MFINet: Multi-view Fusion and 2D-3D Interaction Enhancement for Real-Time LiDAR Semantic Segmentation

Nan Ma, Zhijie Liu, Yiheng Han*

Beijing University of Technology, Beijing, China
manan123@bjut.edu.cn, liuzhijie@emails.bjut.edu.cn, hanyiheng@bjut.edu.cn

Abstract

LiDAR semantic segmentation is a key task in advanced autonomous driving systems. Projection-based methods exhibit real-time potential due to their efficiency, but suffer from inevitable 3D information loss and rely on time-consuming post-processing, limiting overall performance. To address this, we propose MFINet, a real-time semantic segmentation network based on multi-view fusion and 2D-3D interaction enhancement. It adopts a three-branch architecture that integrates 3D Point View (3D-PV), 2D Bird's Eye View (2D-BEV) and 2D Range View (2D-RV) to make full use of 2D and 3D representation. From 3D to 2D, we design a 3D Point Feature Projector (3DPFP), which injects 3D features into the 2D BEV and RV pseudo-images to retain effective 3D information. From 2D to 3D, a Feature Enhancement (FE) module is designed to leverage the advantages of 2D information in extracting geometric and semantic features. We also introduce a 2D-3D Fusion Head (FH) to aggregate point features from multiple views. Besides, we incorporate a Multi-Scale Dilated Attention (MSDA) module with a sliding window strategy to enhance feature discrimination. Extensive experiments on the SemanticKITTI and NuScenes benchmarks demonstrate that MFINet outperforms existing methods on the SemanticKITTI, NuScenes val set and achieves competitive results on the NuScenes test set.

Code — https://github.com/anonymous-ai26/MFINet

Introduction

In autonomous driving and robotic perception systems, LiDAR has become a crucial sensor for 3D environmental perception due to its ability to provide high-precision, structured three-dimensional spatial information. Compared to 2D images, point cloud data offers more accurate representation of positions and distances, enabling more precise characterization of object shapes and spatial relationships. Among these, point cloud semantic segmentation aims to assign precise semantic labels to each point, serving as a key step for environment modeling and scene understanding.

In recent years, deep learning has driven rapid advances in point cloud semantic segmentation. Depending on the form of data representation, existing methods can be broadly categorized into three types: point-based (Zhao et al. 2021; Hui et al. 2023; Han et al. 2024), voxel-based (Cheng et al. 2021; Yan et al. 2021; Li, Dai, and Ding 2022; Ibrahim et al. 2023), and projection-based (Zhang et al. 2020; Cortinhal, Tzelepis, and Erdal Aksoy 2020; Zhao, Bai, and Huang 2021; Li et al. 2022). Point-based methods operate directly on raw point clouds, preserving 3D structure to the greatest extent. However, their difficulty in efficiently modeling irregular data leads to high computational costs, making real-time inference challenging. Voxel-based methods discretize point clouds into regular grids to improve computational efficiency but face trade-offs between accuracy and memory consumption. In contrast, projection-based methods map 3D point clouds into 2D pseudo-images, allowing efficient modeling using mature 2D CNN architectures. These methods offer a good balance between real-time performance and deployment flexibility, making them a mainstream choice for lightweight applications. However, the projection process inherently involves dimensionality reduction, inevitably causing 3D information loss. For example, the 2D Range View (2D-RV) has advantages in preserving depth information but loses part of the spatial structural information. The

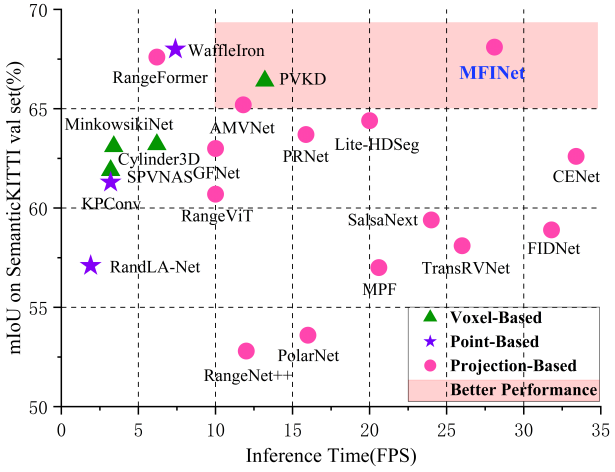

Figure 1: LIDAR semantic segmentation accuracy vs speed on SemanticKITTI val set.

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

2D Bird's Eye View (2D-BEV) provides a clear planar geometric layout and is well-suited for modeling road topology, but it fails to retain feature information along the vertical direction. Additionally, due to the sparsity of LiDAR data, the resulting pseudo-images contain a lot of empty pixels, which can introduce noise during feature learning and degrade the model's accuracy and stability.

Motivated by the above findings, We propose MFINet, a real-time semantic segmentation network based on multi-view fusion and 2D-3D interactive enhancement, which adopts a three-branch architecture comprising representations of 3D Point View (3D-PV), 2D Bird's Eye View (2D-BEV) and Range View (2D-RV). This design fully leverages the complementary information across different views to alleviate the loss of 3D information caused by the projection process. From 3D to 2D, we design a 3D Point Feature Projector (3DPFP) to inject 3D features into the 2D-BEV and 2D-RV pseudo-images, effectively enhancing geometric fidelity during the projection process. From 2D to 3D, a Feature Enhancement (FE) module is employed to leverage the advantages of 2D information in extracting geometric and semantic features. We also introduce a 2D-3D Fusion Head (FH) to aggregate point features from multiple views and semantic levels to enable end-to-end and real-time semantic prediction. In addition, we incorporate Multi-Scale Dilated Attention (MSDA) module into the shallow layers of the 2D-BEV and 2D-RV encoder branches. This module leverages a sliding window strategy to efficiently model sparse features and enhance feature discrimination. To ensure real-time performance, the key of our pipeline is to combine a brief 3D-PV branch and optimal 2D branches.

