information

MDPI

Article

PVI-Net: Point–Voxel–Image Fusion for Semantic Segmentation of Point Clouds in Large-Scale Autonomous Driving Scenarios

Zongshun Wang, Ce Li *, Jialin Ma, Zhiqiang Feng and Limei Xiao

School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China; 212085400049@lut.edu.cn (Z.W.); jialinm@lut.edu.cn (J.M.); jonathan_fzq@163.com (Z.F.); xlm@lut.edu.cn (L.X.)

* Correspondence: xjtulice@gmail.com

Abstract: *In this study, we introduce a novel framework for the semantic segmentation of point clouds in autonomous driving scenarios, termed PVI-Net.* This framework uniquely integrates three different data perspectives—point clouds, voxels, and distance maps—executing feature extraction through three parallel branches. Throughout this process, we ingeniously design a point cloud–voxel cross-attention mechanism and a multi-perspective feature fusion strategy for point images. These strategies facilitate information interaction across different feature dimensions of perspectives, thereby optimizing the fusion of information from various viewpoints and significantly enhancing the overall performance of the model. The network employs a U-Net structure and residual connections, effectively merging and encoding information to improve the precision and efficiency of semantic segmentation. We validated the performance of PVI-Net on the SemanticKITTI and nuScenes datasets. The results demonstrate that PVI-Net surpasses most of the previous methods in various performance metrics.

Keywords: semantic segmentation; multi-perspective; cross-attention; LiDAR point clouds

check for updates

Citation: Wang, Z.; Li, C.; Ma, J.; Feng, Z.; Xiao, L. PVI-Net: Point–Voxel–Image Fusion for Semantic Segmentation of Point Clouds in Large-Scale Autonomous Driving Scenarios. *Information* **2024**, *15*, 148. https://doi.org/10.3390/info15030148

Academic Editors: Marco Leo and Gholamreza Anbarjafari

Received: 26 January 2024
Revised: 27 February 2024
Accepted: 6 March 2024
Published: 7 March 2024

1. Introduction

In recent years, with the rapid development of artificial intelligence technology, 3D point cloud processing has become an important branch in the field of computer vision. Especially in outdoor scenes, such as autonomous driving, urban planning, and *Geographic Information Systems (GISs),* LiDAR point cloud segmentation technology plays a crucial role. For autonomous vehicles, accurate point cloud segmentation is key to safe navigation and decision making. Due to the working principle of LiDAR sensors, the collected point cloud data may have uneven density and occlusion issues. These characteristics make extracting accurate and reliable semantic information from these data a challenging task.

Recent advancements in point cloud semantic segmentation have substantially contributed to the field, particularly within large-scale autonomous driving scenarios [1–4]. *These advancements predominantly revolve around the effective processing and analytical representation of voluminous point cloud data, captured through LiDAR technology. Our work introduces a novel conceptualization within this domain, where a single point cloud dataset is represented through three distinct but complementary perspectives: point-based, voxel-based, and distance map representations. This unique approach aims to enhance the model's feature extraction capabilities by leveraging the intrinsic advantages of each representation method, thereby enriching the semantic segmentation process.* Among these, voxel-based methods convert point clouds into three-dimensional grids and use 3D convolutional neural networks for processing, *which* is convenient for capturing spatial information but requires high resolution when dealing with sparse point clouds, increasing computational and storage burdens. Direct point-based methods retain the precision of the original structure but are computationally inefficient when dealing with unstructured data, while image-based methods accelerate processing, but they may lose three-dimensional spatial information when projecting point clouds into two-dimensional images, affecting segmentation accuracy.

Therefore, we found that, in building models for large scene point cloud segmentation, the fusion of point cloud, voxel, and distance map perspectives is not just a simple data overlay but a multi-dimensional information fusion strategy. Point clouds, as a high-fidelity representation of raw data, maintain the original precision of spatial information and the integrity of microscopic details, directly reflecting the depth perception of scenes. Voxelization, though introducing some quantization errors, provides an intuitive and operable geometric expression for the macro form and volumetric characteristics of the data. Distance maps, as an advanced representation of the spatial relationships in point clouds, provide a key perspective for understanding the geometric continuity and topological structure of scenes by encoding the spatial distances between points. This multi-dimensional data representation strategy lays the foundation for in-depth analysis and accurate segmentation of large-scale point cloud scenes. A point cloud segmentation model that integrates different perspectives shows outstanding robustness and accuracy in processing complex and large-scale scenes. This fusion is not just a simple stacking of data but a deep integration of information.

In this study, we propose an adaptive point–voxel–distance map feature fusion framework, PVI-Net, to optimize the semantic segmentation of point clouds in outdoor scenes. This framework combines the advantages of point cloud, voxel, and distance map perspectives, providing a comprehensive perspective for processing complex large-scale data. PVI-Net uses a multi-layer feature extraction and fusion mechanism, combining multi-layer perceptron (*MLP*), 3D sparse convolution, and 2D convolution, implementing effective feature fusion and information encoding retention through a U-Net structure and residual connections, thereby improving the accuracy and efficiency of semantic segmentation. Specifically, the point cloud–voxel cross-attention mechanism and point–image multi-perspective feature fusion strategy effectively handle the structural differences and information fusion between different perspectives, enhancing the overall performance of the model. For computational efficiency, PVI-Net reduces the computational cost of multi-perspective fusion through optimization strategies. Voxelization processing quickly filters point clouds in the early stage of data processing, reducing the processing burden on high-density information, while the high-level spatial relationship expression provided by distance maps helps the model quickly identify scene features, reducing the need for point-by-point analysis of complex data. These strategies collectively contribute to effectively improving computational efficiency and resource management balance while maintaining high segmentation accuracy. *The experimental results show that PVI-Net performs excellently in processing point cloud data of complex outdoor scenes. The evaluation results on two key datasets, SemanticKITTI and nuScenes, show that PVI-Net performs excellently in terms of point cloud semantic segmentation accuracy in large-scale autonomous driving scenarios.*

Our work offers the following key contributions:

- Proposing PVI-Net, a semantic segmentation framework for large-scale point cloud scenes, which integrates three different data perspectives—point cloud, voxel, and distance map—achieving an adaptive multi-dimensional information fusion strategy.
- Designing point–voxel cross-attention and *Multi-perspective Fusion Attention (MF-Attention)* mechanisms in the network structure, effectively addressing the structural differences and information fusion issues between different perspectives.
- Designing a multi-perspective feature post-fusion module. This module can effectively combine features from point clouds, voxels, and distance maps. In the post-fusion stage, the model integrates information from different perspectives, enhancing semantic understanding of complex outdoor scenes.

