symmetry

MDPI

Article

LSSCC-Net: Integrating Spatial-Feature Aggregation and Adaptive Attention for Large-Scale Point Cloud Semantic Segmentation

Wenbo Wang [1,2], Xianghong Hua [1,*], Cheng Li [1], Pengju Tian [3], Yapeng Wang [2] and Lechao Liu [2]

[1] The School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China; 2024282140133@whu.edu.cn (W.W.); lc.whu@whu.edu.cn (C.L.)
[2] 61365 Troops, Tianjin 300140, China
[3] The Engineering Research Center of Environmental Laser Remote Sensing Technology and Application of Henan Province, Nanyang Normal University, Nanyang 473061, China; pjtian@whu.edu.cn
* Correspondence: xhhua@sgg.whu.edu.cn

Abstract

Point cloud semantic segmentation is a key technology for applications such as autonomous driving, robotics, and virtual reality. Current approaches are heavily reliant on local relative coordinates and simplistic attention mechanisms to aggregate neighborhood information. This often leads to an ineffective joint representation of geometric perturbations and feature variations, coupled with a lack of adaptive selection for salient features during context fusion. On this basis, we propose LSSCC-Net, a novel segmentation framework based on LACV-Net. First, the spatial-feature dynamic aggregation module is designed to fuse offset information by symmetric interaction between spatial positions and feature channels, thus supplementing local structural information. Second, a dual-dimensional attention mechanism (spatial and channel) is introduced to symmetrically deploy attention modules in both the encoder and decoder, prioritizing salient information extraction. Finally, Lovász-Softmax Loss is used as an auxiliary loss to optimize the training objective. The proposed method is evaluated on two public benchmark datasets. The mIoU on the Toronto3D and S3DIS datasets is 83.6% and 65.2%, respectively. Compared with the baseline LACV-Net, LSSCC-Net showed notable improvements in challenging categories: the IoU for "road mark" and "fence" on Toronto3D increased by 3.6% and 8.1%, respectively. These results indicate that LSSCC-Net more accurately characterizes complex boundaries and fine-grained structures, enhancing segmentation capabilities for small-scale targets and category boundaries.

Keywords: spatial-feature dynamic aggregation; Adaptive Attention Fusion Mechanism; loss function; LACV-Net network; point cloud semantic segmentation

Academic Editor: Jie Yang

Received: 1 December 2025
Revised: 3 January 2026
Accepted: 6 January 2026
Published: 8 January 2026

1. Introduction

Rapid advancements in 3D scanning technologies, including LiDAR and RGB-D depth cameras, have significantly increased interest in the intelligent processing of point cloud data. Point cloud semantic segmentation, a fundamental task in 3D scene understanding and analysis, aims to assign semantic labels to each point in 3D space accurately. This technology is crucial for applications such as autonomous driving and intelligent robotics and has played an important role in their rapid development [1]. Unlike 2D images,

3D point clouds are unordered, sparse, and unstructured. These properties reduce the efficiency of traditional convolutional neural networks (CNNs) in processing such data [2].

Traditional point cloud semantic segmentation relies on manually extracting neighborhood geometric features, followed by machine learning classification. Recently, deep learning has made significant strides in image segmentation and natural language processing, which have been adapted to point cloud processing. This adaptation has led to notable improvements in semantic segmentation for small-scale point clouds. However, large-scale point clouds often consist of millions or even billions of points, which significantly increases scene complexity. This complexity has slowed the development of segmentation technologies for large-scale scenes [3,4].

Deep learning-based point cloud semantic segmentation methods are primarily categorized into projection-based [5–8], voxel-based [9–12], point-based [13–23], and related fusion methods [24–27]. Projection-based methods segment point clouds by converting them into 2D images, utilizing advanced 2D image segmentation algorithms. Voxel-based methods transform point clouds into voxel grids, similar to image pixels, and apply convolutional networks for feature extraction. However, projecting point clouds into 2D images inevitably leads to the loss of 3D spatial information. As a result, the intrinsic three-dimensional characteristics cannot be fully exploited. Similarly, voxelization causes detail loss due to quantization and increases computational complexity, reducing algorithm performance. In contrast, point-based methods directly use raw point clouds, preserving their fine-grained geometric structures and spatial relationships. Despite this advantage, they suffer from low processing efficiency, high computational costs, and a limited ability to extract global features.

PointNet [13], the first method for direct point-cloud processing, applies shared multilayer perceptrons to each point and aggregates global context with symmetric pooling functions. Because it processes points independently, PointNet cannot capture the local structure needed for fine-grained detail. PointNet++ [14] addressed this by introducing multi-level feature extraction and adaptive density sampling, which improved learning efficiency and robustness, but it still did not model relationships among neighboring points. RandLA-Net [15] introduced a simple, efficient design based on random sampling and local feature aggregation that markedly enhanced scalability for large-scale point clouds. Zhao et al. observed that self-attention mechanisms are highly compatible with 3D vision tasks and were the first to introduce self-attention into point-cloud processing by proposing the Point Transformer [18] architecture. The framework leverages self-attention to capture local and global context, employs vector attention (VA) within local kNN neighborhoods for information exchange, and uses learnable relative positional encodings to model spatial relationships. Consequently, it achieved strong performance in point-cloud semantic segmentation. However, the Point Transformer has several limitations. First, the parameter count of vector attention increases rapidly with channel dimensionality, making the model prone to overfitting. Second, kNN search and relative positional encoding are computationally expensive, which reduces efficiency. Finally, the overall framework is difficult to scale. To address these issues, Wu et al. proposed Grouped Vector Attention (GVA) in the improved Point Transformer V2 [19] network, replacing VA with group-shared weights to reduce parameters and mitigate overfitting. Nevertheless, Point Transformer V2 still depended on kNN search and complex positional encoding, which constrained efficiency and receptive field expansion. Point Transformer V3 [20] further improved efficiency by using serialized neighborhoods based on space-filling curves instead of kNN queries, simplifying attention interactions, and replacing relative positional encoding with enhanced conditional positional encoding (xCPE), thereby accelerating inference and reducing memory consumption. Fan et al. proposed SCF-Net [21], which learns spatial contextual features to

capture correlations between points and thereby improves segmentation performance. Liu et al. introduced DG-Net [2], which fuses geometric structure with semantic features and models long-range dependencies, enabling better segmentation of objects with complex geometries. NeiEA-Net [28] optimizes local neighborhoods in 3D Euclidean space through two modules: Neighborhood Feature Enhancement (NFE) and Neighborhood Feature Aggregation (NFA), effectively learning local details in point clouds. LACV-Net [16] addresses local perception ambiguity and insufficient global features in large-scale point cloud semantic segmentation through three components: the Local Adaptive Feature Augmentation (LAFA) module, aggregation loss function, and Comprehensive Vector of Locally Aggregated Descriptors (C-VLAD) module. However, its segmentation accuracy is still limited in scenes with complex scale variations.