The main contributions of this paper are summarized as follows:

- We propose MFINet, a real-time semantic segmentation network based on multi-view fusion and 2D-3D interactive enhancement, which adopts a three-branch architecture comprising representations of 3D Point View (3D-PV), 2D Bird's Eye View (2D-BEV) and Range View (2D-RV).
- We design a novel framework incorporating a 3D Point Feature Projector (3DPFP), Feature Enhancement module (FE), 2D-3D Fusion Head (FH), and Multi-Scale Dilated Attention (MSDA), which significantly enhances the performance of LiDAR semantic segmentation through 2D-3D interactive feature enhancement. Meanwhile, the 3D-PV branch maintains maximal simplicity, thereby ensuring real-time efficiency.
- We conducted extensive experiments on the SemanticKITTI and NuScenes benchmarks, demonstrating that MFINet outperforms existing methods on the SemanticKITTI and NuScenes validation sets and achieves competitive performance on the NuScenes test set with faster inference speed.

Related Works

Point-based methods take raw point clouds as input, offering excellent geometric fidelity. Representative works include PointNet (Qi et al. 2017a), which introduces a shared MLP and global max pooling for point-wise encoding; PointNet++ (Qi et al. 2017b), which adopts a hierarchical structure to capture local and multi-scale contextual features; KPConv (Thomas et al. 2019), which performs continuous deformable convolution in Euclidean space; Point Transformer (Zhao et al. 2021), which leverages self-attention to model long-range dependencies; PAConv (Xu et al. 2021), which dynamically generates convolution kernels based on point-wise positional relationships to handle irregular distributions; SuperLiDAR (Hui et al. 2023), which learns "super-point" representations via Euclidean distance-based aggregation; and PCB-RandNet (Han et al. 2024) introduces a cylindrical partitioning strategy in the polar coordinate system for balanced sampling across distance ranges, enhancing distant object recognition and improving segmentation accuracy with efficient computation.

Voxel-based methods discretize point clouds into regular grids, enabling efficient modeling with 3D convolutional networks. Cylinder3D (Zhu et al. 2021) adopts cylindrical partitioning and asymmetric 3D convolutions to balance point distribution and enhance structure representation, while introducing a point-wise refinement module to mitigate semantic loss from voxel label encoding. PVKD (Hou et al. 2022) combines point-voxel dual-branch outputs with knowledge distillation and leverages supervoxel partitioning and difficulty-aware sampling to improve sparsity handling and semantic supervision. SphereFormer (Lai et al. 2023) proposes a radial-window self-attention mechanism that divides 3D space into elongated regions, facilitating fusion between sparse distant and dense nearby points. 3D-ARSS (Wang et al. 2023) integrates spatial and channel attention modules and adopts sparse tensors to reduce memory and computation, enabling plug-and-play structural enhancement. SFPNet (Wang et al. 2024) presents a voxel-based framework that builds a sparse 3D grid via sparse voxelization and adopts an SSCN-based encoder–decoder to balance receptive field and boundary detail.

Projection-based methods map sparse 3D point clouds into dense 2D pseudo-images, enabling efficient modeling with mature 2D semantic segmentation architectures and offering fast deployment and inference. SqueezeSeg (Wu et al. 2018) pioneered spherical projection with fully convolutional networks for per-pixel prediction. Subsequent works address geometric distortion and information loss: GFNet (Qiu, Yu, and Tao 2022) introduces a geometric flow module for alignment and bidirectional information fusion between views; CENet (Cheng, Han, and Xiao 2022) employs large kernels and auxiliary supervision to improve nonlinear modeling and efficiency. Recently, Transformer-based architectures have been increasingly adopted: RangeFormer (Kong et al. 2023) uses a Transformer framework to capture global context in RV images, addressing issues such as semantic inconsistency and shape distortion from many-to-one projection; TransRVNet (Cheng, Han, and Xiao 2023) combines Transformer and CNN to enhance multi-scale fusion; RangeViT (Ando et al. 2023) applies Vision Transformer to further boost performance. WaffleIron (Puy, Boulch, and Marlet 2023) proposes a lightweight framework based on MLPs and dense 2D convolutions.