2. Related Works

2.1. Point Processing in Point Cloud Segmentation

Point-based methods [5–8] are renowned for their ability to learn global features directly from raw point clouds. However, they fall short in capturing details and local structures within point clouds. To address this deficiency, multi-scale processing methods [9] have been proposed. Such methods enhance the understanding of complex structures by analyzing point cloud features at different scales. Nevertheless, these methods often increase computational burdens. *Graph-based methods* [10], on the other hand, have turned to a new processing strategy, transforming point cloud data into graph structures and utilizing graph neural networks to capture complex relationships between points. This approach is particularly suitable for processing unstructured point cloud data but faces high computational costs in graph construction and processing. Overall, in the field of large-scale autonomous driving point cloud processing, point cloud data are unstructured, meaning the data points are unordered, and the number of neighbors for each point can vary. This irregularity poses significant challenges to point cloud processing.

2.2. Voxel Processing in Point Cloud Segmentation

Voxel-based point cloud *segmentation* [11–13] has garnered widespread attention in the understanding of autonomous driving scenes. *Park et al.* [14] proposed an Efficient Point Cloud Transformer (EPT) based on local self-attention to understand large-scale 3D scenes. EPT, due to its voxel structure, offers faster inference speeds compared with point-based work. *Wang et al.* [15] introduced a Dynamic Sparse Voxel Transformer (DSVT), a Voxel Transformer backbone based on a single-step window for outdoor 3D perception. This method divides a series of local regions in each window according to sparsity and then computes the features of all regions in a fully parallel manner. Although these methods use sparse voxel grids to reduce memory occupancy and employ layered and multi-scale voxel representations to capture more details, the conversion of point cloud data into voxel format faces detail loss due to voxelization. Our proposed PVI-Net bridges this gap through multiple perspectives.

2.3. Range Image Processing in Point Cloud Segmentation

Recent advancements in point cloud segmentation have highlighted the potential of range images as a complementary representation to traditional point-based and voxel-based methods. Range images, derived from point clouds through spherical projection, maintain depth information in a structured, image-like format, facilitating the application of mature 2D image-processing techniques. The transformation of point clouds into range images involves projecting 3D points onto a 2D plane based on their azimuth and elevation angles relative to a specific viewpoint, typically the sensor origin. This process preserves the spatial locality and depth information, offering a compact representation that is particularly beneficial for capturing surface geometries and contours. Several notable studies have leveraged range images for enhancing point cloud analysis. For instance, RangeNet++ [16] employs a deep neural network to segment range images semantically, exploiting their structured nature for efficient processing. Similarly, SqueezeSeg [17] and its successors demonstrate the efficacy of convolutional neural networks in interpreting range images for tasks like semantic segmentation and object detection within point clouds. Despite their advantages, range images are not without challenges. The projection process can introduce distortions, particularly at large distances or near the edges of the field of view. Therefore, considering how to further narrow the gap between 2D image processing and 3D point cloud analysis is a potential objective.

2.4. Multi-Perspective Fusion

The advantages of multi-perspective *point cloud segmentation* [18–20] are primarily manifested in its ability to provide a more comprehensive spatial understanding than a single perspective. In multi-perspective point cloud segmentation, data from different angles are fused to form more complete three-dimensional representations of target objects or environments. Chen et al. [21] explored interactive fusion between point cloud and image data, using an autoencoder structure to enhance the performance of 3D object detection through simultaneously learning features of point clouds and images. Tang et al. [8] focused on finding efficient 3D architectures. They combined sparse point and voxel convolutions, aiming to create a network that is both efficient and accurate for processing point cloud data. These methods can significantly reduce the occlusions and blind spots caused by single perspectives, especially in complex environments. Compared with previous methods, our approach proposes a point–voxel–image tri-perspective point cloud semantic segmentation framework, which enables capturing more information about shape, size, and other important features from multiple angles.

3. Methodology

In this section, we provide a comprehensive introduction to the PVI-Net framework for point cloud processing in outdoor scene segmentation. In Section 3.1, we outline the overall structure of the network and data flow. Following this, in Section 3.2, we detail the input data sources and feature extraction processes of the network's three key branches. Further, in Section 3.3, we delve into the fusion methods of these three branches during the feature extraction stage and the key modules designed for the post-fusion stage. This chapter aims to offer an in-depth understanding of the details of the PVI-Net framework, showcasing its efficiency and innovation in processing complex outdoor scene point cloud data.

3.1. Overview

Figure 1 shows our newly developed PVI-Net network, a tri-branch feature fusion network for point cloud semantic segmentation. For the input point cloud data, we first map point cloud features into voxel grid features, providing input for the voxel feature learning branch. Then, point cloud data are transformed into range images through spherical projection, serving as input for the image feature learning branch. The point cloud branch employs a basic PointNet structure and several *MLPs* to generate multi-resolution features. The voxel and image branches utilize 3D sparse convolution and 2D convolution, respectively, and employ a U-Net structure for featuring the encoding and decoding of each branch, simultaneously achieving a fusion of features from three perspectives. Additionally, in the decoding stage, we apply residual connections to ensure that information learned during the encoding stage is effectively transferred to the output. Finally, using an innovative multi-perspective feature post-fusion module, we perform post-fusion of features from the three branches, accurately restoring the semantic information of each point cloud.

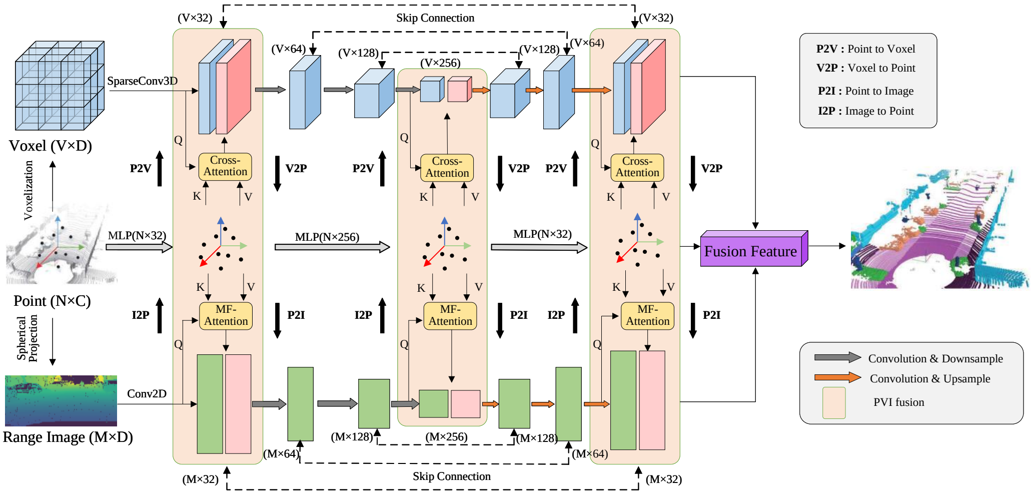

Figure 1. PVI-Net. Point cloud–voxel–image fusion point cloud semantic segmentation network structure diagram.