Recent methods have advanced modeling of inter-point dependencies and long-range context, substantially improving large-scale point cloud semantic segmentation. However, explicit modeling of geometric perturbations—such as point position offsets and variations in surface normals—remains limited in many existing approaches. Consequently, accurately characterizing complex geometric boundaries (for example, wall corners and object edges) and extracting discriminative features for small structures (e.g., thin pole, fence) are still challenging. In addition, downsampling often removes fine details, making it difficult to reconcile global context with preservation of local structure. Existing attention-based point-cloud networks also have notable limitations. First, attention mechanisms are often applied along a single dimension and thus optimize only one aspect of feature representation, which is inadequate for the multi-dimensional feature expression required in complex scenes. For example, the attention mechanism in RandLA-Net [15] focuses solely on spatial local feature aggregation and cannot suppress redundant channel-wise information. Although NeiEA-Net [28] enhances neighborhood spatial features, it does not perform adaptive weighting in the channel dimension. Second, attention deployment is imbalanced: most networks concentrate attention modules in the encoder for local feature extraction, while the decoder relies mainly on simple skip connections for feature fusion and lacks effective attention guidance. Consequently, important features compressed during encoding are easily corrupted by redundant information during decoding, which limits segmentation accuracy. Finally, large-scale point-cloud datasets often exhibit severe class imbalance, and existing models generally have insufficient capacity to learn categories that contain few sample points. To address these challenges, this paper further optimized the LACV-Net [16] architecture. During the encoding stage, offset information was extracted separately from spatial coordinates and feature channels, and contextual information was fused via max pooling and weighted summation. This symmetric interaction strategy alleviated semantic ambiguity among neighboring points and strengthened local contextual representation. Meanwhile, a Spatial Attention Mechanism (SAM) was introduced to generate attention maps using 1×1 convolution and transposed convolution, followed by Sigmoid activation to weight local features and emphasize key spatial regions. In the decoding stage, a Channel Attention Mechanism (CAM) was incorporated; it learned channel-wise weights through a shared MLP applied to globally averaged and max-pooled features, substantially enhancing feature expressiveness. The symmetric deployment of spatial and channel attention mechanisms effectively overcame the dimensional and structural limitations of existing attention-based networks. Moreover, Lovász-Softmax Loss [29] is introduced as an auxiliary optimization target alongside the original loss. This complementary mechanism enhances the model's ability to segment small targets, class boundaries, and large object regions. It overcomes the limitations of using a single loss for region-level optimization, thereby improving segmentation accuracy. Key innovations include the following:

(a) A Spatial-Feature Dynamic Aggregation (SFDA) module is constructed. Offset information is extracted through symmetric interaction in two dimensions (spatial position and feature channel), enhancing the local context representation of point clouds.
(b) An Adaptive Attention Fusion Mechanism (AAFM) module is designed. Spatial Attention Mechanism (SAM) and Channel Attention Mechanism (CAM) are deployed symmetrically in the encoder and decoder, respectively. This builds symmetric saliency perception across spatial and channel dimensions, improving the model's feature expression capability.
(c) An IoU-oriented loss function collaborative optimization framework is established. Lovász-Softmax Loss is introduced as an auxiliary optimization objective. Through a complementary mechanism, the model's ability to segment small targets, class boundaries, and complete regions of large objects is enhanced, overcoming the limitations of a single loss in region-level optimization.

2. Methodology

The LSSCC-Net point cloud semantic segmentation network, introduced in this paper, builds on the LACV-Net architecture [16]. LACV-Net employs an encoder–decoder framework, utilizing the Local Adaptive Feature Augmentation (LAFA) module to reduce local perception ambiguity. It integrates global features with the Comprehensive Vector of Locally Aggregated Descriptors (C-VLAD) module and enhances model convergence with an aggregation loss function, excelling in large-scale point cloud segmentation tasks. To address the limitations of LACV-Net in local structure modeling, feature selection, and loss optimization, three key improvements are introduced while preserving the original encoder–decoder architecture. First, a Spatial–Feature Dynamic Aggregation (SFDA) module is introduced during the encoding stage to enhance robustness to spatial structure variations through symmetric interaction modeling. Second, this study incorporated the Adaptive Attention Fusion Mechanism (AAFM) module, adding the Spatial Attention Mechanism (SAM) in the encoding stage to emphasize salient regions, and the Channel Attention Mechanism (CAM) in the decoding stage to enhance feature expression comprehensively. Third, this study employed Lovász-Softmax Loss as an auxiliary loss optimization objective to improve segmentation capabilities for small objects, category boundaries, and large object regions. This section details the LSSCC-Net network and its enhanced modules. The improved LACV-Net architecture is termed the LSSCC-Net, as depicted in Figure 1.

LSSCC-Net comprises an encoder, a decoder, a classifier, and a C-VLAD (Comprehensive Vector of Locally Aggregated Descriptors) module that connects the encoder and decoder. In the encoding phase, five layers are implemented. Each layer processes input features using the Local Adaptive Feature Augmentation (LAFA) and SFDA modules. After merging outputs from these modules, the Spatial Attention Mechanism (SAM) applies feature weighting to emphasize critical spatial features. The encoder progressively reduces the number of points through downsampling ($N \rightarrow N/4 \rightarrow N/16 \rightarrow N/64 \rightarrow N/256$) while increasing the channel size ($16 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$). These downsampled features serve as input for the next encoder layer. Positioned between the encoder and decoder, the C-VLAD module aggregates local features from all preceding encoding layers. The decoder restores the point cloud count through upsampling. Each decoding layer incorporates the Channel Attention Mechanism (CAM) and a Multi-Layer Perceptron (MLP), using skip connections to transfer features between the encoder and decoder, thus preserving local detail. Finally, the classifier employs three fully connected layers and a Dropout layer to predict semantic labels, completing the semantic segmentation process.

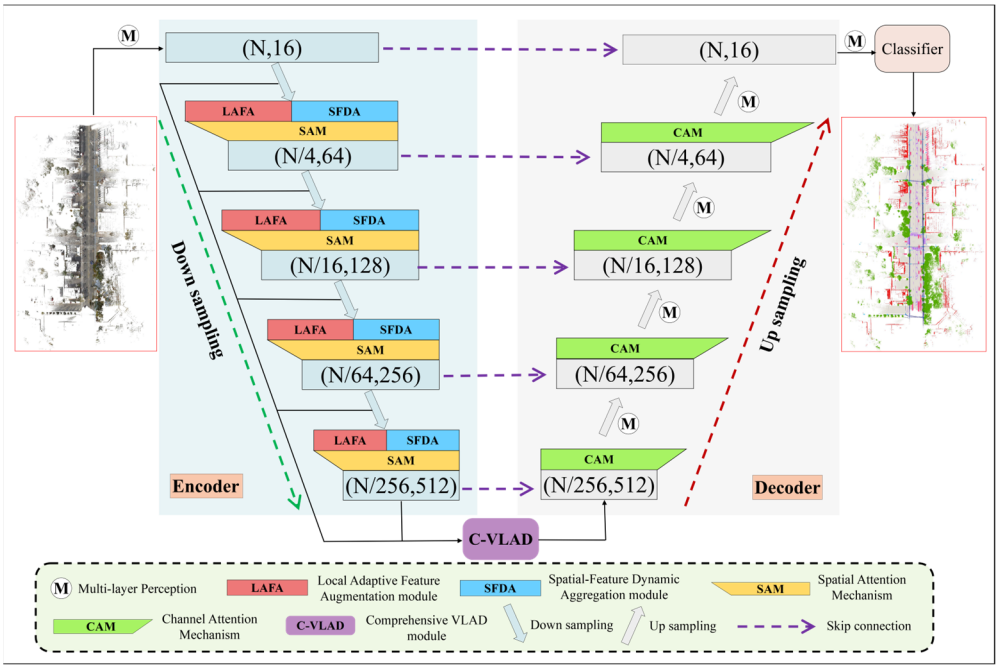

Figure 1. Architecture of the LSSCC-Net.