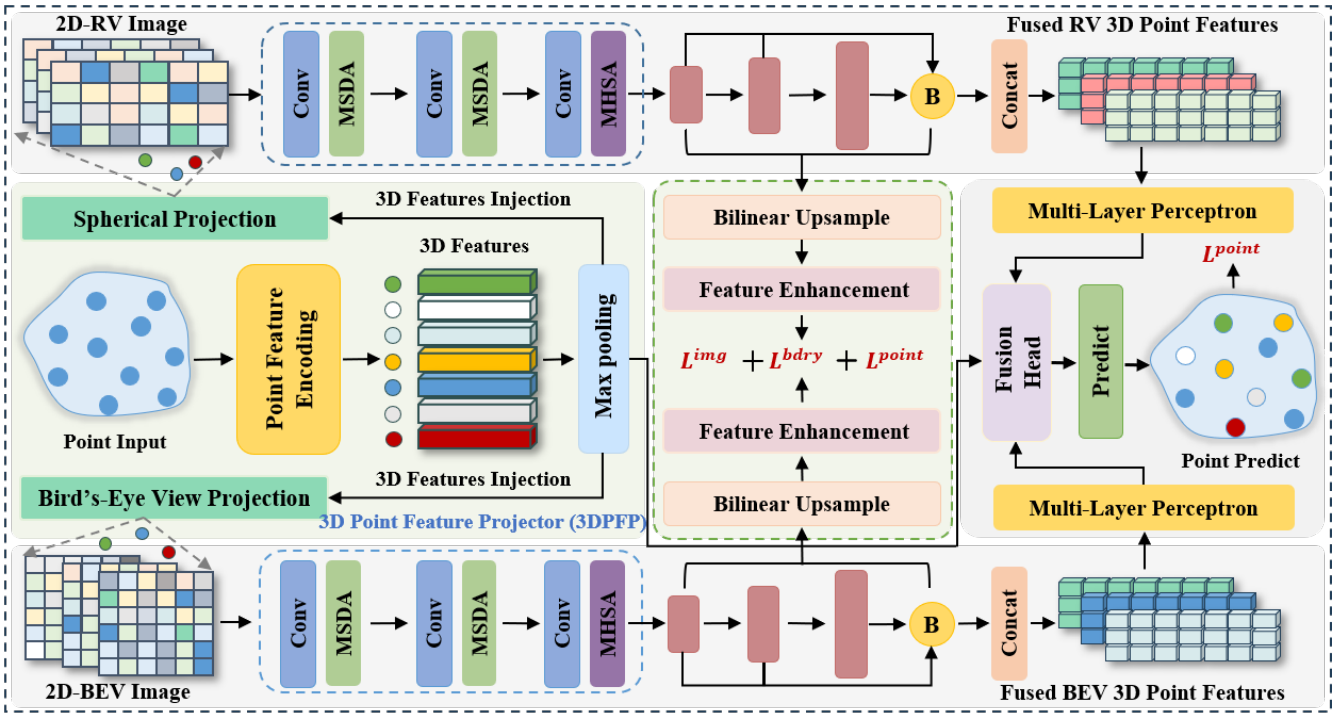

Figure 2: Overview of MFINet, which employs a three-branch architecture based on 3D Point View (3D-PV), 2D Bird's Eye View (2D-BEV), and Range View (2D-RV). Its core components include the 3D Point Feature Projector (3DPFP), Multi-Scale Dilated Attention (MSDA), 2D-3D Feature Enhancement (FE), and 2D-3D Fusion Head (FH).

Methodology

The overall architecture of MFINet is shown in Fig. 2. The 3D Point Feature Projector (3DPFP) extracts 3D information and injects it into the 2D-RV and 2D-BEV branches by projection to construct pseudo-images. In the shallow layers of the 2D branches, the Multi-Scale Dilated Attention (MSDA) is used to efficiently model sparse features and enhance discrimination. During decoding, the 2D-3D Feature Enhancement (FE) is used to extract geometric and semantic features. The 2D-3D Fusion Head (FH) aggregates multi-view point-level features for semantic prediction. Details are described in the following subsections.

2D Projection

2D-RV Image. Spherical projection transforms the unstructured 3D point cloud from the cartesian coordinate system (x, y, z) into a structured 2D image coordinate system (u, v) to generate a pseudo-image, referred to as the RV. Specifically, each point is projected based on its azimuth angle $\theta = \arctan(y/x)$, elevation angle $\phi = \arctan\left(z/\sqrt{x^2+y^2+z^2}\right)$, and depth $d = \sqrt{x^2+y^2+z^2}$, as follows:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left(1-\frac{\theta}{\pi}\right) W_{\text{rv}} \\ \left(1-\frac{\phi+f_{\text{up}}}{f_{\text{up}}+f_{\text{down}}}\right) H_{\text{rv}} \end{pmatrix} \tag{1}$$

where W_{rv} and H_{rv} represent the width and height of the RV pseudo-image, and f_{up} and f_{down} represent the upper and lower field of view angles of the LiDAR.

2D-BEV Image. The BEV projection maps the unstructured 3D point cloud onto the horizontal plane to generate a pseudo-image from the BEV perspective. Each point's plane coordinates (x, y) are mapped to pixel positions (u, v) in the BEV image, as calculated below:

$$(u, v) = \left(\frac{x - x_{\min}}{x_{\max} - x_{\min}} \cdot W_{\text{bev}}, \frac{y - y_{\min}}{y_{\max} - y_{\min}} \cdot H_{\text{bev}}\right) \tag{2}$$

where W_{bev} and H_{bev} represent the width and height of the BEV image, and $(x_{\min}, x_{\max})$ and $(y_{\min}, y_{\max})$ represent the range of the point cloud in the x and y directions.

3D Point Feature Projector (3DPFP)

Accurate semantic prediction relies on the preservation of fine-grained 3D details during the projection process. However, as a dimensionality reduction operation, projection inevitably leads to 3D spatial information loss. Furthermore, the unordered nature of point clouds leads to degradation of geometric information in projection. To alleviate this issue, we design a 3DPFP that extracts 3D features before projection, enhancing the geometric fidelity of the pseudo-images in both 2D-BEV and 2D-RV.

Specifically, given a set of 3D points $P = \{p_i\}_{i=1}^{N}$, where each $p_i \in \mathbb{R}^d$ denotes the 3D attributes (e.g., x, y, z, r, d), we apply multilayer perceptron (MLP) to obtain point-wise

3D features f_i. Since multiple 3D features may be projected to the same 2D position, we apply max pooling across all 3D features sharing the same pixel to obtain a unique descriptor per location. Then, based on the spatial locations $(u_i^{\text{2D-RV}}, v_i^{\text{2D-RV}})$ and $(u_i^{\text{2D-BEV}}, v_i^{\text{2D-BEV}})$, the pooled features are projected onto the 2D RV and BEV planes. The per-point features are embedded as follows:

$$I_{\text{2D-BEV/RV}} = \text{Proj}\left(\text{Pool}(\text{MLP}_3(P))\right) \tag{3}$$

where $\text{Proj}(\cdot)$ denotes spherical or BEV projection. The resulting pseudo-images enriched with 3D features are used as inputs to the 2D-BEV and 2D-RV branches.

Multi-Scale Dilated Attention (MSDA)

Convolutional neural networks excel at capturing local features, but their limited receptive fields restrict multi-scale global modeling. Traditional vision Transformers (Dosovitskiy et al. 2020) excel at modeling long-range dependencies between arbitrary image patches. However, for shallow 2D pseudo-image feature maps that exhibit sparsity and locality, the global attention mechanism not only incurs quadratic computational overhead, but also introduces redundant dependencies and noise, thereby undermining the model's efficiency and representation capability. To handle these issues, inspired by (Jiao et al. 2023), we introduce an MSDA module into the shallow encoders of both the 2D-BEV and 2D-RV branches, while deep stages adopt standard Multi-Head Self-Attention (MHSA). As shown in Fig. 3, MSDA enhances noise feature discrimination and efficiently models semantic contextual dependencies through a multi-scale sliding window strategy, achieving a good balance between representational power and computational efficiency. The process is as follows:

Given an input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, we obtain the corresponding queries, keys, and values by linear projection:

$$\mathbf{Q} = W_Q\mathbf{X}, \quad \mathbf{K} = W_K\mathbf{X}, \quad \mathbf{V} = W_V\mathbf{X} \tag{4}$$

where W_Q, W_K, W_V denote the projection matrices.