3.2. Tri-Branch Feature Learning

3.2.1. Point Cloud Feature Extraction Branch

In the point cloud branch of PVI-Net, given an unordered set of points $P = \{p_P^i\}_{i=1}^N$, where each point in the point cloud $p_P^i \in \mathbb{R}^C$ includes the coordinates $c_P^i = [x_i, y_i, z_i]$ and the point cloud features. The direct use of MLP to extract features in the point cloud branch helps to reduce the high computational load and memory consumption caused by searching for neighboring relationships, thereby enabling efficient processing of large-scale data and simplifying the network structure. Each point in the point cloud is individually processed with MLP, which effectively extracts and learns the features of each point, and can be represented as follows:

$$F_p^i = \begin{cases} MLP(P), & l = 1 \\ MLP(F_p^{l-1}) + F_p^{l-1}, & l > 1 \end{cases} \tag{1}$$

where l denotes the layer of the MLP, and F_p^i represents the features extracted via the MLP at layer l. The point cloud feature extraction involves processing each point in the point cloud individually. *MLP layers, including linear transformations and nonlinear activations,* allow the network to learn complex patterns in the data. This process is crucial for capturing the complex geometric details of the point cloud, and these features are subsequently integrated with the voxel and range image branches through the fusion process.

3.2.2. Voxel Feature Extraction Branch

For the input point cloud $P = p_{P\,i=1}^{i\ \,N}$, a three-dimensional voxel grid covering the entire range of the point cloud is first defined. This grid consists of many small cubes (voxels), each with a fixed size. *Furthermore, the point cloud data are mapped onto the three-dimensional voxel grid to obtain voxel features with a voxel resolution of $L_V \times H_V \times W_V$, denoted as $F_V \in \mathbb{R}^{L_V \times H_V \times W_V}$.* The voxel index for each point is calculated based on its coordinates in three-dimensional space. For a point $p_P^i(x, y, z)$ and a voxel grid in which each voxel's size is $\Delta x \times \Delta y \times \Delta z$, the voxel index (i, j, k) of point p_P^i can be calculated as follows:

$$\begin{cases} i = \left\lfloor \frac{x - x_{\min}}{\Delta x \times c_k} \right\rfloor \\ j = \left\lfloor \frac{y - y_{\min}}{\Delta y \times c_k} \right\rfloor \\ k = \left\lfloor \frac{z - z_{\min}}{\Delta z \times c_k} \right\rfloor \end{cases} \tag{2}$$

where $x_{\min}$, $y_{\min}$, and $z_{\min}$ are the minimum coordinate values of the voxel grid in each direction, $\lfloor \cdot \rfloor$ denotes the floor function, and c_k is the downsampling stride of the 3D CNN. This approach ensures that each point in the point cloud is allocated to a corresponding voxel, establishing a mutual correspondence between points and voxels, facilitating feature interaction between point cloud and voxels. To avoid the memory loss caused by empty voxels, we use 3D sparse convolution to downsample and encode voxel features:

$$\begin{cases} SConv3D(F_V), & l = 1 \\ SConv3D(F_V^{l-1}) + F_V^{l-1}, & l > 1 \end{cases} \tag{3}$$

where $SConv3D(\cdot)$ contains a 3D sparse convolution and an activation function, and F_V^l represents the voxel features extracted via 3D sparse convolution at layer l. We use 3D sparse convolution to downsample and encode voxel features, preserving feature maps of three downsampling voxel directions. The voxel features are then upsampled to restore voxel features.

3.2.3. Image Feature Extraction Branch

The method of converting point cloud data into range images is achieved through spherical projection, where the position of each point is mapped onto a two-dimensional plane. *Given a three-dimensional point cloud $P = p_{P\,i=1}^{i\ \,N}$ with coordinates (x_i, y_i, z_i) in the three-*

dimensional Cartesian coordinate system, the corresponding two-dimensional coordinates $[u_i, v_i]$ *of the two-dimensional image* $I \in \mathbb{R}^{H_I \times W_I \times C}$*, with height* H_I*, width* W_I*, and dimension C, through spherical projection, can be expressed as follows:*

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left[1 - \arctan(y_i, x_i)\pi^{-1}\right] W_I \\ \left[1 - \frac{\arcsin(z_i, d^{-1}) - R_d}{R}\right] H_I \end{pmatrix} \tag{4}$$

where $d = \sqrt{x_i^2 + y_i^2 + z_i^2}$ is the Euclidean distance from point P to the reference origin in the LiDAR coordinate system, as well as the straight-line distance to the projection center. R represents the vertical field of perspectives of the LiDAR sensor, and R_d is the lower boundary of the vertical field of perspectives. Spherical projection is a non-bijective process in which each point, p_i, in the point cloud maps to a pixel position in the projected image. However, due to the nature of this mapping, multiple three-dimensional points may correspond to the same pixel in the image, leading to a one-to-many mapping relationship.

In the image feature extraction branch, convolutional operations are used to extract features from the two-dimensional image obtained through spherical projection, which can be represented as follows:

$$F_I^l = \begin{cases} Conv(I), & l = 1 \\ Conv(F_I^{l-1}) + F_I^{l-1}, & l > 1 \end{cases} \tag{5}$$

where $Conv(\cdot)$ *contains a 2D convolution and an activation function.* F_I^l represents the image features extracted via 2D convolution at layer l, similarly preserving feature maps of three downsampling image directions for the encoding–decoding process.

3.3. Multi-Perspective Feature Fusion

In the previous section, we first introduced the projection system, establishing corresponding index systems between point–voxel–range and the feature extraction process of the three branches. In this section, we construct interactions between the representations based on points, voxels, and ranges.

The distinct characteristics and advantages of point clouds, voxels, and depth maps necessitate different fusion strategies, based on their properties and complementarity in fusion. Point cloud data are irregular, while voxels partition space into regular grids. This structural difference makes simple addition or concatenation fusion insufficient for capturing their complex relationships.