2.1. Spatial-Feature Dynamic Aggregation Module

The original model aggregates neighborhood information primarily using local relative coordinates and simple attention pooling. This design limits its ability to handle geometric perturbations and feature variations between points, thereby hindering accurate characterization of complex boundaries and fine-grained structures. To overcome these limitations, the network presented here incorporates a SFDA module into the encoding stage, enhancing local structure perception by modeling symmetric interactions between spatial positions and feature channels (Figure 2). The central idea of the SFDA module is to extract offset information through symmetric interactions along two dimensions—spatial location and feature channel—to enrich local contextual representations of point clouds and reduce semantic ambiguity among neighboring points (for example, classification confusion arising from densely distributed points near boundaries and the difficulty of identifying local geometric structures of small objects).

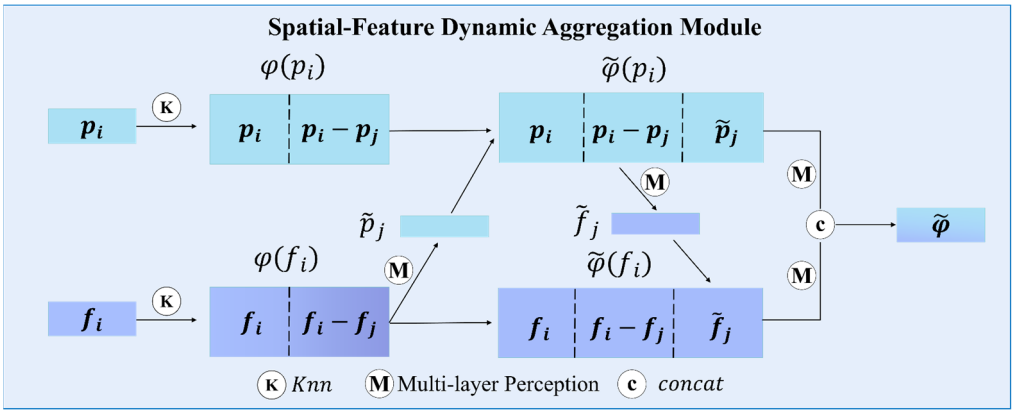

Figure 2. Spatial-Feature Dynamic Aggregation Module.

This module initially establishes spatial and feature offsets using local coordinate differences. It then encodes and fuses these offsets to create enhanced features capable of perturbation perception. Specifically, first, the K-nearest neighbor (KNN) algorithm based

on Euclidean distance is used to find the neighboring points $p_j \in N(p_i)$ of the center point p_i, and the corresponding feature information is denoted as f_i and f_j. The absolute position of the center point and the relative positions of its neighboring points are concatenated into local contextual information φ. Correspondingly, the local spatial contextual information is denoted as $\varphi(p_i)$, and the local feature contextual information is denoted as $\varphi(f_i)$, which is represented as follows:

$$\varphi(p_i) = \left| p_i, p_i - p_j \right| \tag{1}$$

$$\varphi(f_i) = \left| f_i, f_i - f_j \right| \tag{2}$$

Relying solely on a fixed structure φ and neighborhood constraints in the three-dimensional space of point cloud data often results in poor generalization and feature redundancy. To address this, the module learns offsets symmetrically from both spatial and feature dimensions to enhance local contextual point information. This process involves two main steps. One is spatial offset learning: based on the rich feature information of $\varphi(f_i)$, the spatial offset is learned through a Multi-Layer Perceptron (MLP). Adjusting the coordinates of neighboring points can obtain enhanced local spatial contextual information, which can be expressed as

$$\widetilde{\varphi}(p_i) = \left| p_i, p_i - p_j, \widetilde{p}_j \right| \tag{3}$$

$$\widetilde{p}_j = M(\varphi(f_i)) + p_j \tag{4}$$

where $\widetilde{\varphi}(p_i)$ represents the enhanced local spatial contextual information, and $\widetilde{p}_j$ is the coordinate of the neighboring point after offset. M refers to a Multi-Layer Perceptron (MLP). The second component involves feature offset learning: utilizing the enhanced local spatial contextual information, the feature offset is further refined. By adjusting the features of neighboring points, enhanced local feature contextual information can be obtained, expressed as

$$\widetilde{\varphi}(f_i) = \left| f_i, f_i - f_j, \widetilde{f}_j \right| \tag{5}$$

$$\widetilde{f}_j = M\left(\widetilde{\varphi}(p_i)\right) + f_j \tag{6}$$

The enhanced local feature contextual information $\widetilde{\varphi}(f_i)$ is derived from several components: f_i, representing the feature information at the center point, and f_j, representing the feature information at a neighboring point. After applying an offset, the neighboring point's feature becomes $\widetilde{f}_j$. Subsequently, $\widetilde{\varphi}(p_i)$ and $\widetilde{\varphi}(p_i)$ are integrated using a Multi-Layer Perceptron (MLP) to produce the local spatial-feature contextual information $\widetilde{\varphi}$.

$$\widetilde{\varphi} = concat\left(M\left(\widetilde{\varphi}(p_i)\right); M\left(\widetilde{\varphi}(f_i)\right)\right) \tag{7}$$

The integration of the SFDA module into each encoding layer allows for the step-by-step processing of point cloud data at varying resolutions, effectively capturing multi-scale information. Each SFDA module receives spatial coordinates and feature information as inputs, producing enriched feature representations through enhanced symmetric information interaction and aggregation in both spatial and feature dimensions.

2.2. Adaptive Attention Fusion Mechanism Module

Existing networks often struggle to automatically identify key spatial regions or informative feature dimensions in point cloud data, which limits their ability to capture salient information. An AAFM module is proposed in this study to overcome this limitation. The module symmetrically integrates a Spatial Attention Mechanism (SAM) in the encoder

and a Channel Attention Mechanism (CAM) in the decoder, forming a saliency perception module for both spatial and channel dimensions. The encoder's primary role is feature extraction. Spatial Attention Mechanism (SAM) enhances this by generating an attention map using 1×1 convolution and transposed convolution. It then applies the Sigmoid function to activate and weight local features, enabling the model to focus on crucial spatial regions and extract more representative features. Conversely, the decoder's main task is feature recovery. Channel Attention Mechanism (CAM) contributes by learning channel weights through a shared Multi-Layer Perceptron (MLP) after performing global average pooling and max pooling. This process helps adjust the weights of different channels, thereby recovering more accurate semantic information. By applying the Spatial Attention Mechanism (SAM) and the Channel Attention Mechanism (CAM) to the encoder and decoder, respectively, the module leverages complementary strengths through symmetric collaboration. This approach optimizes feature extraction and recovery, enhancing the model's performance and generalization ability.