After that, we divide the channels of the feature map to n subspaces and perform Sliding Window Dilated Attention (SWDA) in different subspaces with different dilation rates. For each subspace i, we define the sparse receptive field centered at position p with dilation r_i and kernel size $k \times k$ as:

$$\mathcal{N}_i(p) = \{j \in \Omega \,|\, j \in \text{Window}(p, r_i, k)\} \tag{5}$$

where Ω denotes valid positions in the feature map.

For each subspace i, we compute the attention output using SWDA mechanism:

$$\mathbf{S}_i(p) = \sum_{j\in\mathcal{N}_i(p)} \text{Softmax}_j\left(\frac{\mathbf{Q}_i(p)\cdot\mathbf{K}_i(j)}{\sqrt{d}}\right)\mathbf{V}_i(j) \tag{6}$$

Finally, outputs from all subspaces are concatenated and linearly projected back to the original channel dimension:

$$\mathbf{Z} = \text{Concat}\left(\mathbf{S}_1(\mathbf{X}_1), \ldots, \mathbf{S}_n(\mathbf{X}_n)\right) \tag{7}$$

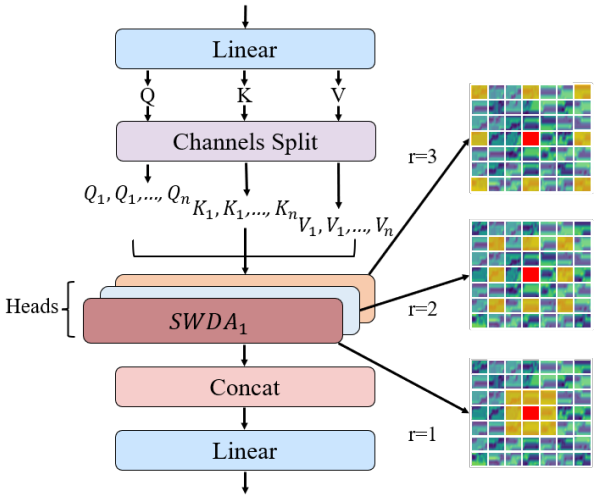

Figure 3: Multi-Scale Dilated Attention(MSDA).By default, we use a 3 × 3 kernel size with dilation rates r = 1, 2 and 3, and the sizes of attended receptive fields in different heads are 3 × 3, 5 × 5 and 7 × 7.

$$\text{MSDA}(\mathbf{X}) = \text{Linear}(\mathbf{Z}) \tag{8}$$

where $\text{Concat}(\cdot)$ concatenates attention outputs along the channel dimension to form $\mathbf{Z}$; $\text{Linear}(\cdot)$ is a learnable projection mapping $\mathbf{Z}$ back to the original channel dimension.

2D-3D Feature Enhancement (FE)

To fully exploit 2D information for geometric and semantic feature extraction, we introduce FE into the decoders of the 2D-RV and 2D-BEV branches. The FE jointly optimizes image-level semantics, boundary details, and point-level features. As shown in Fig. 4, the FE is activated only during the training phase.

In the decoders of the 2D-RV and 2D-BEV branches, the intermediate feature maps $\{F^i\}_{i=1}^N$ from multiple scales are first upsampled to match the original input resolution and then fed in parallel into the FE. The detailed procedure is as follows: First, a segmentation head is applied to each upsampled feature map F_{up}^{i} to generate a semantic segmentation map $\hat{S}_i$. Then, based on $\hat{S}_i$, multi-scale boundary detail maps are extracted using 3×3 Laplacian convolutions with 1, 2, 4 strides. These detail maps are upsampled to the original resolution and fused via a learnable 1×1 convolution. A threshold of 0.1 is applied to binarize the fused map, resulting in a binary detail map $\hat{B}_i$ that captures both boundary and corner information. Meanwhile, each feature map F_{up}^{i} is back-projected into the 3D space to obtain the corresponding point-wise semantic prediction $\hat{P}_i$.

Finally, we compute the losses for the 2D semantic segmentation, boundary detail map, and 3D point-wise semantic prediction, and formulate the overall loss as:

$$\mathcal{L}_i^{\text{FE}} = \lambda_1\mathcal{L}_i^{\text{img}} + \lambda_2\mathcal{L}_i^{\text{bdry}} + \lambda_3\mathcal{L}_i^{\text{point}} \tag{9}$$

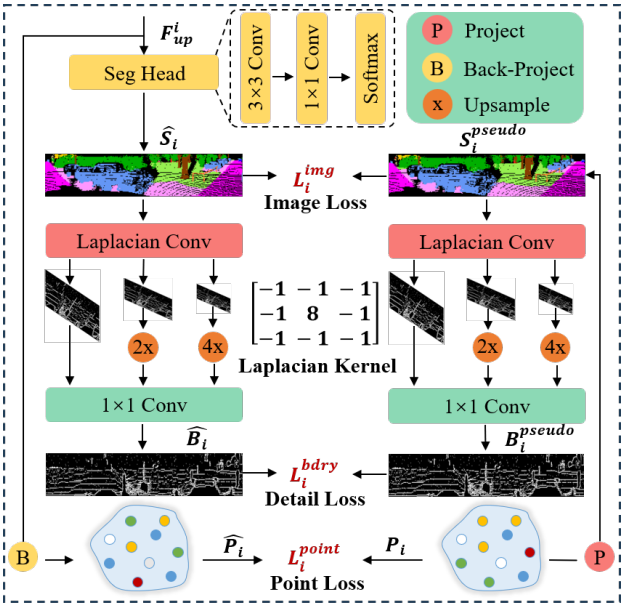

Figure 4: 2D-3D Feature Enhancement (FE). Achieves joint enhancement of image-level, boundary-level, and point-level features.

where $\mathcal{L}_i^{\text{img}}$ denotes the 2D segmentation loss combining weighted cross-entropy (WCE) and Lovasz-Softmax. $\mathcal{L}_i^{\text{bdry}}$ is a binary cross-entropy (BCE) loss to guide the network to learn details of features such as boundaries and corners. $\mathcal{L}_i^{\text{point}}$ denotes the point-wise loss, also based on WCE. The weights are set as $\lambda_1 = 0.4$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.4$.

