23] achieve good results while maintaining computational efficiency. RangeNet++ [32] proposes an accelerated KNN for postprocessing, which plays an important role in the following range-based methods. Salsanext [16] transforms point clouds into RV images with five channels through spherical projection and introduces a context module to capture the global context information based on [13]. Meta-RangeSeg [33] adds a range residual image to the five channels extracted in Salsanext. Zhang et al. [34] propose a BEV representation of the point cloud and represent it with regular grids. Considering the imbalanced spatial distribution of points, PolarNet [17] assigns points to polar grid based on quantized azimuth and radius.

While projection-based methods strike a balance between performance and accuracy, it is important to note that these representations may result in the loss of some raw point cloud information due to the reduction of data during the projection process. RV methods can project multiple 3-D points onto a common pixel, and BEV methods are not sufficiently sensitive to those vertically extended objects. Considering the above elements, we can achieve better label predictions by leveraging the strengths of both views. TornadoNet [35] adopts a unique approach by employing a pillar-projection-learning module (PPL) to extract features from the BEV representation. These features are then fed into a reverse-ordered encoder-decoder CNN to model the RV data. In contrast, MPF [27] utilizes two separate models to process RV and BEV data independently, and the predicted softmax probabilities from both branches are combined for the final predictions. Taking a step further, AMVNet [24] refines the uncertain predictions obtained from RV and BEV by employing a point head. The point head

focuses on refining the predictions that exhibit disagreements between the two branches, thereby improving the overall accuracy of the segmentation. Different from our proposed method, Li et al. [25] and Alnaggar et al. [27] process BEV and RV separately during encoding and decoding and only do the late fusion at the output level. Especially at those distant and sparse points, our method performs better than single BEV or RV networks and those multiview-based methods that ignore feature-level fusion.

III. METHODOLOGY

RV loses some important information of distant objects due to the staggered and occlusion distributions of the scene elements. However, BEV projection can compensate for the shortcomings of range projection in the mentioned aspect. As shown in Fig. 3, we propose multilevel multiview fusion network, which fuses the features of BEV and RV at the feature level by the GMFM. The details are described in the following.

A. RV Branch

We project a 3-D LiDAR scan onto spherical RV pseudo images as the input of the RV branch, which consists of five channels including (x, y, z) , intensity, and range. Before doing the projection, we first project a point (x, y, z) from cartesian space to spherical space as follows.

In order to ensure both real-time performance and accuracy of our algorithm, we refer to Salsanext's encoder-decoder structure as the backbone of our RV branch. However, we observe that different channels within the range image contain distinct information. To address this issue, we introduce the channel-attention module (CAM), which aims to enhance the channel features by learning the relationships between the various channels while preserving the original features. Another challenge we encounter is the lack of contextual information across the network. To overcome this limitation, we devise a spatial-attention module (SAM) to capture global context alongside more detailed spatial information. By incorporating larger receptive fields, this module enables the network to grasp complex correlations among different classes.

As shown in Fig. 4, X_{in} is the input of the RV branch; X_{out} is the enhanced features processed by the SCAM. SCAM consists of SAM and CAM, which denote the spatial-attention module and channel-attention module.

B. BEV Branch

To learn the effective features from BEV pseudo images, top-down projection is employed to transform a point cloud to BEV image. Inspired by Zhang et al. [17], we use polar partitioning instead of Cartesian partitioning, which makes the points more evenly distributed across each BEV grid. Furthermore, considering that several points may exist in one BEV grid, to avoid arbitrarily handcrafting the features for each grid, we use an MLP to capture the distribution of points in each grid with a higher dimensional feature representation. Then, a max pooling layer is used to aggregate the features of

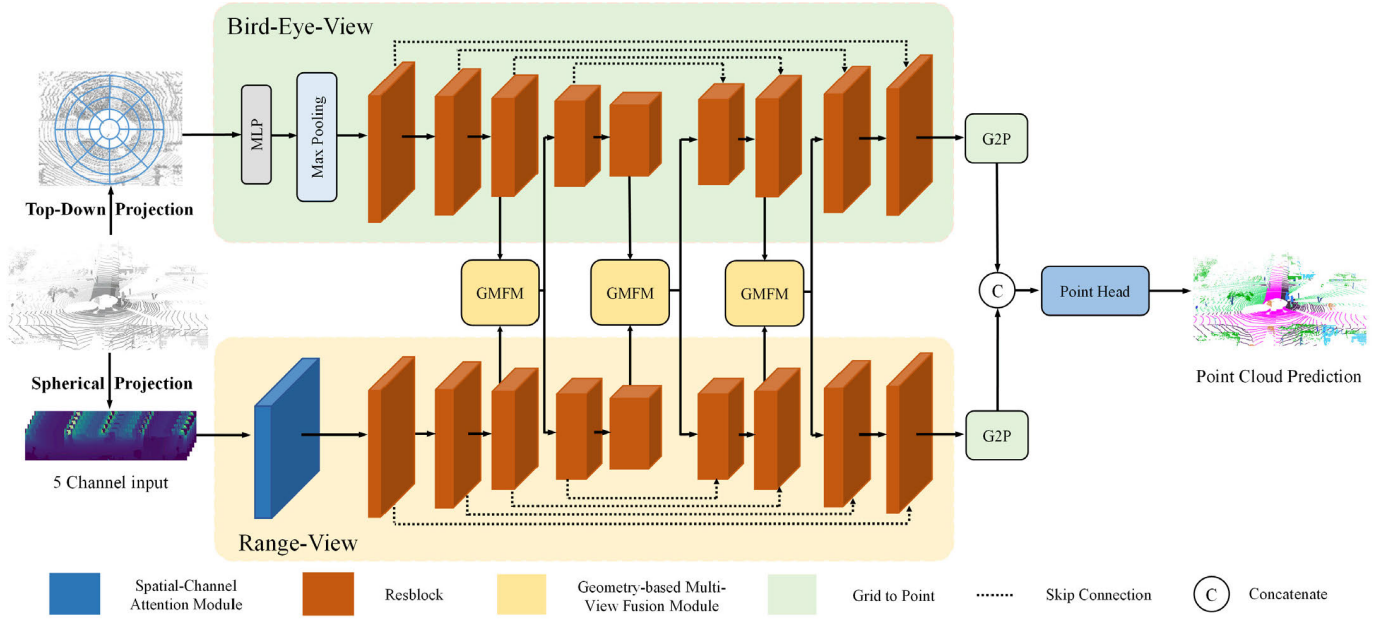


Fig. 3. Overall architecture of the proposed network. Here, the points are projected into RV and BEV, which are separately fed into two branches for multilevel feature extraction. Then, intermediate features from three layers are fused through GMFM and the fused features of each branch are fed into next layer, respectively. Finally, the outputs of both branches are fed into G2P module and point head to obtain the final point cloud prediction.