2.2.1. Spatial Attention Mechanism

During the collection of 3D point cloud data, various sources of noise and irrelevant information often arise, degrading point cloud semantic segmentation performance. To address this issue, our proposed network incorporates the Spatial Attention Mechanism (SAM) during the encoding stage. This mechanism enhances feature extraction by concentrating on key regions and filtering out redundant information, thereby aiding the model in capturing essential spatial structure features. As shown in Figure 3, the Spatial Attention Mechanism (SAM) performs 1×1 transposed convolution on the feature F processed by the LAFA module and the SFDA module to generate an intermediate feature map M with the same number of channels as the input feature:

$$M = Conv_{1\times1}^{Transpose}(F) \tag{8}$$

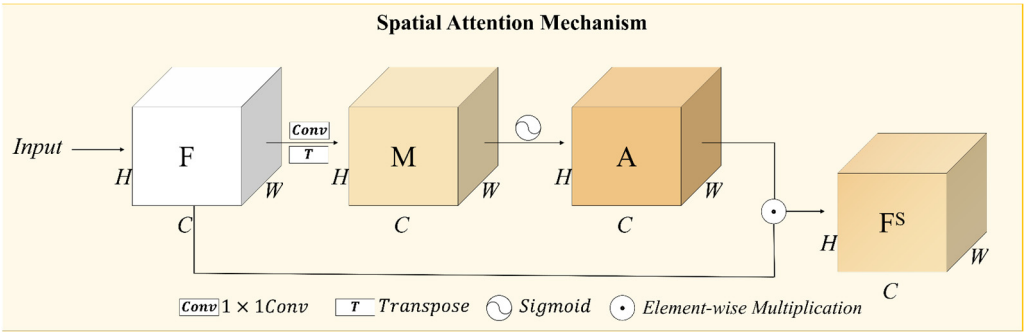

Figure 3. Spatial Attention Mechanism.

The *sigmoid* activation function is applied to the intermediate feature map M to map the values of M to the range [0, 1], obtaining the spatial attention map A:

$$A = sigmoid(M) \tag{9}$$

Subsequently, the spatial attention map A is element-wise multiplied by the input feature F to obtain the feature F^S weighted by the Spatial Attention Mechanism (SAM).

$$F^S = F \odot A \tag{10}$$

2.2.2. Channel Attention Mechanism

Processing point cloud data in large-scale scenes often suffers from inter-class ambiguity. Objects with similar shapes and structures can be challenging to label correctly with semantic tags. To address this, the paper incorporates the Channel Attention Mechanism (CAM) during the decoding stage. Channel Attention Mechanism (CAM) enhances the network's feature expression capability by adjusting channel weights and learning inter-channel correlations to highlight crucial channel information. As shown in Figure 4, the Channel Attention Mechanism (CAM) performs global average pooling and global max pooling operations on the input feature F respectively, compressing the feature information of each channel into a scalar to obtain the average-pooled feature vector Z_{avg} and the max-pooled feature vector Z_{max}. Two fully connected operations are performed on Z_{avg} and Z_{max} respectively to obtain intermediate feature vectors Y_{avg} and Y_{max}:

$$Y_{avg} = W_2 \cdot ReLU(W_1 \cdot Z_{avg}) \tag{11}$$

$$Y_{max} = W_2 \cdot ReLU(W_1 \cdot Z_{max}) \tag{12}$$

where W_1 and W_2 are the weight matrices of the fully connected layers, and $ReLU$ is the activation function. Then, Y_{avg} and Y_{max} are added, and the $Sigmoid$ activation function is applied to obtain the channel attention map B:

$$B = Sigmoid(Y_{avg} + Y_{max}) \tag{13}$$

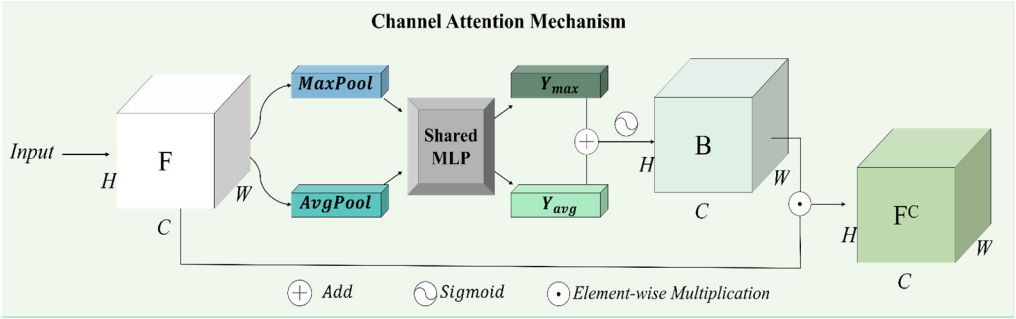

Figure 4. Channel Attention Mechanism.

Subsequently, the channel attention map B is element-wise multiplied by the input feature F to obtain the feature F^C enhanced by the Channel Attention Mechanism (CAM).

$$F^C = F \odot B \tag{14}$$

2.3. Constructing an IoU-Oriented Loss Function Collaborative Optimization Framework

An IoU-oriented optimization strategy is adopted, incorporating Lovász-Softmax Loss as an auxiliary component to form a collaborative optimization framework. This framework combines aggregation loss L_{agg} with the Lovász-Softmax Loss L_{lov}. By employing a complementary mechanism, the model enhanced its segmentation abilities for small objects, category boundaries, and large object regions. This approach overcame the limitations associated with using a single loss in region-level optimization.

The Lovász-Softmax Loss aims to reformulate the optimization of the Jaccard index, also known as Intersection over Union (IoU), into a continuously differentiable problem. IoU is a crucial metric for evaluating point cloud semantic segmentation, effectively reflecting segmentation quality. However, its non-convex and non-differentiable nature complicates direct optimization within neural networks. By applying the Lovász

extension to sorted errors, the Lovász-Softmax Loss transforms IoU loss into a convex and differentiable surrogate loss function, making it amenable to optimization through gradient descent.

The Lovász-Softmax Loss utilizes the Lovász extension to prioritize misclassified samples based on their potential to enhance the IoU. In multi-class semantic segmentation, it optimizes the IoU loss for each category c and combines these losses to determine the overall loss. The Lovász-Softmax Loss is represented as

$$L_{lov} = \frac{1}{C}\sum_{C=1}^{C} \overline{\Delta_{J_{nc}}}(m(i)) \tag{15}$$

where $m(i)$ denotes the point-wise error for the i-th class among the C classes, and $\overline{\Delta_{J_{nc}}}$ denotes the Lovász extension of the Jaccard index. The final total loss function can be expressed as

$$L = L_{agg} + L_{lov} \tag{16}$$

3. Experiments and Analysis

To comprehensively evaluate the performance of the proposed LSSCC-Net architecture for large-scale point cloud semantic segmentation, experiments were conducted on two widely used benchmark datasets: the outdoor Toronto3D [30] and the indoor S3DIS [31]. Ablation experiments were also conducted on modified network modules to validate the contributions of the proposed components.