2D-3D Fusion Head (FH)

We observe that fusing the final outputs of the 3D-PV, 2D-BEV, and 2D-RV branches yields satisfactory results but is not optimal. Features from different semantic levels are complementary: shallow features emphasize 3D structures, while deeper features are richer in semantic context. To exploit these multi-level features effectively, we design an FH to integrate representations from the 3D-PV, 2D-BEV, and 2D-RV branches at different semantic levels. Specifically, we first back-project the multi-scale 2D feature maps $F^l_{\text{2D-BEV/RV}}$ from the decoders of the 2D-BEV and 2D-RV branches into 3D space and concatenate them. Then, the fused and compressed features $F^{\text{point}}_{\text{2D-BEV/RV}}$ are integrated with the 3D features F_{point} extracted by the 3DPFP. The overall operation can be formulated as:

$$F^{\text{point}}_{\text{2D-BEV/RV}} = \text{MLP}\left(\text{Concat}\left(\prod_{l=1}^{3} F^l_{\text{2D-BEV/RV}}\right)\right) \tag{10}$$

$$F_{\text{fused}} = \text{Concat}\left(F^{\text{point}}_{\text{2D-BEV}}, F^{\text{point}}_{\text{2D-RV}}, F_{\text{point}}\right) \tag{11}$$

where l indexes the feature maps from three different levels of the decoder, $\Pi(\cdot)$ denotes the back-projection, and F_{fused} is used to generate the final semantic scores with a linear head for the point over the entire point cloud.

The above fusion process fully exploits the complementarity among features from different views and semantic levels. The 3D-PV, 2D-BEV, and 2D-RV branches provide diverse spatial attentions from distinct perspectives, yielding a more comprehensive semantic representation. Hierarchical fusion progressively enhances features—from geometric details to semantic abstraction, and from local to global understanding—thereby improving prediction accuracy.

Loss Function

The overall loss function consists of a main loss and an auxiliary loss, formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \mathcal{L}_{\text{aux}} \tag{12}$$

where $\mathcal{L}_{\text{main}}$ is the point-wise semantic prediction loss from the 2D-3D FH, computed using cross-entropy, and $\mathcal{L}_{\text{aux}}$ is the auxiliary loss from the 2D-3D FE.

Methods	**mean-IoU**	**FPS (Hz)**	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
RangeNet++	52.8	12	91.0	25.0	47.1	40.7	25.5	45.2	62.9	0.0	93.8	46.5	81.9	0.2	85.8	54.2	84.2	52.9	72.7	53.2	40.0
PolarNet	53.6	16	91.5	30.7	38.8	46.4	24.0	54.1	62.2	0.0	92.4	47.1	78.0	**1.8**	89.1	45.5	85.4	59.6	72.3	58.1	42.2
SalsaNext	59.4	24	90.5	44.6	49.6	86.3	54.6	74.0	81.4	0.0	93.4	40.6	69.1	0.0	84.6	53.0	83.6	64.3	64.2	54.4	39.8
TransRVNet	58.1	26	94.5	44.2	48.1	81.7	34.8	49.8	73.7	0.0	94.8	48.2	81.3	0.7	83.2	51.6	85.4	59.0	73.2	62.7	37.2
PCB-RandNet	56.5	-	94.5	6.8	35.5	76.7	37.6	48.3	79.0	0.0	93.4	42.4	80.8	1.1	89.0	56.9	87.6	68.4	74.4	61.0	40.8
SPVNAS	61.9	3.2	96.2	20.1	60.4	79.7	59.0	65.3	80.6	0.0	93.3	46.9	80.3	0.0	90.6	62.3	88.1	66.1	74.6	63.9	48.3
Cylinder3D	63.2	6.2	96.5	41.6	65.8	87.7	55.9	71.7	89.3	0.0	93.5	37.1	78.3	1.1	88.4	49.2	87.9	66.8	72.3	64.9	52.4
MinkowsikiNet	63.1	3.4	96.8	20.4	63.6	85.5	63.0	68.6	84.1	0.0	93.5	47.3	81.1	0.0	90.6	61.7	88.0	66.5	74.1	64.6	49.5
GFNet	63.0	10	94.2	49.7	63.2	74.9	32.1	69.3	83.2	0.0	**95.7**	53.8	**83.8**	0.2	**91.2**	62.9	88.5	66.1	76.2	64.1	48.3
AMVNet	65.2	11.8	95.6	48.8	65.4	88.7	54.8	70.8	86.2	0.0	95.5	**53.9**	83.2	0.3	90.9	62.1	87.9	66.8	74.2	64.7	49.3
PC-BEV	66.4	-	94.3	**67.6**	71.4	75.4	67.5	**77.4**	81.3	0.1	95.2	39.7	83.0	0.1	86.3	49.7	85.1	**70.2**	68.2	**70.2**	**64.2**
MFINet(Ours)	**68.1**	**28.1**	**96.9**	56.0	**79.0**	**92.4**	**68.0**	75.5	**94.4**	0.0	93.6	49.6	81.3	0.0	91.0	**63.8**	**89.1**	64.3	**77.2**	64.7	49.6

Table 1: Quantitative results on the SemanticKITTI val set, with the best and second-best bolded and underlined.