Therefore, we designed an adaptive point–voxel cross-attention feature interaction method to handle this irregularity and structural difference better. It computes the relationship between point cloud and voxel features, *enabling more flexible weighting of these features and a more effective combination of their information. As shown in Figure 2,*

$$f_{PV} = \sum_{k=1}^{K} MLP[((f_V + f_P^k) + \delta) \odot f_P^k] + f_V \tag{6}$$

where $MLP(\cdot)$ denotes a feature encoding function, $\odot$ represents element-wise multiplication, and δ is the positional encoding, *defined as follows:*

$$\delta = MLP\Big(\text{Concat}\Big[\sigma\Big(p_P^k - \mu_c\Big), \sigma\Big(p_P^k\Big)\Big]\Big) \tag{7}$$

where p_P^k is the 3D coordinates of a point P, $\mu_c = \frac{1}{K}\sum_{i=1}^{K} p_i$ is the mean of all projected point coordinates, σ is a nonlinear activation function, and $Concat(\cdot)$ denotes vector concatenation. This combines both relative and absolute position information, passed through nonlinear activation and then concatenated as input to the MLP, capturing the spatial relationships of points in both local and global contexts.

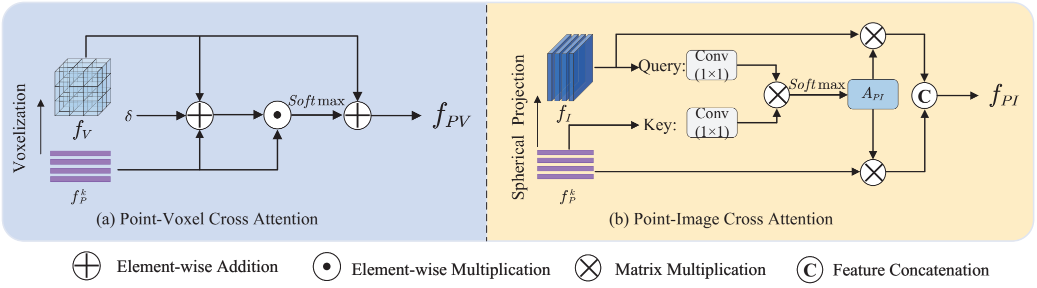

Figure 2. Multi-perspective feature deep fusion structure.

3.3.1. MF-Attention Feature Fusion Module

We process point cloud data, mapping them to a two-dimensional image. In this process, multiple points in the point cloud may map to the same pixel position in the two-dimensional image. To consider information comprehensively from different perspectives of points and images and dynamically balance their contributions, we designed an MF-Attention feature fusion module. Suppose a set of points, $p_{Pk=1}^{k\ K}$, in the point cloud maps to a pixel, P_I, in the two-dimensional image, then each point, p_P^k, in the set has a corresponding feature vector, f_P^k, and each pixel, P_I, also has a feature vector, f_I. The goal of MF-Attention fusion is to update point features, f_P^k, to reflect their relationship with the corresponding pixel feature, f_I. Firstly, we calculate the attention weights between point cloud features and image features:

$$A_{PI} = Softmax(\frac{f_I W_q \times (f_P^k W_k)^T}{\sqrt{d_k}}) \tag{8}$$

where W_I, W_P are learnable weight matrices for further transforming the mapped features into the attention computation space. *The dimension size of the key vectors is represented by d_k. Employing the scaling factor $\sqrt{d_k}$ aids in preserving the numerical stability within the attention mechanism.* Then, the final MF-Attention fusion feature is represented as:

$$f_{PI} = Concat((A_{PI} \times f_I), (A_{PI}^T \times f_P^k)) \tag{9}$$

where $Concat(\cdot)$ is used to concatenate features. The point–image attention fusion mechanism provides an effective way to synthesize and utilize information from point clouds and images, enabling the model to discover and leverage their inherent connections when processing multi-perspective data. This method is particularly useful in combining point cloud and image data for semantic prediction of point clouds.

3.3.2. Multi-Perspective Feature Post-Fusion Module

We extract features from each branch and design a deep fusion method for the features of the three branches to enhance the feature representation ability of each branch. As shown in Figure 3. *Furthermore,* we post-fuse the final prediction results of point cloud, voxel, and depth map features to provide a richer and more comprehensive feature representation for point cloud semantic segmentation tasks. For the final features obtained from the point cloud branch, $F_P \in \mathbb{R}^{N \times D}$, the voxel branch, $F_V \in \mathbb{R}^{L_V \times H_V \times W_V \times D}$, and the image branch, $F_I \in \mathbb{R}^{H_I \times W_I \times D}$, *the corresponding semantic segmentation pseudo-probabilities are represented as follows:*

$$\begin{cases} O_P = Softmax(MLP(F_p)) \\ E_V = Softmax(SConv3D(F_V)) \\ E_I = Softmax(Conv(F_I)) \end{cases} \tag{10}$$

where $O_P \in \mathbb{R}^{N\times T}$*, where T represents the number of semantic categories. For* $E_V \in \mathbb{R}^{L_V\times H_V\times W_V\times T}$ and $E_I \in \mathbb{R}^{H_I\times W_I\times T}$, they are mapped back to the original point cloud position according to the hash table built in the voxelization and spherical projection processes:

$$\begin{cases} E_V \to O_V \\ E_I \to O_I \end{cases} \tag{11}$$

where $O_V \in \mathbb{R}^{N\times T}$, $O_I \in \mathbb{R}^{N\times T}$*.* To associate global features, we weight the features of each branch globally, allowing the model to learn key features automatically in each perspective. The weighted features of each branch are represented as follows:

$$\begin{cases} G_P = MLP(g[MaxPool(F_P); AvgPool(F)]) \\ G_V = MLP(3DGAP(F_V)) \\ G_I = MLP(GAP(F_I)) \end{cases} \tag{12}$$

where $g(\cdot)$ *denotes* $(2,1)$ *linear mapping,* $3DGAP(\cdot)$ represents 3D global average pooling, and $GAP(\cdot)$ represents global average pooling. Thus, the final fusion result is represented as follows:

$$Y = G_P O_P + G_V O_V + G_I O_I \tag{13}$$

Fusing the features of point clouds, voxels, and depth maps utilizes each perspective's unique advantages to provide a more comprehensive and powerful data representation, thus achieving better performance in specific tasks.