All experiments were conducted on a hardware platform featuring an Intel Core i7-14700KF CPU and an NVIDIA GeForce RTX 4070 Ti SUPER GPU (16 GB VRAM), operating on Ubuntu 20.04. The model was constructed and trained using TensorFlow 2.4 within a Python 3.8 environment. Point cloud visualization primarily utilized CloudCompare (version 2.13.2) software. For the Toronto3D and S3DIS datasets, grid sampling resolutions were 0.06 m and 0.04 m, respectively. Hardware limitations necessitated different hyperparameter settings from the baseline network: the initial number of input points for training was 30,720, with a batch size of 4, over 100 epochs. The Adam optimizer was used with an initial learning rate of 0.01, and the neighborhood search range K was set to 16.

To quantitatively evaluate semantic segmentation performance, three metrics were adopted: overall accuracy (OA), class-wise Intersection-over-Union (IoU), and mean Intersection-over-Union (mIoU). The formulas for these metrics are outlined below:

$$\mathrm{OA} = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$\mathrm{IoU} = \frac{TP}{FN + FP + TP} \tag{18}$$

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i}{FN_i + FP_i + TP_i} \tag{19}$$

In this context, TP represents true positives, indicating positive samples correctly predicted by the model. TN stands for true negatives, referring to negative samples accurately identified as negative. FP denotes false positives, where negative samples are incorrectly predicted as positive. FN signifies false negatives, where positive samples are mistakenly identified as negative. N indicates the number of categories in the dataset.

3.1. Analysis of Point Cloud Semantic Segmentation Results

3.1.1. Semantic Segmentation for Outdoor Scenes

The Toronto3D outdoor laser point cloud dataset was gathered on Toronto roads using a mobile laser scanning (MLS) system mounted on a vehicle. It features various urban street scenes, covering an area of about 1 km, divided into four regions, each approximately 250 m long. The dataset comprises over 78.3 million points, each with XYZ coordinates, RGB color information, and manually annotated labels for 8 semantic categories, including building, road, tree, and car. In this study, we adhered to the official training and test set division: Region 2 was the test set, while the other three regions served as the training set.

As shown in Table 1, LSSCC-Net achieves the best overall performance compared with other methods in terms of Overall Accuracy (OA, 97.7%) and mean Intersection over Union (mIoU, 83.6%). Specifically, it attains the optimal IoU values for three categories, namely "road", "road mark", and "fence", which are 0.4%, 3.6%, and 3.9% higher than those of the second-ranked algorithm, respectively. Under identical experimental settings, LSSCC-Net improves OA by 0.4% and mIoU by 2.1% compared with the baseline LACV-Net. Furthermore, for six categories (road, road mark, natural, building, pole, and fence), the IoU values of the improved LSSCC-Net are 0.4%, 3.6%, 0.7%, 0.6%, 3.6%, and 8.1% higher than those of the baseline LACV-Net, respectively. These results demonstrate the effectiveness of the proposed improvements.

Table 1. Quantitative comparison results of different methods on Toronto3D (Area 2) unit: %.

Method	Year	OA	mIoU	Road	Road Mark	Natural	Building	Util.line	Pole	Car	Fence
ResDLPS-Net [4]	2021	96.5	80.3	95.8	59.8	96.1	90.9	86.8	79.9	89.4	43.3
RandLA-Net [32]	2021	94.4	81.8	96.7	64.2	96.9	94.2	88.0	77.8	93.4	42.9
BEMF-Net [33]	2023	97.0	81.4	96.2	60.0	**97.6**	94.1	87.9	**85.7**	94.1	35.9
DG-Net [2]	2024	96.8	82.1	97.1	65.2	97.2	92.6	88.1	84.2	93.6	38.7
LACV-Net [16]	2024	97.4	82.7	97.1	66.9	97.3	93.0	87.3	83.5	93.4	43.1
HFA-Net [23]	2025	96.7	81.1	96.1	60.9	96.7	91.9	87.4	82.4	91.8	41.9
ELMF-Net [1]	2025	97.0	81.7	96.1	59.0	97.2	**94.3**	**88.5**	84.6	94.2	39.8
RandLA-Net *	-	97.4	82.1	97.1	67.6	96.6	91.9	87.4	83.2	93.8	39.3
LACV-Net *	-	97.3	81.5	97.3	68.8	96.3	91.7	87.5	76.9	**94.5**	39.1
LSSCC-Net (Ours) *	-	**97.7**	**83.6**	**97.7**	**72.4**	97.0	92.3	87.4	80.5	94.2	**47.2**

Bold numbers denote the best performance for each metric, while underscored numbers (_) indicate the second-best performance. Methods marked with an asterisk (*) were trained and tested using the experimental environment and hyperparameters specified in this study. Performance results for other methods were sourced from publicly available literature.

To effectively compare the segmentation performance of various methods on the Toronto3D dataset, Figure 5 illustrates the overall scene segmentation visualizations for RandLA-Net, LACV-Net, and LSSCC-Net. Additionally, Figure 6 displays the local semantic segmentation results for the same methods on the Toronto3D dataset.

Figure 5 demonstrates that LSSCC-Net (ours) more accurately characterizes complex boundaries and detailed structures, significantly enhancing segmentation capabilities for small objects, class boundaries, and large object regions. This improvement can be mainly attributed to the Spatial–Feature Dynamic Aggregation (SFDA) module in LSSCC-Net. By jointly modeling local spatial features, it strengthens local structural perception, improving segmentation of dense, confusing points near boundaries and small objects with complex geometries. Additionally, the model optimizes feature extraction and restoration through an Adaptive Attention Fusion Mechanism (AAFM), enhancing performance and generalization. A–D denote representative local regions, whose enlarged visualizations are shown in Figure 6.

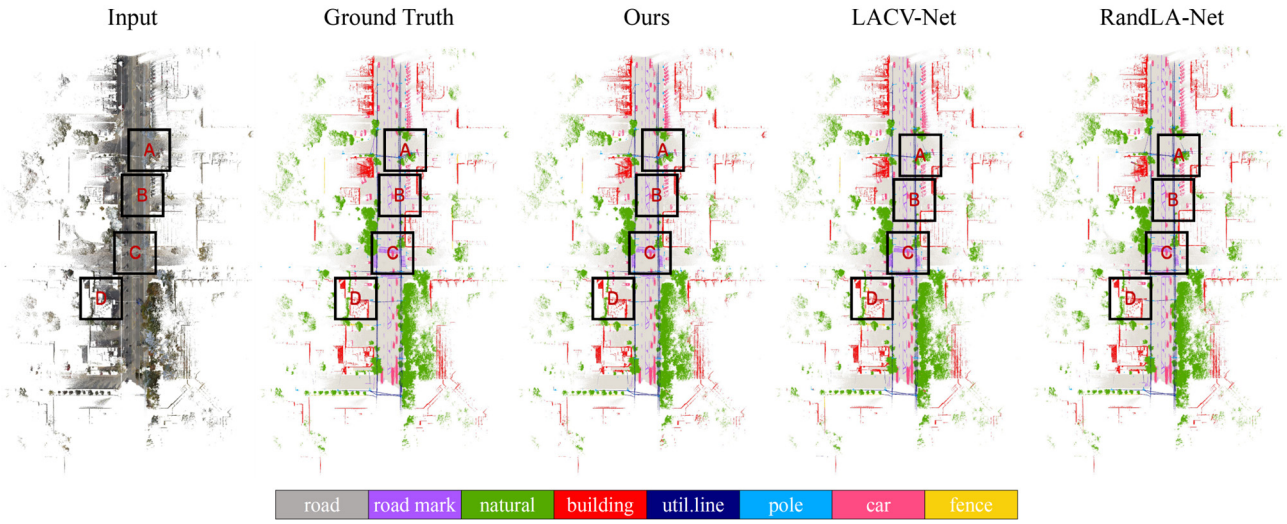

Figure 5. Result of Semantic Segmentation on Toronto3D.