Methods	**mean-IoU**	Barrier	Bicycle	Bus	Car	Construction	Motorcycle	Pedestrian	Traffic-cone	Trailer	Truck	Driveable	Other-flat	Sidewalk	Terrain	Manmade	Vegetation
PolarNet	71.0	74.7	28.2	85.3	90.9	35.1	77.5	71.3	58.8	57.4	76.1	96.5	71.1	74.7	74.0	87.3	85.7
Salsanext	72.2	74.8	34.1	85.9	88.4	42.2	72.4	72.2	63.1	61.3	76.5	96.0	70.8	71.2	71.5	86.7	84.4
PCSCNet	72.0	73.3	42.2	87.8	86.1	44.9	82.2	76.1	62.9	49.3	49.3	95.2	66.9	69.5	72.3	83.7	82.5
Cylinder3D	76.1	76.8	40.3	91.3	93.8	51.3	78.0	78.9	64.9	62.1	84.4	96.8	71.6	76.4	75.4	90.5	87.4
SVASeg	74.7	73.1	44.5	88.4	86.6	48.2	80.5	77.7	65.6	57.5	82.1	96.5	70.5	74.7	74.6	87.3	86.9
PVKD	76.0	76.2	40.0	90.2	**94.0**	50.9	77.4	78.8	64.7	62.0	84.1	96.6	71.4	76.4	**76.3**	90.3	86.9
AMVNet	76.1	**79.8**	32.4	82.2	86.4	62.5	81.9	75.3	**72.3**	**83.5**	65.1	**97.4**	67.0	**78.8**	74.6	90.8	87.9
RangeViT	75.2	75.5	40.7	88.3	90.1	49.3	79.3	77.2	66.3	65.2	80.0	96.4	71.4	73.8	73.8	89.9	87.2
RPVNet	77.6	78.2	43.4	92.7	93.2	49.0	85.7	80.5	66.0	66.9	84.0	96.9	73.5	75.9	76.0	90.6	88.9
PRNet	78.0	78.0	40.9	92.5	93.4	54.1	85.4	80.6	63.2	69.8	84.7	97.3	75.1	77.7	76.1	91.0	89.3
RangeFormer	78.1	78.0	45.2	94.0	92.9	58.7	83.9	77.9	69.1	63.7	85.6	96.7	74.5	75.1	75.3	89.1	87.5
PC-BEV	78.8	78.2	46.3	92.5	93.4	55.0	87.1	81.0	65.4	69.2	85.7	97.1	**76.8**	77.0	**76.3**	90.6	88.5
SDSeg3D	78.7	78.2	52.8	94.5	93.1	54.5	88.1	82.2	69.4	67.3	86.6	96.4	74.5	75.2	75.3	87.1	84.1
WaffleIron	79.1	**79.8**	**53.8**	94.3	87.6	49.6	**89.1**	**83.8**	70.6	72.7	84.9	97.1	75.8	76.5	75.9	87.8	86.3
SphereFormer	79.5	78.7	46.7	95.2	93.7	54.0	88.9	81.1	68.0	74.2	86.2	97.2	74.3	76.3	75.8	**91.4**	**89.7**
SFPNet	80.1	78.8	49.7	**95.3**	93.5	**63.1**	86.4	82.9	68.6	72.8	**86.7**	97.0	74.7	76.0	75.3	91.2	89.5
MFINet(Ours)	**80.3**	79.5	49.2	93.9	93.9	58.6	87.0	82.5	68.9	79.8	85.9	97.2	75.0	77.5	76.0	91.2	89.3

Table 2: Quantitative results on the NuScenes val set, with the best and second-best bolded and underlined.

Experiment

Dataset

SemanticKITTI (Behley et al. 2019) is a large-scale LiDAR point cloud dataset from the University of Bonn, Germany, comprising over 43,000 scans with 360° coverage and 19 semantic categories. Sequences 00–10 are for training (with 08 as validation), and 11–21 for testing. NuScenes (Caesar et al. 2020) is a LiDAR semantic segmentation dataset comprising 1,000 scenes collected from diverse urban areas in Boston and Singapore. The dataset contains 28,130 training samples, 6,019 validation samples, and 6,008 test samples.

Implementation Details

MFINet is implemented using the PyTorch framework. The model is trained using an NVIDIA RTX 4090 GPU, and for fairness, evaluated on a single NVIDIA RTX 2080Ti GPU. The model is trained with the Adam optimizer using an initial learning rate of 0.006 and a momentum of 0.9. A OneCycle learning rate policy is adopted to dynamically adjust the learning rate. The input pseudo-image resolutions are set to 64×2048 for the 2D-RV and 600×600 for the 2D-BEV.

Results and Discussion

Table 1 presents the experimental results on the SemanticKITTI val set. MFINet achieves the highest mIoU of 68.1% while maintaining a real-time inference speed of 28.1 FPS. This validates the effectiveness of our proposed multi-view fusion, 2D-3D interactive enhancement strategy, and lightweight architecture in balancing accuracy and efficiency. Table 2 reports the performance comparison in the NuScenes val set. MFINet achieves the highest mIoU of 80.3% among all methods, outperforming AMVNet (Liong et al. 2020), RangeFormer (Kong et al. 2023), and PC-BEV (Qiu et al. 2025) by 4.2%, 2.2%, and 1.5%, further validating its robustness and broad applicability. Table 3 presents the results on the NuScenes test set. MFINet achieves an mIoU of 79.7%, surpassing all compared methods except RangeFormer (Kong et al. 2023). It is worth noting that all the compared methods mentioned above are LiDAR-only approaches.

As shown in Table 4, MFINet achieves competitive performance on both SemanticKITTI val and NuScenes val/test using only LiDAR input. While multi-modal methods LCPs (Zhang et al. 2023), EPMF (Tan et al. 2024) and 2DPASS (Yan et al. 2022) rely on additional RGB information, our model attains comparable or better mIoU with significantly fewer parameters and FLOPs. Compared with state-of-the-art LiDAR-only model LSK3DNet (Feng et al. 2024), MFINet further reduces computational cost while maintaining strong accuracy. Meanwhile, we conducted FPS tests on the Nvidia Jetson TX2, where LSK3DNet, LSK3DNet-S(kernel size is 7), and MFINet achieved 1.2, 3.8, and 5.1 FPS. These results indicate that despite adopting a simple projection-based design, our method exhibits strong generalization capability and practical deployment value.