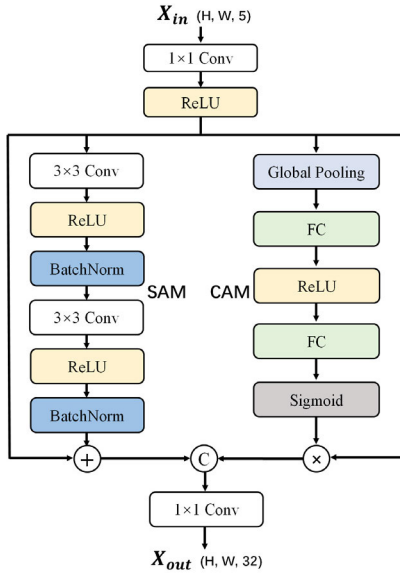


Fig. 4. Architecture of the SCAM.

points in each grid to represent the features possessed by that grid.

After extracting the features of BEV images, to achieve a balance between computation costs and segmentation accuracy, we choose ResNet [37] with 34 layers as an encoder network, which has a maximum of 256 channels. The decoder blocks apply transpose convolutions and fuse upsampled features with those of the early blocks via skip connections.

C. Feature-Level Multiview Fusion Module

To achieve the complementary advantages of RV and BEV images in LiDAR semantic segmentation, we design a GMFM

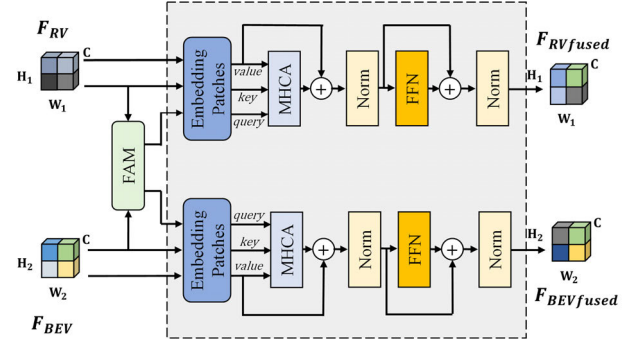


Fig. 5. Overall architecture of GMFM. FAM aligns the two branches' features using geometric relationship. Then, the aligned features are fed into MHCA to derive the fused features.

to fuse the two views' features based on the geometric association between RV and BEV. As shown in Fig. 5, the GMFM divided into two branches, where one is to obtain the fused RV features and the other is for the fused BEV features. Here, I will take the RV branch as an example to analyze the process of fusion.

First, the input RV and BEV features are fed into the feature alignment module (FAM), which implements the feature alignment from BEV to RV by calculating the geometric transformation matrices $\mathbf{M}_{B \rightarrow R}$. During the transformation, we utilize the original point cloud as an intermediary agent. We first calculate the matrices $\mathbf{M}_{B \rightarrow P} \in \mathbb{Z}^{H_b \times W_b}$ in the following equation, which align the features of RV images to original point clouds, as shown in Fig. 6:

$$\mathbf{M}_{B \rightarrow P} = \begin{bmatrix} n_{(0,0)} & \cdots & n_{(0,W_b-1)} \\ \vdots & \ddots & \vdots \\ n_{(H_b-1,0)} & \cdots & n_{(H_b-1,W_b-1)} \end{bmatrix} \quad (1)$$

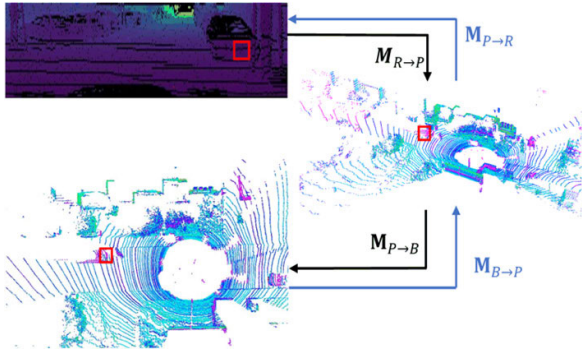


Fig. 6. Architecture of FAM.

where (H_b, W_b) are the width and height of the 2-D BEV images, and $\{n_{(i,j)} | 0 \leq i \leq H_b - 1, 0 \leq j \leq W_b - 1\}$ is the $n_{(i,j)}$ th point, which projects on (i, j) coordinates; then, the transformation matrices $\mathbf{M}_{P \rightarrow R} \in \mathbb{Z}^{N \times 2}$ from original points to RV image can be seen as follows:

$$\mathbf{M}_{P \rightarrow R} = [\mathbf{r}_0 \quad \cdots \quad \mathbf{r}_{N-1}] = \begin{bmatrix} u_0 & \cdots & u_{N-1} \\ v_0 & \cdots & v_{N-1} \end{bmatrix} \quad (2)$$

where N is the number of points, and $\{\mathbf{r}_k = (u_k, v_k) | 0 \leq k \leq N - 1\}$ represents the projected pixel coordinates of 2-D RV image, corresponding to the k th point. By calculating the $\mathbf{M}_{B \rightarrow P}$ and $\mathbf{M}_{P \rightarrow R}$, we obtain the geometric transformation matrices $\mathbf{M}_{B \rightarrow R} \in \mathbb{Z}^{H_r \times W_r \times 2}$

$$\mathbf{M}_{B \rightarrow R} = \begin{bmatrix} \mathbf{r}_{n_{(0,0)}} & \cdots & \mathbf{r}_{n_{(0,W_b-1)}} \\ \vdots & \ddots & \vdots \\ \mathbf{r}_{n_{(H_b-1,0)}} & \cdots & \mathbf{r}_{n_{(H_b-1,W_b-1)}} \end{bmatrix}. \quad (3)$$