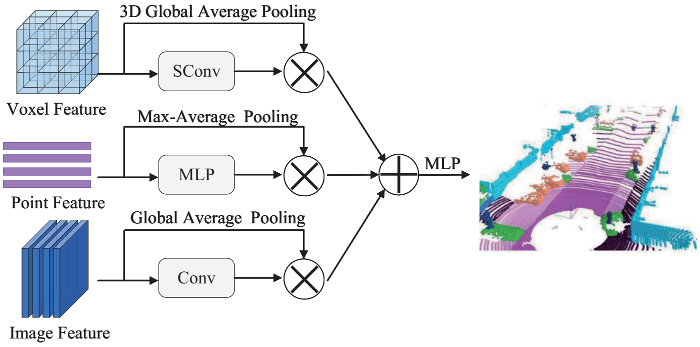

Figure 3. Multi-perspective feature fusion module.

4. Experiments

In this section, we extensively explore the PVI-Net network and its application in autonomous driving. In Section 4.1, we provide a thorough introduction to the two key datasets used in our experiments—SemanticKITTI and nuScenes—elucidating their importance in network testing and evaluation. Following this, in Section 4.2, we delve into the various components of the PVI-Net architecture, detailing the key aspects and experimental settings of the network to ensure transparency and reproducibility in our experiments. In this section, to intuitively understand the impact of various indicators on network performance, we use "↓" and "↑" to denote that smaller or larger values of the indicators, respectively, lead to better network performance. Finally, in Section 4.3, we conduct a comprehensive performance evaluation of the PVI-Net model. In addition, we perform a series of ablation experiments to verify the superiority and effectiveness of the model in its key constituent steps.

4.1. Datasets

SemanticKITTI. The SemanticKITTI dataset, an extension of the KITTI Vision Benchmark Suite, is a leading dataset in the fields of autonomous driving and robotics vision. Its key feature is the provision of a large-scale, time-sequenced LiDAR scanning dataset, comprising over 43.5 billion finely annotated point clouds distributed across more than

22,000 scene sequences, covering various road types and climatic conditions. The point clouds in the dataset are subdivided into 25 categories, with training and test sets composed of sequences from 00 to 10 and 11 to 21, respectively, to test and optimize their models, ensuring their effective operation in various environments and an accurate understanding of their surroundings.

nuScenes. The nuScenes dataset, released by Aptiv Autonomous Mobility, is a widely used multi-perspective dataset in the field of autonomous driving research. It was collected in diverse urban environments in Boston and Singapore, providing rich information on roads, traffic, and climate conditions. This dataset combines data from six cameras, five radars, and one LiDAR, achieving 360-degree comprehensive environmental capture, greatly facilitating an in-depth understanding of complex scenes and supporting tasks such as object detection, tracking, and segmentation. nuScenes includes over 1 million precise 3D bounding box annotations, covering 23 different object categories, totaling 40,000 frames of high-quality data. *These data are meticulously divided into 8130 training samples, 6019 validation samples, and 6008 test samples, ensuring extensive training and evaluation coverage.* Additionally, to enhance its applicability in real-world scenarios, the dataset specially optimized its category annotations, focusing on 16 primary categories for LiDAR semantic segmentation.

4.2. Implementation Details and Settings

Architecture Settings. As shown in Figure 1, we propose a multi-perspective point cloud segmentation network architecture. This architecture first converts point cloud data into quantized voxels with a high resolution of $1600 \times 1408 \times 40 \times 8$. At the core of voxel processing, the backbone network employs 3D sparse convolution, generating feature maps of voxel directions at four different scales with output dimensions of 32, 64, 128, and 256, respectively. Subsequently, these feature maps are restored by a decoder symmetrical to the dimensions of the encoder to recover voxel features. In our experiments, the resolution of voxels is set to a 5 cm edge length for each voxel. For image branch processing, when dealing with the SemanticKITTI dataset, the input range–image size is set to 64×2048. When handling the nuScenes dataset, the initial input range–image size of 32×2048 is later adjusted to 64×2048 to align with the dimensions of the SemanticKITTI dataset.

Training Strategies. In our experiments, we trained the model for 120 epochs using the Adam optimizer, with the initial learning rate set to 0.01. *This process was conducted on a system equipped with 4× RTX 3090 GPUs, with the batch size set to 4.* To prevent overfitting, we used data augmentation techniques, including GT-sampling technology and random flipping, rotation, and scaling, within the range of [0.95, 1.05]. During training, we also employed a cosine annealing strategy to adjust the learning rate and implemented global scaling and random rotation around the Z-axis as enhancement measures to increase data diversity and the model's generalization capability.

4.3. Results

4.3.1. Evaluation on SemanticKITTI Dataset

In our research, we conducted comprehensive experiments on the newly proposed PVI-Net network using the SemanticKITTI dataset and compared it with some of the latest advanced methods, as shown in Table 1. The results show that PVI-Net achieved a significant improvement of over 10% in the mean intersection over union (mIOU) metric compared with previous classic single-perspective input networks (such as point-based, voxel-based, and image-based methods). In comparison with mixed-perspective methods, PVI-Net also exhibited the best mIOU performance. Notably, PVI-Net outperformed RPVNet by 0.6% in mIOU, highlighting the effectiveness and practical value of the cross-attention mechanism and the proposed MF-Attention multi-perspective fusion strategy used in our network compared with the direct averaging fusion approach of RPVNet.

Table 1. Experimental results of the model on the SemanticKITTI dataset. To compare the performance of different models clearly, we divide the compared models into four groups based on the type of input data: point-based input, image-based input, voxel-based input, and mixed-view input. In the table, we specifically highlight the highest mIOU score in each category in red and the second highest score in blue.