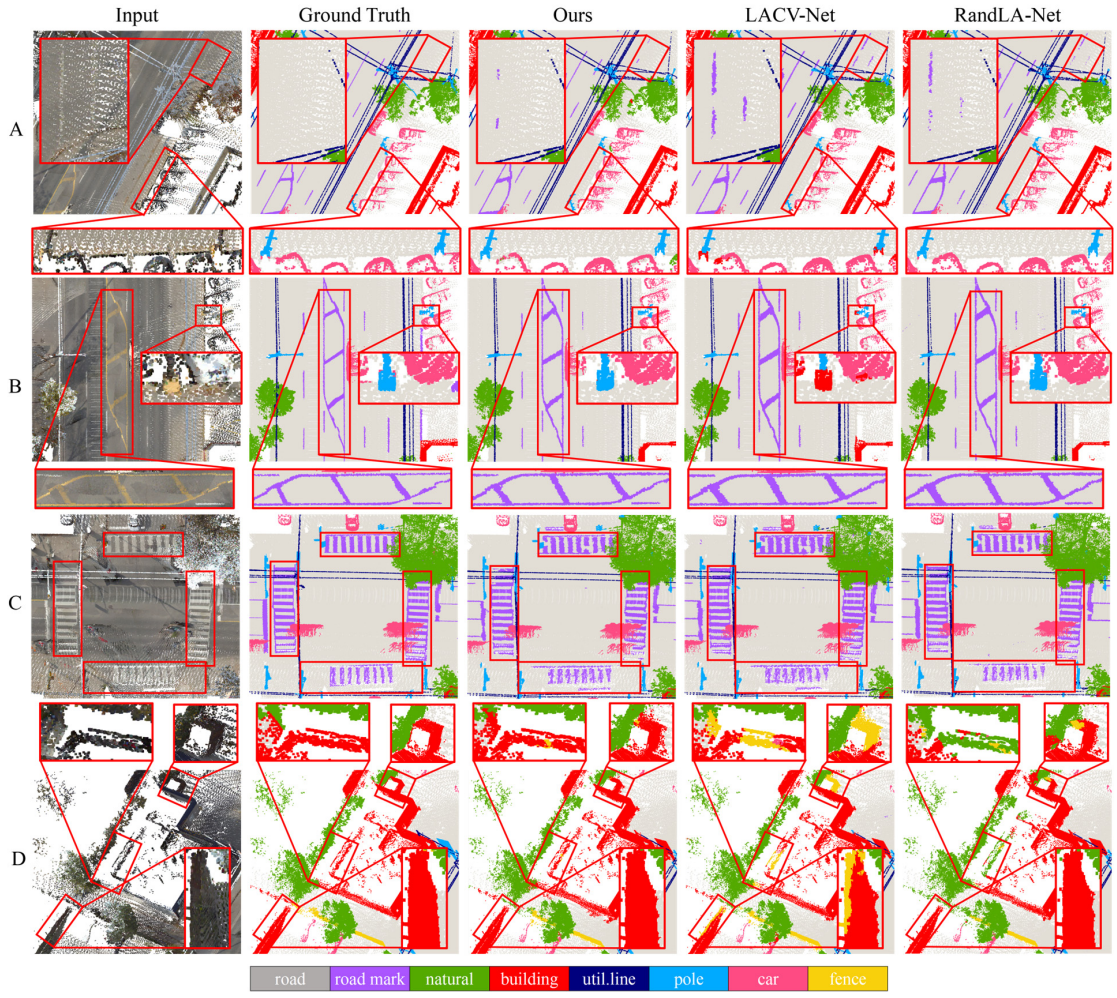

Figure 6. Local result of Semantic Segmentation on Toronto3D.

Figure 6 illustrates that LSSCC-Net outperforms LACV-Net and RandLA-Net in distinguishing small objects and accurately segmenting boundaries in the three local detailed regions. The red boxes highlight areas where LSSCC-Net provides more precise segmentation and clearer boundary delineation than the baseline LACV-Net. The experimental results show that LSSCC-Net offers detailed segmentation of boundaries between road and

road mark, as well as accurate segmentation of junctions between util.line and pole. Additionally, it excels in segmenting small objects that are often confused with other categories, such as pole bases and fence.

3.1.2. Semantic Segmentation for Indoor Scenes

The S3DIS (Stanford Large-Scale 3D Indoor Spaces) dataset, released by Stanford University, is a widely used resource for indoor scene understanding. It comprises six areas, each containing 11 room types, totaling 272 rooms. Scenes range from 0.5 to 2.5 million points, with an overall count of about 290 million points. Each point includes XYZ coordinates, RGB color data, and manually annotated labels, covering 13 semantic categories like ceiling, floor, wall, beam, and column.

We employed two evaluation methods to assess the proposed network: using Area 5 as an independent test set and conducting 6-fold cross-validation. Area 5 was chosen due to its unique characteristics, such as distinctive room layouts, object distributions, and point cloud densities, which make the segmentation results more representative. In the 6-fold cross-validation, each of the six large indoor areas in the dataset served as the test set once, with the remaining five areas used for training. This process was repeated six times, and the results were averaged to provide a comprehensive evaluation metric.

Table 2 shows that, relative to other recent top-performing methods, LSSCC-Net attains slightly lower overall accuracy (OA) than LGGCM and DG-Net, and its mean IoU (mIoU) is marginally below DG-Net. Nevertheless, the proposed network achieves the highest IoU in three categories—column, window, and board—exceeding the second-best method by 2.8%, 0.2%, and 1.5%, respectively, and it ranks second in IoU for an additional four categories. Under identical experimental settings and hyperparameters, the improved LSSCC-Net matches the baseline LACV-Net in OA (88.3%) while improving mIoU to 65.2%, a gain of 0.9%. Compared with LACV-Net, LSSCC-Net increases IoU in eight categories: ceiling (+0.1%), floor (+0.7%), wall (+0.5%), column (+8.9%), window (+0.2%), door (+5.2%), table (+0.3%), and board (+5.4%). These gains validate the effectiveness of the proposed improvements.