Finishing the feature alignment, we record the transformed BEV features as $\mathbf{F}_{B2R}$, the concatenating result of $\mathbf{F}_{B2R}$ and $\mathbf{F}_{RV}$ as $\mathbf{F}_{RB}$. Then, $\mathbf{F}_{RV}$ and $\mathbf{F}_{RB}$ are, respectively, reshaped into the sequences of flattened 2-D patches. We select $\mathbf{F}_{RB}$ as queries q and $\mathbf{F}_{RV}$ as keys k and values v . Then, MHCA and feed-forward network in (5) are used to aggregate the features of RV and BEV as $\mathbf{F}_{RV}^{\text{fused}}$. The BEV branch can be processed similar to get the fused BEV features

$$\begin{cases} \mathbf{F}_{\text{MHCA}} = \text{Norm}(\mathbf{F}_{RB} + \text{MHCA}(\mathbf{F}_{RB}, \mathbf{F}_{RV}, \mathbf{F}_{RV})) \\ \mathbf{F}_{RV}^{\text{fused}} = \text{Norm}(\mathbf{F}_{\text{MHCA}} + \text{FFN}(\mathbf{F}_{\text{MHCA}))). \end{cases} \quad (4)$$

D. Output-Level Multiview Fusion Module

The output-level multiview fusion module is accountable for transforming the two views' output of decoder module into final semantic predictions of all points in the input 3-D point cloud. Inspired by Qiu et al. [36], we introduce grid-to-point (G2P) module to back project the RV output and BEV output to point view, which consists of a grid sampling and KPConv [12]. The results are, respectively, recorded as $\mathbf{Q}_R$ and $\mathbf{Q}_B$. As shown in (5), we feed the concatenated result of $\mathbf{Q}_R$ and $\mathbf{Q}_B$ into the point head composed of an MLP, a max pooling layer, and a fully connected (FC) layer to obtain the final 3-D point cloud semantic predictions $\mathbf{Q}_P$

$$\mathbf{Q}_P = \text{pointhead}(\text{cat}(\mathbf{Q}_R, \mathbf{Q}_B)). \quad (5)$$

E. Loss Function

Let $\mathbf{Y}_P$ to be the pointwise 3-D semantic segmentation labels. Meanwhile, $\mathbf{Q}_R$, $\mathbf{Q}_B$, and $\mathbf{Q}_P$ [see (5)] are the final 3-D predictions of RV branch, BEV branch, and fusion output.

Following [16], [23], we adopt a combination of cross-entropy loss $\mathcal{L}_{\text{CE}}$ and iousoftmax loss $\mathcal{L}_{\text{iousoftmax}}$ [38] to calculate the loss of three parts. $\mathbf{L}_{\text{rb}}$, $\mathbf{L}_{\text{bev}}$, and $\mathbf{L}_{\text{rv}}$, respectively, represent the loss of fusion output, BEV output, and RV output. Each part is defined as

$$\begin{cases} \mathbf{L}_{\text{rb}} = \mathcal{L}_{\text{CE}}(\mathbf{Q}_P, \mathbf{Y}_P) + \mathcal{L}_{\text{iousoftmax}}(\mathbf{Q}_P, \mathbf{Y}_P) \\ \mathbf{L}_{\text{bev}} = \mathcal{L}_{\text{CE}}(\mathbf{Q}_B, \mathbf{Y}_P) + \mathcal{L}_{\text{iousoftmax}}(\mathbf{Q}_B, \mathbf{Y}_P) \\ \mathbf{L}_{\text{rv}} = \mathcal{L}_{\text{CE}}(\mathbf{Q}_R, \mathbf{Y}_P) + \mathcal{L}_{\text{iousoftmax}}(\mathbf{Q}_R, \mathbf{Y}_P). \end{cases} \quad (6)$$

To optimize all the modules in the proposed segmentation model, the total loss can be calculated into $\mathbf{L}$ as $\alpha \mathbf{L}_{\text{rb}} + \beta \mathbf{L}_{\text{bev}} + \gamma \mathbf{L}_{\text{rv}}$, where α , β , and γ are set as 1.0, 0.5, and 0.5 to balance the loss terms.

IV. EXPERIMENT

A. Dataset

We select the public urban dataset Semantic-KITTI [6] and off-road dataset RELLIS-3D [7] to verify our method's effectiveness and robustness. The detailed information of the two datasets is as follows.

Semantic-KITTI dataset is a widely used benchmark dataset for semantic segmentation and scene understanding tasks in autonomous driving. The dataset provides high-resolution point cloud data collected by a Velodyne LiDAR sensor.

RELLIS-3D dataset consists of five sequences collected from three unpaved roads in the Texas A&M University RELLIS campus, which contains annotations for 13 556 LiDAR scans. Our experiments use only 14 classes with annotations (except void) for training and testing.

For further validating the segmentation performance and the scenario adaptation of the proposed method, we collected 100 frames of the unstructured scene. We consider outdoor parking lots and roads without clear road boundaries or lacking marking lines as unstructured scenes. The unstructured scene test set we build is collected by a vehicle equipped with a Velodyne 32-line LiDAR and a forward-facing monocular camera, ensuring synchronized data acquisition between the two sensors.

B. Evaluation Metric

The mean intersection over union (mIoU) is an evaluation metric commonly used for point cloud semantic segmentation tasks. It measures the quality of the predicted segmentation masks by assessing the overlap between the predicted and ground truth labels. To calculate mIoU, each predicted point is assigned a label, and the intersection and union of points with the same label are computed. The IoU score for each class is then calculated as the ratio of the intersection to the union. Finally, the average IoU scores across all classes are computed to obtain the mIoU.