Methods	Data	mIoU (%) ↑	Car	Bicycle	Motorcycle	Truck	Other-Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalk	Other-Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic-Sign
PointNet [22]	Point	14.6	46.3	1.3	0.3	0.1	0.8	0.2	0.2	0.0	61.6	15.8	35.7	1.4	41.4	12.9	31.0	4.6	17.6	2.4	3.7
RandLANet [23]	Point	53.9	94.2	26.0	25.8	40.1	38.9	49.2	48.2	7.2	90.2	60.3	73.7	20.4	86.9	56.3	81.4	61.3	66.8	49.2	47.7
KPConv [24]	Point	58.8	96.0	30.2	42.5	33.4	44.3	61.5	61.6	11.8	88.8	61.3	72.7	31.6	90.5	64.2	84.8	69.2	69.1	56.4	47.4
SqueezeSegv3 [25]	Range	55.9	92.5	38.7	36.5	29.6	33.0	45.6	46.2	20.1	91.7	63.4	74.8	26.4	89.0	59.4	82.0	58.7	65.4	49.6	58.9
RangeNet++ [16]	Range	52.2	91.4	25.7	34.4	25.7	23.0	38.3	38.8	4.8	91.8	65.0	75.2	27.8	87.4	58.6	80.5	55.1	64.6	47.9	55.9
SalsaNext [26]	Range	59.5	91.9	48.3	38.6	38.9	31.9	60.2	59.2	19.4	91.7	63.7	75.8	29.1	90.2	64.2	81.8	63.6	66.5	54.3	62.1
PolarNet [27]	Voxel	54.3	93.8	40.3	30.1	22.9	28.5	43.2	40.2	5.6	90.8	61.7	74.4	21.7	90.0	61.3	84.0	65.5	67.8	51.8	57.5
MinkowskiNet [28]	Voxel	63.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cylinder3D [29]	Voxel	67.8	97.1	67.6	64.0	59.0	58.6	73.9	67.9	36.0	91.4	65.1	75.5	32.3	91.0	66.5	85.4	71.8	68.5	62.6	65.6
AF2S3 [30]	Voxel	69.7	94.5	65.4	86.8	39.2	41.1	80.7	80.4	74.3	91.3	68.8	72.5	53.5	87.9	63.2	70.2	68.5	53.7	61.5	71.0
FusionNet [31]	Fusion	61.3	95.3	47.5	37.7	41.8	34.5	59.5	56.8	11.9	91.8	68.8	77.1	30.8	92.5	69.4	84.5	69.8	68.5	60.4	66.5
TornadoNet [32]	Fusion	63.1	94.2	55.7	48.1	40.0	38.2	63.6	60.1	34.9	89.7	66.3	74.5	28.7	91.3	65.6	85.6	67.0	71.5	58.0	65.9
AMVNet [33]	Fusion	65.3	96.2	59.9	54.2	48.8	45.7	71.0	65.7	11.0	90.1	71.0	75.8	32.4	91.4	69.1	85.6	67.0	71.5	58.0	65.9
SPVCNN [34]	Fusion	63.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PVNAS [35]	Fusion	67.0	97.2	50.6	50.4	56.6	58.0	67.4	67.1	50.3	90.2	67.6	75.4	21.8	91.6	66.9	86.1	73.4	71.0	64.3	67.3
RPVNet [36]	Fusion	70.3	97.6	68.4	68.7	44.2	61.1	75.9	74.4	73.4	93.4	70.3	80.7	33.3	93.5	70.2	86.5	75.1	71.7	64.8	61.4
PIV-Net	Fusion	70.9	97.4	67.2	68.9	43.7	61.5	76.6	75.0	73.6	92.3	71.2	80.1	32.8	92.6	70.8	86.9	74.5	72.5	64.8	62.5

4.3.2. Evaluation on nuScenes Dataset

For a comprehensive validation of our model's robustness, we carried out a series of detailed experiments on the nuScenes dataset. *As shown in Table* 2, PVI-Net demonstrated exceptional performance, especially in the key metric of mIOU, where it surpassed other classic single-perspective and multi-perspective networks, achieving a leading position. This result further confirms the enormous potential of multi-perspective data fusion in the field of point cloud semantic segmentation. Notably, by combining point cloud and voxel data, our network effectively overcomes geometric distortions that may occur during point cloud projection, significantly enhancing the accuracy of point cloud segmentation. Moreover, Figure 4 presents the semantic segmentation visualization results of the PVI-Net network on the nuScenes dataset. These experimental results not only showcase the efficient performance of PVI-Net but also emphasize the importance of multi-perspective fusion in enhancing point cloud processing capabilities in complex environments.

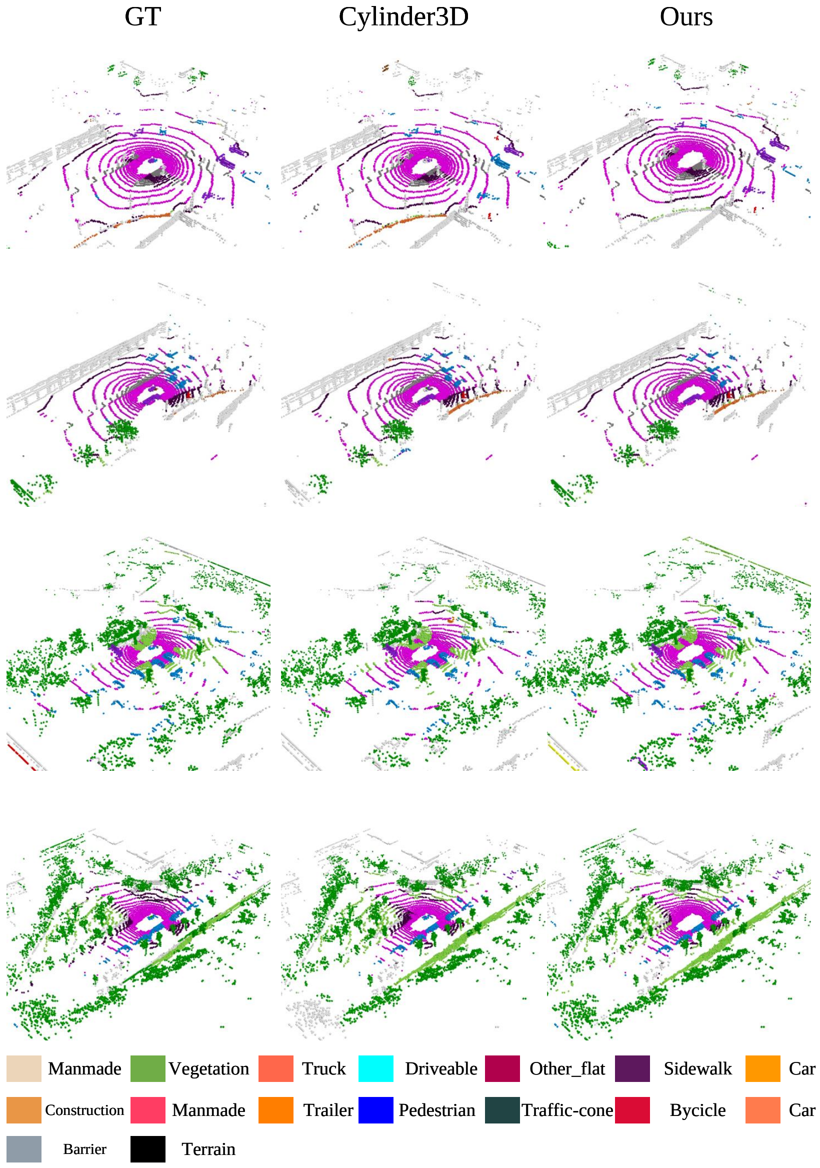

Figure 4. A visual comparison of the results from the model on the nuScenes dataset.