Notably, compared with the Toronto3D dataset, the overall performance improvement of LSSCC-Net on the S3DIS dataset was relatively limited. This difference may be attributed to intrinsic variations in point-cloud scene characteristics. A category-level analysis showed that the proposed method produced more pronounced IoU gains for slender or geometrically complex classes, such as columns, windows, and boards, indicating an advantage in modeling complex geometric structures. Furthermore, the SFDA module proved more effective at characterizing objects with clear boundaries or elongated shapes, while its contribution to large planar surfaces (e.g., floor and ceiling) was relatively limited. Overall, experimental results across both datasets suggested that LSSCC-Net had greater benefits in scenes with complex geometry, imprecise boundaries, and a high proportion of small-scale objects. This trait made it particularly effective in outdoor urban environments such as Toronto3D, whereas performance gains were comparatively modest in indoor scenes dominated by regular planar structures.

Table 3 presents the quantitative evaluation results from a 6-fold cross-validation on the S3DIS dataset. Using the experimental setup and hyperparameters specified in this study, the enhanced LSSCC-Net method showed increases of 0.4% in OA (reaching 88.5%) and 0.8% in mIoU (achieving 70.9%) compared to the baseline LACV-Net. These results suggest that the improved method exhibits robust performance and adaptability across various point cloud scenarios.

Table 2. Quantitative comparison results of different methods on S3DIS (Area 5) unit: %.

Method	Year	OA	mIoU	Ceil.	Floor	Wall	Beam	Col.	Wind.	Door	Table	Chair	Sofa	Book.	Board	Clut.
RandLA-Net [15]	2020	87.2	62.4	91.1	95.6	80.2	0.0	24.7	62.3	47.7	76.2	83.7	60.2	71.1	65.7	53.8
SCF-Net [21]	2021	87.2	63.7	90.8	97.0	80.9	0.0	19.9	60.7	44.6	79.4	87.9	**76.1**	**71.5**	68.8	50.4
DPFA-Net [34]	2022	88.0	55.2	93.0	**98.6**	80.2	0.0	14.7	55.8	42.8	72.3	73.5	27.3	55.9	53.0	50.5
LGGCM [35]	2024	88.8	63.3	**94.8**	98.3	81.5	0.0	35.9	63.3	43.5	80.2	**88.4**	68.8	55.8	64.6	47.8
DG-Net [2]	2024	**90.2**	**65.9**	93.1	97.8	**82.6**	0.0	33.1	62.2	**52.3**	**80.4**	87.5	72.3	70.7	69.8	54.5
HFA-Net [23]	2025	88.0	63.0	90.9	96.9	79.3	0.0	21.0	63.1	37.6	76.5	86.5	75.9	67.0	72.5	52.1
RandLA-Net *	-	86.6	62.5	92.8	97.4	78.8	0.0	33.7	60.2	33.3	75.8	86.6	68.8	71.2	62.6	50.7
LACV-Net *	-	88.3	64.3	93.0	97.1	81.8	0.0	29.8	63.5	37.7	77.7	87.3	72.0	71.3	68.6	**55.8**
LSSCC-Net (Ours) *	-	88.3	65.2	93.1	97.8	82.3	0.0	**38.7**	**63.7**	42.9	78.0	86.0	67.3	69.7	**74.0**	54.6

Bold numbers denote the best performance for each metric, while underscored numbers (_) indicate the second-best performance. Methods marked with an asterisk (*) were trained and tested using the experimental environment and hyperparameters specified in this study. Performance results for other methods were sourced from publicly available literature.

Table 3. Quantitative comparison results of different methods on S3DIS (six-fold cross-validation) unit: %.

Method	Year	OA	mIoU	Ceil.	Floor	Wall	Beam	Col.	Wind.	Door	Table	Chair	Sofa	Book.	Board	Clut.
PointNet [13]	2017	78.6	47.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
RandLA-Net [15]	2020	88.0	70.0	93.1	96.1	80.6	62.4	48.0	64.4	69.4	69.4	76.4	60.0	64.2	65.9	60.1
BAF-LAC [36]	2021	88.2	71.7	92.5	95.9	81.3	63.2	57.8	63.0	79.9	70.3	74.6	60.6	67.2	65.3	60.4
SCF-Net [21]	2021	88.4	71.6	93.3	96.4	80.9	64.9	47.4	64.5	70.1	71.4	81.6	67.2	64.4	67.5	60.9
DPFA-Net [34]	2022	89.3	61.7	94.6	98.0	79.2	40.7	36.6	52.2	70.8	65.9	74.7	27.7	49.8	51.6	60.6
LACV-Net	2024	89.7	72.7	94.5	96.7	82.1	65.2	48.6	69.3	71.2	72.7	78.1	67.3	67.2	70.9	61.6
LACV-Net *	-	88.1	70.1	93.9	96.6	81.0	56.4	50.1	65.3	68.9	70.8	72.1	65.8	64.5	65.8	60.2
LSSCC-Net (Ours) *	-	88.5	70.9	93.7	96.8	80.9	60.4	50.9	66.0	68.3	71.4	81.3	59.9	63.7	67.1	61.8

Methods marked with an asterisk (*) were trained and tested using the experimental environment and hyperparameters specified in this study. Performance results for other methods were sourced from publicly available literature.

Figure 7 illustrates the segmentation outcomes of RandLA-Net, LACV-Net, and LSSCC-Net on Area 5 of the S3DIS dataset. Red boxes highlight areas where LSSCC-Net surpasses LACV-Net in segmentation accuracy. The results show that LSSCC-Net excels in differentiating walls from columns, accurately segments similar categories, and enhances boundary segmentation between distinct objects. The performance advantage primarily stems from LSSCC-Net's inclusion of an additional SFDA module during the encoding phase. This module extracts offset information from both spatial positions and feature channels, facilitating interaction between these dimensions. Consequently, it effectively resolves semantic ambiguity in neighboring points.