C. Implementation Details

For Semantic-KITTI, we set the input shape of RV branch as $(W_{\text{rv}} = 2048 \text{ and } H_{\text{rv}} = 64)$. Meanwhile, the BEV branch's

TABLE I

RESULTS (IN %) AND TIME OF TYPICAL STREAMS OF METHODS AND OUR PROPOSED METHOD ON SEMANTICKITTI. THE METHODS ARE GROUPED AS POINT-BASED, VOXEL-BASED, SINGLE-PROJECTION-BASED, AND MULTIPROJECTION FUSION NETWORKS. *: RESULT REPRODUCED BY [26]

Methods		road	parking	bicycle	car	bicycle	motorcycle	vegetation	trunk	terrain	pole	fence	truck	Other-vehicle	person	bicyclist	motorcyclist	sidewalk	other-ground	traffic-sign	mIoU(%)	FPS(Hz)
Point	RandLA-Net [11]	90.7	60.3	26.0	94.2	26.0	25.8	81.4	61.3	56.3	81.4	56.3	40.1	38.9	49.2	48.2	7.2	73.7	20.4	61.3	53.9	1.3
	KPConv [12]	88.8	61.3	30.2	96.0	30.2	42.5	84.8	69.2	69.1	56.4	64.2	33.4	44.3	61.5	61.6	11.8	72.7	31.6	47.4	58.8	3.8
Voxel	MinkowskiNet [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.1*	3.4
	SPVNAS [31]	90.2	67.6	50.6	97.2	50.6	50.4	86.1	73.4	71.0	64.3	66.9	56.6	58.0	67.4	67.1	50.3	75.4	21.8	67.3	67.0	3.9
	Cylinder3D [13]	91.4	65.1	67.6	97.1	67.6	64.0	85.4	71.8	68.5	62.6	66.5	59.0	58.6	73.9	67.9	36.0	75.5	32.3	65.6	67.8	5.6
Single-Projection	PolarNet [17]	90.8	61.7	40.3	93.8	40.3	30.1	84.0	65.5	67.8	51.8	61.3	22.9	28.5	43.2	40.2	5.6	74.4	21.7	57.5	54.3	16
	SalsaNext [16]	91.7	63.7	48.3	91.9	48.3	38.6	81.8	63.6	66.5	54.3	64.2	38.9	31.9	60.2	59.0	19.4	75.8	29.1	62.1	59.5	24
	LENet [22]	92.4	68.8	55.7	93.9	55.7	48.6	84.5	69.4	69.1	59.7	68.8	40.1	44.0	67.4	64.4	27.7	78.2	32.3	62.4	64.2	26
Multi-Projection	MPF[27]	90.5	64.7	28.2	92.7	30.5	30.5	83.4	56.6	69.8	46.0	59.0	26.9	25.2	42.5	45.5	9.5	74.3	32.0	54.9	53.6	28.5
	GFNet [36]	93.6	72.5	53.2	96.0	53.2	48.3	85.2	71.1	69.3	61.8	73.9	31.7	47.3	62.8	57.3	44.7	80.8	31.2	68.0	65.4	10
	AMVNet [24]	90.1	71.0	59.9	96.2	59.9	54.2	85.6	71.7	69.6	62.7	69.1	48.8	45.7	71.0	65.7	11.0	75.8	32.4	67.2	65.3	-
	Ours	93.8	74.5	54.3	96.4	54.3	51.2	86.2	72.1	69.0	65.2	70.6	48.1	49.1	65.3	64.8	36.2	78.2	33.7	69.3	67.5	12

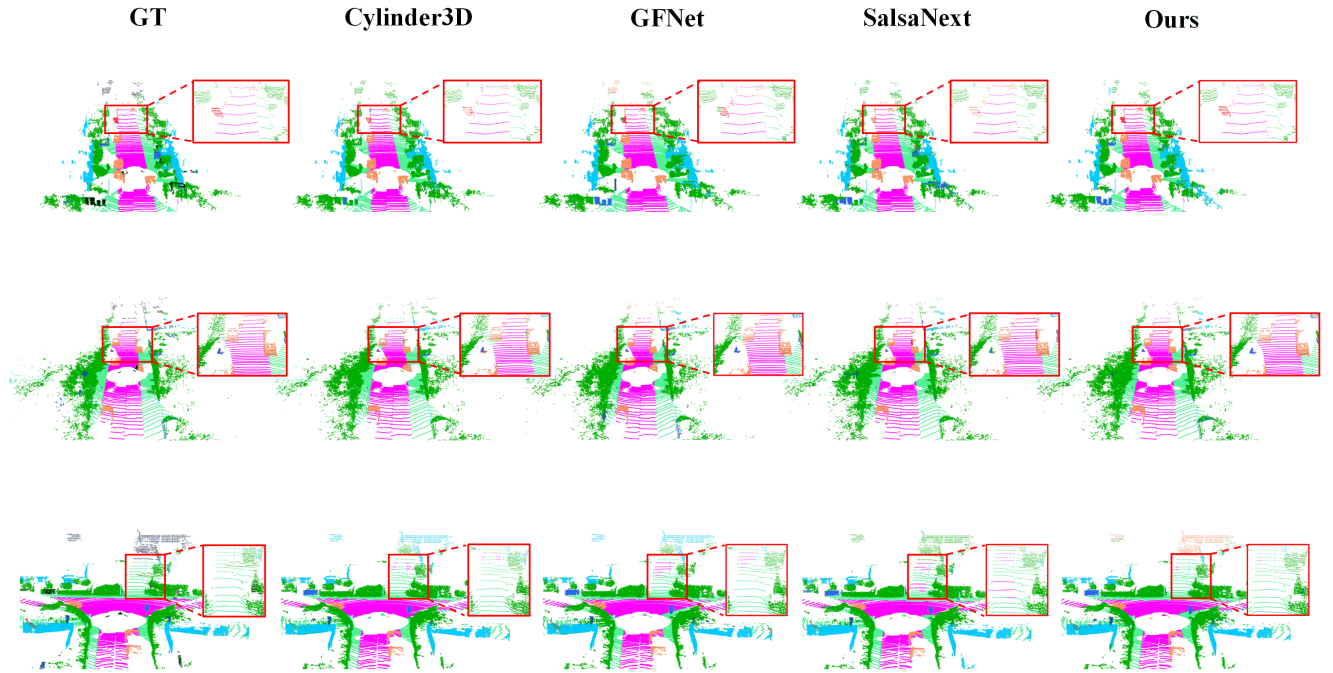


Fig. 7. Visual comparison with different state-of-the-art methods on the Semantic KITTI dataset.

inputs are ($W_{bev} = 320$ and $H_{bev} = 480$). The input of RELLIS3D is the same as that of Semantic-KITTI.