Table 2. Experimental data on PVI-Net for the nuScenes dataset. We highlight the highest score in red and the second-highest score in blue.

Methods	Data	mIoU (%) ↑	Barrier	Bicycle	Bus	Car	Construction	Motorcycle	Pedestrian	Traffic-Cone	Trailer	Truck	Driveable	Other_Flat	Sidewalk	Terrain	Manmade	Vegetation
RangeNet++ [16]	Range	65.5	66.0	21.3	77.2	80.9	30.2	66.8	69.6	52.1	54.2	72.3	94.1	66.6	63.5	70.1	83.1	79.8
PolarNet [27]	Voxel	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 [26]	Range	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
AMVNet [33]	Fusion	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	87.4
Cylinder3D [29]	Voxel	76.1	76.4	40.3	91.2	92.8	51.3	78.0	78.9	64.9	62.1	84.4	96.8	71.6	76.4	75.4	90.5	87.4
RPVNet [36]	Fusion	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
PVI-Net	Fusion	78.1	78.8	43.8	93.5	93.1	48.6	87.0	80.4	65.9	67.5	85.1	97.0	74.5	75.8	76.4	90.6	89.0

4.4. Ablation Study

In this section, we delve into the key components of the PIV-Transformer, conducting a series of fusion experiments to analyze the impact of each branch, the multi-perspective feature deep fusion modules, and the post-fusion modules within the network. Additionally, we evaluate the computational efficiency and parameter count of PVI-Net under various branch combinations. All the aforementioned experiments are implemented on the SemanticKITTI dataset, and we showcase the test results of these methods on the validation part (sequence 08) of this dataset.

4.4.1. Impact of Different Perspectives on Network Performance

A shown in Table 3, we conducted a series of independent and interactive ablation experiments on three different branches. Furthermore, we detailed the required parameter count and model inference speed for each ablation experiment network. For the sake of uniformity, all ablation experiments in Table 3 use the same hardware settings and batch sizes as the PVI-Net network experiments (see Section 4.2). Our experimental results clearly show that, compared with single-perspective inputs, multi-perspective inputs demonstrate better performance in segmentation tasks. Specifically, regarding the point cloud segmentation network's interaction with multi-perspective features, we found that voxel features, as opposed to image features, provide a richer and more comprehensive feature supplement for the point cloud branch.

Table 3. Impact of different perspectives on network performance.

View	mIoU (%) ↑	Params (M) ↓	Latency (ms) ↓
Point	15.3	0.065	13.8
Voxel	65.5	23.3	97.6
Image	50.8	3.36	23.2
Point+Voxel	68.1	24.8	125.4
Point+Image	56.8	3.32	41.3
Point+Voxel+Image	**70.9**	**28.2**	**158.7**

4.4.2. Impact of Multi-Perspective Feature Deep Fusion Modules

In Table 4, we present a series of ablation experiments on the key modules of the PVI-Net network, verifying their contributions in the process of deep feature fusion. In this table, modules marked with a "✓" default to using an averaging method for fusion. Through these experimental results, we observed that each module mentioned in the network positively impacted the model's effectiveness.

Table 4. Impact of different perspectives on network performance.

PVC Attention	MF-Attention	Skip Connection	mIoU (%) ↑
✓			68.8
✓	✓		69.6
✓	✓	✓	70.9

4.4.3. Impact of Multi-Perspective Feature Post-Fusion Module

In Table 5, we specifically compare the multi-perspective feature post-fusion method used in our network with the common Addition (additive fusion) and Concatenation (concatenative fusion) methods. The experimental results show that, on the SemanticKITTI dataset, our fusion method improved the mIoU by 1.7% and 1.4% compared with the Addition and Concatenation methods, respectively. This outcome demonstrates that our fusion strategy more effectively integrates information from different sources when processing multi-perspective data, thereby enhancing the accuracy of semantic segmentation.

Table 5. Impact of different perspectives on network performance.

Method	mIoU (%) ↑
Addition	69.2
Concatenation	69.5
Our fusion	70.9

4.4.4. Multi-Perspective Fusion Addresses Challenges Encountered by Single-Perspective Methods

This paper enhances the understanding of complex 3D scenes by introducing a multi-view fusion approach, addressing the limitations of single-view methods that often miss crucial scene details due to occlusions, scale variations, and viewpoint dependencies. By integrating data from various perspectives, our multi-view fusion technique reconstructs obscured parts, mitigates scale discrepancies, and generates viewpoint-invariant features, leading to improved feature completeness and classification accuracy. Although our initial model, PVI-Net, does not outperform the latest state-of-the-art models in accuracy, it validates the feasibility of multi-view fusion and offers a novel perspective for 3D scene comprehension.

5. Conclusions

In conclusion, PVI-Net stands as a testament to the innovative exploration of point cloud semantic segmentation, particularly within the realm of autonomous driving. Central to our framework is the strategic intra-modal fusion of three distinct representations of a singular point cloud dataset. This fusion, achieved through parallel processing branches, underscores our commitment to extracting a richer, more nuanced feature set from point cloud data. We introduced point cloud–voxel cross-attention and point–image multi-perspective feature fusion strategies, which are innovative approaches that enable effective information interaction between different perspectives, significantly optimizing the process of information fusion between perspectives. Additionally, PVI-Net employs a U-Net architecture and residual connections. These not only enhance the precision and efficiency of semantic segmentation but also present an innovative method for multi-perspective feature post-fusion. This effectively integrates information from different data sources, thereby improving the accuracy of semantic segmentation. Extensive experiments in autonomous driving scenarios confirm that PVI-Net demonstrates outstanding performance in point cloud semantic segmentation.

Author Contributions: Funding acquisition, C.L.; resources, C.L.; validation, J.M. and Z.F.; visualization, L.X.; writing—original draft, Z.W.; writing—review and editing, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: The work of this thesis was partially funded by the National Natural Science Foundation of China (Grant No.: 62363025), the Science and Technology Program of Lanzhou (No.: 2022-2-58), and the Key R&D plan of Science and Technology Plan of Gansu Province—Social Development Field Project (No.: 23YFFA0064).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: [HTML]FE0000Data are contained with in the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yan, X.; Zhan, H.; Zheng, C.; Gao, J.; Zhang, R.; Cui, S.; Li, Z. Let images give you more: Point cloud cross-modal training for shape analysis. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 32398–32411.
2. Yan, X.; Gao, J.; Zheng, C.; Zheng, C.; Zhang, R.; Cui, S.; Li, Z. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 677–695.

3. Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Zhou, J.; Lu, J. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2023, Paris, France, 4–6 October 2023; pp. 21729–21740.
4. Ottonelli, S.; Spagnolo, P.; Mazzeo, P.L.; Leo, M. Improved video segmentation with color and depth using a stereo camera. In Proceedings of the IEEE International Conference on Industrial Technology 2013, Cape Town, South Africa, 25–28 February 2013; pp. 1134–1139.
5. Zhang, Z.; Yang, B.; Wang, B.; Li, B. GrowSP: Unsupervised Semantic Segmentation of 3D Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023, Vancouver, BC, Canada, 17–24 June 2023; pp. 17619–17629.
6. Xia, Y.; Gladkova, M.; Wang, R.; Li, Q.; Stilla, U.; Henriques, J.F.; Cremers, D. CASSPR: Cross Attention Single Scan Place Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2023, Paris, France, 4–6 October 2023; pp. 8461–8472.
7. Fan, S.; Dong, Q.; Zhu, F.; Lv, Y.; Ye, P.; Wang, F.Y. SCF-Net: Learning spatial contextual features for large-scale point cloud segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 14504–14513.
8. Li, L.; He, L.; Gao, J.; Han, X. Psnet: Fast data structuring for hierarchical deep learning on point cloud. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6835–6849. [CrossRef]
9. Nie, D.; Lan, R.; Wang, L.; Ren, X. Pyramid architecture for multi-scale processing in point cloud segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 17284–17294.
10. Phan, A.V.; Le Nguyen, M.; Nguyen, Y.L.H.; Bui, L.T. Dgcnn: A convolutional neural network over large-scale labeled graphs. *Neural Netw.* **2018**, *108*, 533–543. [CrossRef] [PubMed]
11. Yuan, W.; Gu, X.; Li, H.; Dong, Z.; Zhu, S. Monocular Scene Reconstruction with 3D SDF Transformers. *arXiv* **2023**, arXiv:2301.13510.
12. Cui, M.; Long, J.; Feng, M.; Li, B.; Kai, H. OctFormer: Efficient octree-based transformer for point cloud compression with local enhancement. In Proceedings of the AAAI Conference on Artificial Intelligence 2023, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 470–478.
13. Fei, J.; Chen, W.; Heidenreich, P.; Wirges, S.; Stiller, C. SemanticVoxels: Sequential fusion for 3D pedestrian detection using LiDAR point cloud and semantic segmentation. In Proceedings of the 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Virtual, 14–16 September 2020; pp. 185–190.
14. Park, C.; Jeong, Y.; Cho, M.; Park, J. Efficient Point Transformer for Large-Scale 3D Scene Understanding. Available online: https://openreview.net/forum?id=3SUToIxuIT3 (accessed on 1 January 2024)
15. Wang, H.; Shi, C.; Shi, S.; Lei, M.; Wang, S.; He, D.; Schiele, B.; Wang, L. Dsvt: Dynamic sparse voxel transformer with rotated sets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023, Vancouver, BC, Canada, 17–24 June 2023; pp. 13520–13529.
16. Milioto, A.; Vizzo, I.; Behley, J.; Stachniss, C. Rangenet++: Fast and accurate lidar semantic segmentation. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 4213–4220.
17. Wu, B.; Wan, A.; Yue, X.; Keutzer, K. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 1887–1893.
18. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.L.; Han, S. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 2774–2781.
19. Zhang, Z.; Shen, Y.; Li, H.; Zhao, X.; Yang, M.; Tan, W.; Pu, S.; Mao, H. Maff-net: Filter false positive for 3d vehicle detection with multi-modal adaptive feature fusion. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; pp. 369–376.
20. Afham, M.; Dissanayake, I.; Dissanayake, D.; Dharmasiri, A.; Thilakarathna, K.; Rodrigo, R. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 9902–9912.
21. Chen, A.; Zhang, K.; Zhang, R.; Wang, Z.; Lu, Y.; Guo, Y.; Zhang, S. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023, Vancouver, BC, Canada, 17–24 June 2023; pp. 5291–5301.
22. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
23. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 13–19 June 2020; pp. 11108–11117.

24. Thomas, H.; Qi, C.R.; Deschaud, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. Kpconv: Flexible and deformable convolution for point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2019, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6411–6420.
25. Xu, C.; Wu, B.; Wang, Z.; Zhan, W.; Vajda, P.; Keutzer, K.; Tomizuka, M. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 1–19.
26. Cortinhal, T.; Tzelepis, G.; Erdal Aksoy, E. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In Proceedings of the Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, 5–7 October 2020; pp. 207–222.
27. Zhang, Y.; Zhou, Z.; David, P.; Yue, X.; Xi, Z.; Gong, B.; Foroosh, H. 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 2020, Seattle, WA, USA, 13–19 June 2020; pp. 9601–9610.
28. Choy, C.; Gwak, J.; Savarese, S. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, Long Beach, CA, USA, 15–20 June 2019; pp. 3075–3084.
29. Zhou, H.; Zhu, X.; Song, X.; Ma, Y.; Wang, Z.; Li, H.; Lin, D. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. *arXiv* **2020**, arXiv:2008.01550.
30. Cheng, R.; Razani, R.; Taghavi, E.; Li, E.; Liu, B. 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 2021, Nashville, TN, USA, 20–25 June 2021; pp. 12547–12556.
31. Zhang, F.; Fang, J.; Wah, B.; Torr, P. Deep fusionnet for point cloud semantic segmentation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 644–663.
32. Gerdzhev, M.; Razani, R.; Taghavi, E.; Bingbing, L. Tornado-net: Multiview total variation semantic segmentation with diamond inception module. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 9543–9549.
33. Liong, V.E.; Nguyen, T.N.T.; Widjaja, S.; Sharma, D.; Chong, Z.J. Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. *arXiv* **2020**, arXiv:2012.04934.
34. Axelsson, M.; Holmberg, M.; Serra, S.; Ovren, H.; Tulldahl, M. Semantic labeling of lidar point clouds for UAV applications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, Nashville, TN, USA, 20–25 June 2021; pp. 4314–4321.
35. Liu, Z.; Tang, H.; Zhao, S.; Shao, K.; Han, S. Pvnas: 3d neural architecture search with point-voxel convolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 8552–8568. [CrossRef] [PubMed]
36. Xu, J.; Zhang, R.; Dou, J.; Zhu, Y.; Sun, J.; Pu, S. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 16024–16033.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.