Ablation Studies

Table 5 presents an ablation study on the three branches of MFINet: 3D Point View (3D-PV), 2D Range View (2D-RV), and Bird's Eye View (2D-BEV), aiming to verify the effectiveness of the multi-view fusion architecture. We evaluated two dual-branch combinations, 3D-PV + 2D-RV and 3D-PV + 2D-BEV, and compared them against the full three-branch configuration. The results show that the three-branch fusion significantly outperforms all dual-branch combinations in

Methods	**mean-IoU**	Barrier	Bicycle	Bus	Car	Construction	Motorcycle	Pedestrian	Traffic-cone	Trailer	Truck	Driveable	Other-flat	Sidewalk	Terrain	Manmade	Vegetation
PolarNet	69.4	72.2	16.8	77.0	86.5	51.1	69.7	64.8	54.1	69.7	63.4	96.6	67.1	77.7	72.1	87.1	84.4
JS3C-Net	73.6	80.1	26.2	87.8	84.5	55.2	72.6	71.3	66.3	76.8	71.2	96.8	64.5	76.9	74.1	87.5	86.1
AMVNet	76.1	79.8	32.4	82.2	86.4	62.5	81.9	75.3	72.3	83.5	65.1	97.4	67.0	78.8	74.6	90.8	87.9
3D-ARSS	75.8	83.6	41.5	90.2	53.1	70.1	76.7	74.9	61.0	81.5	80.9	95.8	69.4	77.1	**78.5**	89.2	88.4
GFNet	76.1	81.1	31.6	76.0	90.5	60.2	80.7	75.3	71.8	82.5	65.1	97.8	67.0	80.4	76.2	91.8	88.9
SAT3D	76.7	74.3	43.9	91.3	**92.8**	53.0	77.8	**79.7**	64.8	63.3	**85.7**	96.5	**73.4**	74.7	77.3	89.9	88.0
Cylinder3D	77.2	82.8	29.8	84.3	89.4	63.0	79.3	77.2	73.4	84.6	69.1	97.7	70.2	80.3	75.5	90.4	87.6
SPVNAS	77.4	80.0	30.0	91.9	90.8	64.7	79.0	75.6	70.9	81.0	74.6	97.4	69.2	80.0	76.1	89.3	87.1
SphereFormer	78.1	81.5	39.7	**93.4**	87.5	66.4	75.7	77.2	70.6	85.6	73.6	97.6	64.8	79.8	75.0	**92.2**	**89.0**
(AF)-S3Net	78.3	78.9	**52.2**	89.9	84.2	**77.4**	74.3	77.3	72.0	83.9	73.8	97.1	66.5	77.5	74.0	87.7	86.8
RangeFormer	**80.1**	**85.6**	47.4	91.2	90.9	70.7	**84.7**	77.1	74.1	83.2	72.6	97.5	70.7	79.2	75.4	91.3	88.9
MFINet(Ours)	79.7	82.2	44.5	89.7	91.6	69.1	82.2	79.2	**75.1**	**86.6**	73.2	**97.9**	67.0	**80.8**	76.0	92.0	**89.0**

Table 3: Quantitative results on the NuScenes test set, with the best and second-best bolded and underlined.

Methods	Modality	Params	FLOPs	SK-V	NS-V	NS-T
LCPs	L+C	77.7	/	63.2	80.5	78.9
EPMF	L+C	34.2	418.0	65.9	80.6	79.2
2DPASS	L+C	45.6	/	69.3	79.4	80.8
LSK3DNet	L	28.8	763.6	70.2	80.1	79.6
LSK3DNet-S	L	13.6	249.4	68.1	78.3	76.9
MFINet(Ours)	L	10.1	151.9	68.1	80.3	79.7

Table 4: Comparison on SemanticKITTI val and NuScenes val/test with Params, FLOPs, and mIoU. L+C indicates multi-modal methods. L indicates LiDAR-only methods.

Row	3D-PV	2D-RV	2D-BEV	FPS	Params	FLOPs	mIoU
1	✓	✓	×	46.8	6.8	64.9	64.1
2	✓	×	✓	40.9	6.8	88.2	62.6
3	✓	✓	✓	28.1	10.1	151.9	68.1

Table 5: Ablation study on different view branch combinations on the SemanticKITTI validation set.

segmentation accuracy. This demonstrates the complementary nature of the three views in capturing 3D structure, contextual semantics, and spatial distribution, all of which play a crucial role in improving segmentation performance.

We further conduct ablation studies on the key components of MFINet to evaluate their individual contributions to overall model performance, as shown in Table 6. The model without 3DPFP, MSDA, FE and FH serves as the baseline upon which each component is introduced. Experimental results show that incorporating the 3DPFP brings a 0.6% improvement in mIoU, indicating that 3D feature injection enhances the geometric representation of pseudo-images. Adding the MSDA yields an additional 0.9% gain, demonstrating its effectiveness in modeling sparse features. Introducing the FE further improves mIoU by 1.5%, validating the effectiveness of jointly optimizing semantic information at the image, boundary, and point levels. The inclusion of the FH brings an additional 1.6% gain, further enhancing the integration of multi-level semantic features. Compared with the full MFINet, removing only the 3DPFP results in a 2% drop in accuracy, while removing only the MSDA leads to a 1.4% decrease, indicating that both modules not only contribute positively to the overall performance of the model but also enhance the effectiveness of the other components. These results confirm the effectiveness of each module and the scalability of MFINet's modular design in enhancing segmentation accuracy.

Row	3DPFP	MSDA	FE	FH	FPS	Params	FLOPs	mIoU
1	×	×	×	×	42.3	4.9	73.7	63.5
2	✓	×	×	×	32.9	8.4	126.3	64.1
3	✓	✓	×	×	30.9	9.4	141.3	65.0
4	✓	✓	✓	×	29.0	9.9	148.8	66.5
5	✓	✓	✓	✓	28.1	10.1	151.9	68.1
6	×	✓	✓	✓	34.3	6.6	114.5	66.1
7	✓	×	✓	✓	30.8	9.1	138.0	66.7

Table 6: Ablation study on the effectiveness of core components in MFINet.