All experiments are conducted on a desktop with 128-GB RAM, an Intel Xeon¹ E5-2678 clocked at 2.5 GHz under Ubuntu 20.04 using Pytorch. Our network is trained for 50 epochs with a batch size of 16, which takes around 48 h on four NVIDIA RTX 3090 GPUs of 24-GB memory. We utilize stochastic gradient descent (SGD) as the optimizer with a weight decay of 0.001, a momentum of 0.9, and an initial learning rate of 0.02.

D. Experimental Results

We compare the method we propose with those highly competitive methods, which can be separated into point-based,

sparse voxel-based, single-projection-based, and multiview-based methods. We trained each method using the same training strategy in its paper on our experimental platform.

1) *Results on SemanticKITTI*: The test results compared to other state-of-art approaches can be seen in Table I, which are grouped as point-based, voxel-based, and projection-based methods from top to bottom. We analyze the results on SemanticKITTI mainly compared with the following methods: Cylinder3D (voxel-based), Salsanext (single projection-based), and GFNet (multiprojection-based), as is shown in Table I and Fig. 7.

When compared with Cylinder3D, our proposed method gets a slight drop in mIoU (0.3%). This is because projection-based methods cannot avoid the information loss during the projection process. Owing to the multi-view fusion based on hierarchical transformer, our method

¹Registered trademark.

TABLE II

RESULTS (IN %) AND TIME OF TYPICAL STREAMS OF METHODS AND OUR PROPOSED METHOD ON RELLIS-3D. THE METHODS ARE GROUPED AS POINT-BASED, VOXEL-BASED, AND PROJECTION-BASED NETWORKS

Methods		Grass	Tree	Pole	Water	Vehicle	Log	Person	Fence	Bush	Concrete	Barrier	Puddle	Mud	Rubble	mIoU(%)	FPS(Hz)
Point	RandLA-Net	62.1	76.7	42.8	0	34.8	6.3	82.9	10.2	69.8	72.8	54.3	17.6	8.5	1.7	38.6	1.3
Voxel	Cylinder3D	64.9	76.5	63.8	0	50.3	5.5	87.0	11.8	71.8	80.4	80.5	33.0	10.6	2.3	45.6	6.7
	SPVNAS	64.1	76.1	54.4	0	48.6	15.1	85.8	11.0	71.5	70.0	64.8	22.4	8.5	5.2	42.7	6.0
Projection	SalsaNext	65.3	79.6	44.3	0	26.6	22.2	84.4	13.5	73.6	77.9	63.7	26.0	11.6	6.6	42.5	40
	GFNet	64.3	76.4	49.8	0	34.4	21.5	83.1	12.6	72.4	73.5	79.3	24.2	10.1	7.2	43.6	12
	Ours	66.1	80.3	50.4	0	32.4	24.6	85.7	15.2	72.4	73.8	78.3	24.6	10.4	9.2	44.5	14

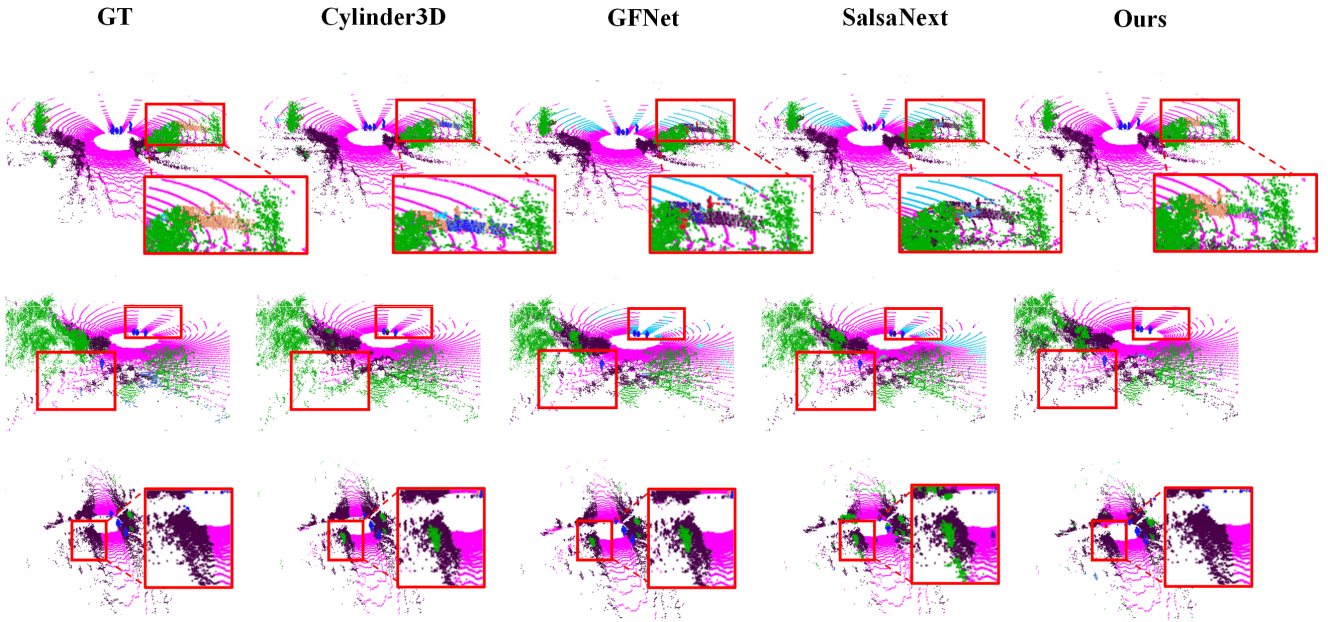


Fig. 8. Visual comparison with different state-of-the-art methods on RELLIS-3D dataset.

achieves an improvement of vertically extended objects, such as pole and fence. Besides, it takes much less time (107.1 ms faster) owing to the implementation of effective feature aggregation operations and lightweight network backbone.

Compared with Salsanext, our method gets an improvement mIoU of 8.0%. Owing to the multiview fusion between RV and BEV, it has an improvement on most categories, especially those objects with large spatial on the xy plane, such as car, parking, and road.

Compared with GFNet, our method obtains a 2.1% increase in mIoU. This is because the proposed SCAM enhances the contextual and channel information. Besides, the GMFM could achieve a deep fusion between two views by using the proposed GMFM.

2) Results on RELLIS-3D: Different from SemanticKITTI, RELLIS-3D dataset is recorded in unstructured environment. Therefore, the mIoU of each algorithm decrease due to the scenario's complexity. In order to verify the effectiveness of our algorithm, we select [11], [13], [16], and [26], which achieve extraordinary results among the different kinds of methods to make comparison. We mainly analyze the results with

Cylinder3D (voxel-based) and Salsanext (projection-based), as is shown in Table II and Fig. 8.

When compared with Cylinder3D, our method gets a slight drop in mIoU (1.1%). This is because its cylindrical partition and asymmetrical 3-D convolution networks could well handle the sparsity and varying density. However, our method achieves an improvement of those objects, which are staggered in the off-road environment, such as grass and tree. Besides, our method takes much less time (78.3 ms faster).

When compared with Salsanext and GFNet, our method achieves an improvement of 2.0% mIoU and 0.9% respectively, and outperforms both of them on most classes. This consistent improvement validates the effectiveness of the proposed GMFM and SCAM.

3) Results on Our Own Test Set: For further verifying the robustness and adaptivity of our method in different and complex scenarios, we perform extra experiments on our unstructured test set. The data were collected at different locations on a totally different platform from the above two training sets using Velodyne 32-line LiDAR sensor, which ensures the distribution of point cloud features (x , y , z , reflectivity, and noise) adaptive for these proposed methods.

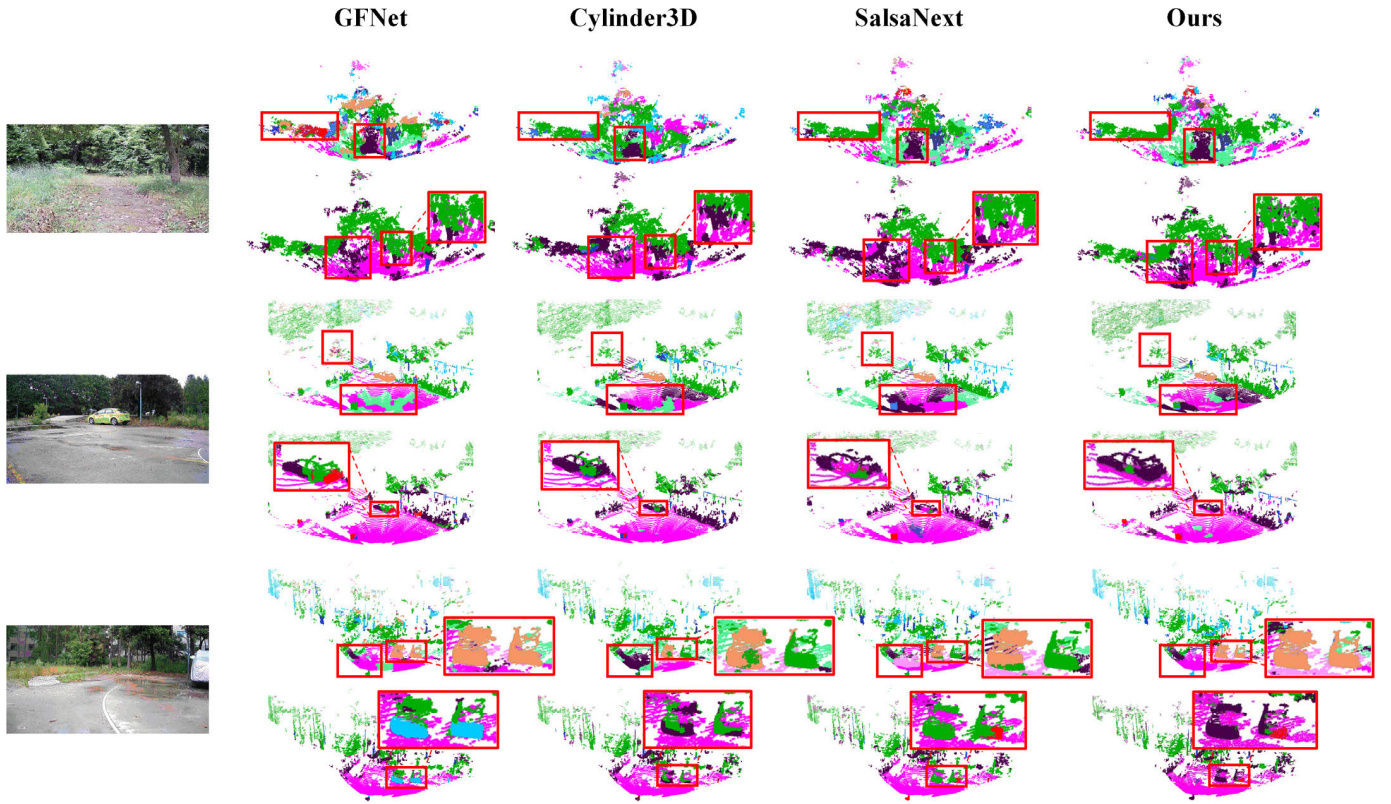


Fig. 9. Sample results of different models on our test set. The first row of each scenario represents the predictions using saved model parameters from SemanticKITTI training set, while the second row is from RELLIS-3D training set.

We test Cylinder3D (voxel-based), Salsanext (single projection-based), and GFNet (multiprojection-based) on our own datasets. The results can be seen in Fig. 9.

When compared with Cylinder3D, our method achieves comparable performance. However, when dealing with occluded objects, the segmentation accuracy of our method is significantly better. Also, our method spends less time than Cylinder3D.

Compared with Salsanext and GFNet, the segmentation performance in most categories improved a lot benefiting from the deep fusion of RV and BEV through hierarchical transformer. It can be seen that the car and road in Fig. 9 selected by the red boxes are segmented more completely.

E. Ablation Study

We design exhaustive ablation experiments to validate the efficacy of our proposed methodologies. These methods are trained with the same parameters in SemanticKITTI validation sets.

To determine the efficacy of our proposed network, we first compare our approach to the RV-single, BEV-single, and ensemble approaches. Compared to the RV-single and BEV-single methods, fusing the two branches only at the output level results in 2.9 and 4.1 mIoU improvements, as shown in Table III. Moreover, after adding our proposed blocks, it outperforms the ensemble method (+3.8 mIoU).