Conclusion

This paper proposes MFINet, a real-time semantic segmentation network based on multi-view fusion and 2D-3D interactive enhancement. We design a novel framework incorporating a 3D Point Feature Projector (3DPFP), Feature Enhancement module (FE), 2D-3D Fusion Head (FH), and Multi-Scale Dilated Attention (MSDA), which significantly enhances the performance of LiDAR semantic segmentation through 2D-3D interactive feature enhancement. Meanwhile, the 3D-PV branch maintains maximal simplicity, thereby ensuring real-time efficiency. Extensive experiments on SemanticKITTI and NuScenes demonstrate that MFINet surpasses state-of-the-art methods. Moreover, ablation studies prove the effectiveness of each module.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2023YFF0615800), the National Natural Science Foundation of China (Nos. 62371013, 62461160309, or N_HKU705/24, 62503024), the Beijing Natural Science Foundation (No. L247007), and in part by the Beijing Postdoctoral Research Foundation (No. 2024-ZZ-23).

References

Ando, A.; Gidaris, S.; Bursuc, A.; Puy, G.; Boulch, A.; and Marlet, R. 2023. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5240–5250.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9297–9307.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Cheng, H.-X.; Han, X.-F.; and Xiao, G.-Q. 2022. Cenet: Toward concise and efficient lidar semantic segmentation for autonomous driving. In *2022 IEEE international conference on multimedia and expo (ICME)*, 01–06. IEEE.

Cheng, H.-X.; Han, X.-F.; and Xiao, G.-Q. 2023. TransRVNet: LiDAR semantic segmentation with transformer. *IEEE Transactions on Intelligent Transportation Systems*, 24(6): 5895–5907.

Cheng, R.; Razani, R.; Taghavi, E.; Li, E.; and Liu, B. 2021. 2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12547–12556.

Cortinhal, T.; Tzelepis, G.; and Erdal Aksoy, E. 2020. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *International Symposium on Visual Computing*, 207–222. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng, T.; Wang, W.; Ma, F.; and Yang, Y. 2024. Lsk3dnet: Towards effective and efficient 3d perception with large sparse kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14916–14927.

Han, X.-F.; Cheng, H.; Jiang, H.; He, D.; and Xiao, G. 2024. Pcb-randnet: Rethinking random sampling for lidar semantic segmentation in autonomous driving scene. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 4435–4441. IEEE.

Hou, Y.; Zhu, X.; Ma, Y.; Loy, C. C.; and Li, Y. 2022. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8479–8488.

Hui, L.; Tang, L.; Dai, Y.; Xie, J.; and Yang, J. 2023. Efficient lidar point cloud oversegmentation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18003–18012.

Ibrahim, M.; Akhtar, N.; Anwar, S.; and Mian, A. 2023. SAT3D: Slot attention transformer for 3D point cloud semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 24(5): 5456–5466.

Jiao, J.; Tang, Y.-M.; Lin, K.-Y.; Gao, Y.; Ma, A. J.; Wang, Y.; and Zheng, W.-S. 2023. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE transactions on multimedia*, 25: 8906–8919.

Kong, L.; Liu, Y.; Chen, R.; Ma, Y.; Zhu, X.; Li, Y.; Hou, Y.; Qiao, Y.; and Liu, Z. 2023. Rethinking range view representation for lidar segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 228–240.

Lai, X.; Chen, Y.; Lu, F.; Liu, J.; and Jia, J. 2023. Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17545–17555.

Li, J.; Dai, H.; and Ding, Y. 2022. Self-distillation for robust LiDAR semantic segmentation in autonomous driving. In *European conference on computer vision*, 659–676. Springer.

Li, X.; Zhang, G.; Jiang, T.; Cai, X.; and Wang, Z. 2022. PRNet: Point-Range Fusion Network for Real-Time LiDAR Semantic Segmentation. In *IJCAI*, 1116–1122.

Liong, V. E.; Nguyen, T. N. T.; Widjaja, S.; Sharma, D.; and Chong, Z. J. 2020. Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. *arXiv preprint arXiv:2012.04934*.

Puy, G.; Boulch, A.; and Marlet, R. 2023. Using a waffle iron for automotive point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3379–3389.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

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

Qiu, H.; Yu, B.; and Tao, D. 2022. Gfnet: Geometric flow network for 3d point cloud semantic segmentation. *arXiv preprint arXiv:2207.02605*.

Qiu, S.; Li, X.; Xue, X.; and Pu, J. 2025. PC-BEV: An Efficient Polar-Cartesian BEV Fusion Framework for LiDAR Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6612–6620.

Tan, M.; Zhuang, Z.; Chen, S.; Li, R.; Jia, K.; Wang, Q.; and Li, Y. 2024. EPMF: Efficient perception-aware multi-sensor fusion for 3D semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 8258–8273.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6411–6420.

Wang, F.; Wu, Z.; Yang, Y.; Li, W.; Liu, Y.; and Zhuang, Y. 2023. Real-time semantic segmentation of LiDAR point clouds on edge devices for unmanned systems. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–11.

Wang, Y.; Zhao, W.; Cao, C.; Deng, T.; Wang, J.; and Chen, W. 2024. Sfpnet: Sparse focal point network for semantic segmentation on general lidar point clouds. In *European Conference on Computer Vision*, 403–421. Springer.

Wu, B.; Wan, A.; Yue, X.; and Keutzer, K. 2018. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *2018 IEEE international conference on robotics and automation (ICRA)*, 1887–1893. IEEE.

Xu, M.; Ding, R.; Zhao, H.; and Qi, X. 2021. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3173–3182.

Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; and Cui, S. 2021. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 3101–3109.

Yan, X.; Gao, J.; Zheng, C.; Zheng, C.; Zhang, R.; Cui, S.; and Li, Z. 2022. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European conference on computer vision*, 677–695. Springer.

Zhang, Y.; Zhou, Z.; David, P.; Yue, X.; Xi, Z.; Gong, B.; and Foroosh, H. 2020. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9601–9610.

Zhang, Z.; Zhang, Z.; Yu, Q.; Yi, R.; Xie, Y.; and Ma, L. 2023. Lidar-camera panoptic segmentation via geometry-consistent and semantic-aware alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3662–3671.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16259–16268.

Zhao, Y.; Bai, L.; and Huang, X. 2021. Fidnet: Lidar point cloud semantic segmentation with fully interpolation decoding. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4453–4458. IEEE.

Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Ma, Y.; Li, W.; Li, H.; and Lin, D. 2021. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9939–9948.