We mainly focus on the segmentation accuracy of objects, which are more than 40 m away from the LiDAR. As is shown in Fig. 10, our method is more efficient than RV-single, BEV-single, and ensemble methods.

TABLE III
RESULTS (IN %) OF ABLATION STUDIES FOR DIFFERENT INPUT TYPES IN SEMANTICKITTI VALIDATION SETS

Methods	mIoU
RV	59.5
BEV	58.3
Ensemble	62.4
Ours	66.2

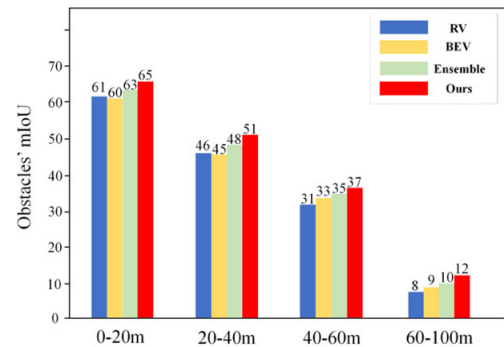


Fig. 10. mIoU at varying radial distance for the SemanticKITTI validation sets.

Subsequently, we analyze the effects of each incorporated component, including the SCAM, the GMFM, and KPConv, as shown in Table IV. SCAM increases the mIoU by 1.2%, proving its efficacy. GMFM combines the characteristics of RV and BEV based on geometric transformation and MHCA at the feature level, thereby enhancing performance (+1.8 mIoU). In addition, when we replace KNN with KPConv to combine the RV and BEV features, the proposed network obtains a

TABLE IV

RESULTS (IN %) OF ABLATION STUDIES FOR NETWORK COMPONENTS IN SEMANTICKITTI VALIDATION SETS

Module				mIoU
Main branch	SCAM	GMFM	KPConv	
✓				62.4
✓	✓			63.6
✓	✓	✓		65.4
✓	✓	✓	✓	66.2

66.2% mIoU in performance for the SemanticKITTI validation sets. These results indicate that the proposed method can effectively propagate complementary information between RV and BEV, thereby enhancing each branch's performance.

V. CONCLUSION

In this article, we present a multiview multilevel fusion method for LiDAR semantic segmentation that accomplishes the deep fusion between BEV and RV methods at both the output and feature levels. In particular, three modules are intended to improve the efficacy of semantic segmentation. First, the SCAM, comprised of the spatial attention module and the channel attention module, is presented to incorporate contextual and channel information of input features. Second, we devise a GMFM that implements the geometric alignments between the features of the RV and BEV branches and then uses an MHCA module to fuse the aligned features. Finally, we replace KNN with KPConv to enhance the precision of postprocessing.

Experiments are conducted on the SemanticKITTI urban dataset, the RELLIS-3D off-road dataset, and our own test dataset. We compare our method to several point-based, voxel-based, single-projection-based, and multiprojection-based methods, respectively. The results demonstrate that our method can account for both efficiency and accuracy in both structured and unstructured scenarios, and the ablation studies demonstrate that each component of our newly introduced method enhances the efficacy of the network.

REFERENCES

- [1] H.-F. Wang et al., "Vehicle-road environment perception under low-visibility condition based on polarization features via deep learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17873–17886, Oct. 2022, doi: [10.1109/TITS.2022.3157901](#).
- [2] D. K. Dewangan and S. P. Sahu, "Driving behavior analysis of intelligent vehicle system for lane detection using vision-sensor," *IEEE Sensors J.*, vol. 21, no. 5, pp. 6367–6375, Mar. 2021, doi: [10.1109/JSEN.2020.3037340](#).
- [3] Z. Qiu, F. Yan, Y. Zhuang, and H. Leung, "Outdoor semantic segmentation for UGVs based on CNN and fully connected CRFs," *IEEE Sensors J.*, vol. 19, no. 11, pp. 4290–4298, Jun. 2019, doi: [10.1109/JSEN.2019.2893892](#).
- [4] Y. Li et al., "Deep learning for LiDAR point clouds in autonomous driving: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3412–3432, Aug. 2021, doi: [10.1109/TNNLS.2020.3015992](#).
- [5] P. Liu, G. Yu, Z. Wang, B. Zhou, R. Ming, and C. Jin, "Uncertainty-aware point-cloud semantic segmentation for unstructured roads," *IEEE Sensors J.*, vol. 23, no. 13, pp. 15071–15080, Jul. 2023, doi: [10.1109/JSEN.2023.3266802](#).
- [6] J. Behley et al., "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 9296–9306, doi: [10.1109/ICCV.2019.00939](#).
- [7] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "RELLIS-3D dataset: Data, benchmarks and analysis," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Xi'an, China, May 2021, pp. 1110–1116, doi: [10.1109/ICRA48506.2021.9561251](#).
- [8] Z. Huang, C. Lv, Y. Xing, and J. Wu, "Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11781–11790, May 2021, doi: [10.1109/JSEN.2020.3003121](#).
- [9] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 77–85.
- [10] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," 2017, *arXiv:1706.02413*.
- [11] Q. Hu et al., "RandLA-Net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 11105–11114.
- [12] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotequi, F. Goulette, and L. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 6410–6419, doi: [10.1109/ICCV.2019.00651](#).
- [13] 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. 2022, doi: [10.1109/TPAMI.2021.3098789](#).
- [14] 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.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4490–4499, doi: [10.1109/CVPR.2018.00472](#).
- [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)*, Nashville, TN, USA, Jun. 2021, pp. 12542–12551.
- [16] T. Cortinhal, G. Tzelepis, and E. E. Aksoy, "SalsaNext: Fast, uncertainty-aware semantic segmentation of LiDAR point clouds for autonomous driving," in *Proc. ISVC*, San Diego, CA, USA, 2020, pp. 207–222.
- [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)*, Seattle, WA, USA, Jun. 2020, pp. 9598–9607.
- [18] 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)*, Brisbane, QLD, Australia, May 2018, pp. 1887–1893, doi: [10.1109/ICRA.2018.8462926](#).
- [19] 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)*, Montreal, QC, Canada, May 2019, pp. 4376–4382, doi: [10.1109/ICRA.2019.8793495](#).
- [20] C. Xu et al., "SqueezeSegV3: Spatially-adaptive convolution for efficient point-cloud segmentation," 2020, *arXiv:2004.01803*.
- [21] 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)*, Las Vegas, NV, USA, Oct. 2020, pp. 926–932, doi: [10.1109/IV47402.2020.9304694](#).
- [22] B. Ding, "LENet: Lightweight and efficient LiDAR semantic segmentation using multi-scale convolution attention," 2023, *arXiv:2301.04275*.
- [23] H. Cheng, X. Han, and G. Xiao, "CENet: Toward concise and efficient LiDAR semantic segmentation for autonomous driving," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Taipei, Taiwan, Jul. 2022, pp. 1–6, doi: [10.1109/ICME52920.2022.9859693](#).
- [24] V. Erin Liong, T. Ngoc Tho Nguyen, S. Widjaja, D. Sharma, and Z. J. Chong, "AMVNet: Assertion-based multi-view fusion network for LiDAR semantic segmentation," 2020, *arXiv:2012.04934*.
- [25] 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)*, Philadelphia, PA, USA, May 2022, pp. 11117–11123, doi: [10.1109/ICRA46639.2022.9811767](#).
- [26] H. Tang, Z. Liu, and S. Zhao, "Searching efficient 3D architectures with sparse point-voxel convolution," in *Proc. 16th Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2020, pp. 685–702.

- [27] Y. Ali 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.
- [28] D. Kochanov, F. K. Nejadasl, and O. Booi, "KPRNet: Improving projection-based LiDAR semantic segmentation," 2020, *arXiv:2007.12668*.
- [29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [30] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal ConvNets: Minkowski convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3070–3079, doi: [10.1109/CVPR.2019.00319](https://doi.org/10.1109/CVPR.2019.00319).
- [31] H. Tang et al., "Searching efficient 3D architectures with sparse point-voxel convolution," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–12.
- [32] 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.
- [33] 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.
- [34] C. Zhang, W. Luo, and R. Urtasun, "Efficient convolutions for real-time semantic segmentation of 3D point clouds," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 399–408.
- [35] M. Gerdzhev, R. Razani, E. Taghavi, and L. Bingbing, "TORNADO-Net: Multiview total variation semantic segmentation with diamond inception module," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Xi'an, China, May 2021, pp. 9543–9549, doi: [10.1109/ICRA48506.2021.9562041](https://doi.org/10.1109/ICRA48506.2021.9562041).
- [36] H. Qiu, B. Yu, and D. Tao, "GFNet: Geometric flow network for 3D point cloud semantic segmentation," 2022, *arXiv:2207.02605*.
- [37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [38] M. Berman, A. R. Triki, and M. B. Blaschko, "The Lovász-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.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4413–4421, doi: [10.1109/CVPR.2018.00464](https://doi.org/10.1109/CVPR.2018.00464).
- [39] 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, doi: [10.1109/TITS.2023.3243643](https://doi.org/10.1109/TITS.2023.3243643).
- [40] F. Guan and L. Guo, "MVPNet: Multiple view pyramid network for LiDAR point cloud prediction based on spatial-temporal feature learning," in *Proc. 2nd Int. Conf. Robot., Autom. Artif. Intell. (RAAI)*, Singapore, Dec. 2022, pp. 244–249, doi: [10.1109/RAAI56146.2022.10092969](https://doi.org/10.1109/RAAI56146.2022.10092969).
- [41] J. Park, C. Kim, S. Kim, and K. Jo, "PCSCNet: Fast 3D semantic segmentation of LiDAR point cloud for autonomous car using point convolution and sparse convolution network," *Expert Syst. Appl.*, vol. 212, Feb. 2023, Art. no. 118815.
- [42] A. Maiti, S. O. Elberink, and G. Vosselman, "TransFusion: Multi-modal fusion network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Vancouver, BC, Canada, Jun. 2023, pp. 6537–6547, doi: [10.1109/CVPRW59228.2023.00695](https://doi.org/10.1109/CVPRW59228.2023.00695).



Wang Xu received the B.S. degree from the School of Instrument Science and Technology, Southeast University, Nanjing, China, in 2022, where he is currently pursuing the M.S. degree. His research interests include multisensor fusion semantic segmentation, computer vision, and intelligent vehicles.



Xu Li received the Ph.D. degree in instrument science and technology from Southeast University, Nanjing, China, in 2006.

He is currently a Professor with the School of Instrument Science and Engineering, Southeast University. His research interests include intelligent driving behavior analysis and modeling, intelligent vehicle navigation, sensing and control, and ADAS.



Peizhou Ni received the B.S. degree in instrument science and technology from Sichuan University, Chengdu, China, in 2016. He is currently pursuing the Ph.D. degree with the Department of Communication Engineering, School of Instrument Science and Engineering, Southeast University, Nanjing, China.

His research interests include intelligent vehicles, deep learning, and semantic segmentation.



Xingxing Guang received the B.S. degree in automation engineering from the Taiyuan Institute of Technology, Taiyuan, China, in 2012, and the M.E. and Ph.D. degrees in control science and engineering from Harbin Engineering University, Harbin, China, in 2015 and 2021, respectively.

She is now working with the China North Artificial Intelligence and Innovation Research Institute and Collective Intelligence and Collaboration Laboratory (CIC). Her research interests

include computer vision, environment perception, localization, and motion planning.



Hang Luo received the B.S. and M.E. degrees in control engineering from the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China, in 2013 and 2016, respectively, and the Ph.D. degree from the Institute of Measurement and Control, Faculty of Mechanical Engineering, Leibniz University Hannover, Hannover, Germany, in 2022.

He is now working with the China North Artificial Intelligence and Innovation Research Institute and Collective Intelligence and Collaboration Laboratory (CIC). His research interests include SLAM, multiview stereo, collaborative mapping, and autonomous exploration.



Xijun Zhao received the B.S. degree in vehicular engineering and the Ph.D. degree in mechanical engineering from Beijing Institute of Technology, Beijing, China, in 2007 and 2011, respectively.

He is now working with the China North Artificial Intelligence and Innovation Research Institute and Collective Intelligence and Collaboration Laboratory (CIC). His research interests include perception, localization, motion planning, and control of unmanned ground vehicles.