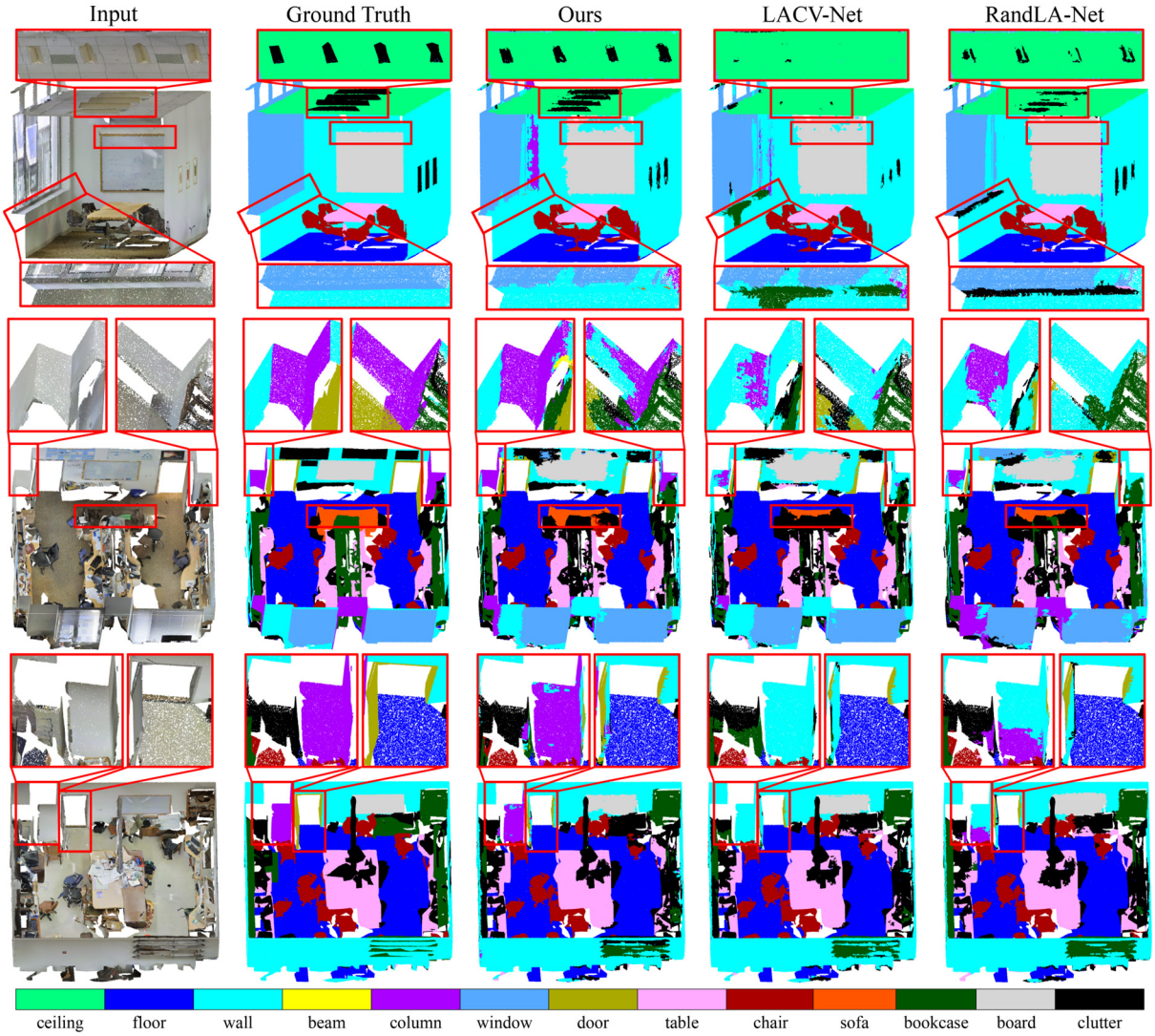

Figure 7. Result of Semantic Segmentation on S3DIS.

3.2. Ablation Experiments

To analyze the contributions of individual components in LSSCC-Net, an ablation study was conducted focusing on three key components: the SFDA module, the AAFM, and the Lovász-Softmax Loss. These experiments were carried out using the Toronto3D dataset, adhering to the study's specific experimental environment and hyperparameters. In line with previous experiments, Region 2 served as the test set, while the other three regions were used for training.

Table 4 displays the results of the ablation experiments, with bold numbers indicating the best performance for each metric. The table uses the following notations: A: This

denotes the removal of the SFDA module. Initially, features in each encoding layer were processed separately by the Local Adaptive Feature Augmentation (LAFA) module and the SFDA module, then combined. Without the SFDA module, features are processed solely by the LAFA module and directly input into the Spatial Attention Module (SAM) for further processing. B: This indicates the removal of the AAFM. It involves eliminating the Spatial Attention Module (SAM) in the encoding layers and the Channel Attention Module (CAM) in the decoding layers, thereby halting the optimization of feature extraction and restoration. C: This refers to the removal of the Lovász-Softmax Loss, retaining only the aggregation loss function of the baseline network.

Table 4. Ablation Studies about LSSCC-Net Network unit: %.

Method	SFDA	AAFM	Lovász	OA	mIoU	Road	Road Mark	Natural	Building	Util.line	Pole	Car	Fence
A		√	√	**97.8**	83.4	97.6	71.0	**97.4**	**93.9**	86.3	**82.2**	95.2	43.6
B	√		√	97.7	83.2	97.6	71.5	97.1	93.5	**88.1**	79.7	**95.3**	43.1
C	√	√		97.6	82.7	97.4	69.6	97.2	**93.9**	86.2	78.8	93.8	44.9
LSSCC-Net	√	√	√	97.7	**83.6**	**97.7**	**72.4**	97.0	92.3	87.4	80.5	94.2	**47.2**

Bold numbers indicate the best performance for each metric. The symbol √ indicates that the corresponding module or loss function is enabled, while its absence indicates that the component is not used.

Table 4 reveals several key observations. When the SFDA module is removed from LSSCC-Net, Model A experiences a 0.2% decrease in mean Intersection over Union (mIoU). Specifically, the IoU for road mark and fence drops by 1.4% and 3.6%, respectively, while the IoU for building and pole increases by 1.6% and 1.7%. This result suggests that the SFDA module enhances local contextual representation, particularly for objects with ambiguous boundaries and complex local geometries. In contrast, removing the AAFM from LSSCC-Net results in Model B showing a 0.4% decline in mIoU. The IoU for road mark decreases by 0.9%, and for fence, it drops by 4.1%. These findings underscore the importance of the AAFM in improving segmentation accuracy for these features. The AAFM enhances the model's ability to identify key regions by integrating the Spatial Attention Module (SAM) in the encoder and the Channel Attention Module (CAM) in the decoder. These modules jointly optimize feature extraction and restoration. In comparison to LSSCC-Net, Model C, which excludes the Lovász-Softmax Loss and retains only the baseline network's aggregation loss function, shows a 0.9% decrease in mIoU. Specifically, the IoU values for road mark, util.line, pole, and fence decrease by 2.8%, 1.2%, 1.7%, and 2.3%, respectively. This indicates that incorporating Lovász-Softmax Loss as an auxiliary loss function complements the original aggregation loss function, thereby enhancing the model's segmentation capability for small targets and class boundaries.

3.3. Efficiency Analysis

We evaluated the computational efficiency of the proposed model on the Toronto3D dataset using per-batch training time and total inference time as efficiency metrics, while OA and mIoU served as performance indicators. All experiments used a batch size of 4. The results are summarized in Table 5.

Table 5. Comparison of computational efficiency on Toronto3D.

Method	Per-Batch Time (ms)	Total Inference Time (s)	OA (%)	mIoU (%)
RandLA-Net	286.35	53.00	97.4	82.1
LACV-Net	322.37	58.67	97.3	81.5
LSSCC-Net (Ours)	496.82	69.66	97.7	83.6

Table 5 shows that LSSCC-Net achieves the best segmentation performance at the cost of reduced efficiency. Its per-batch training time was 496.82 ms and total inference time was 69.66 s, both longer than those of the baseline LACV-Net (322.37 ms and 58.67 s, respectively). The efficiency gap mainly results from the structural enhancements in LSSCC-Net: dual-dimensional offset learning in the Spatial–Feature Dynamic Aggregation (SFDA) module and the symmetric deployment of attention mechanisms in the Adaptive Attention Fusion Mechanism (AAFM) increase computational complexity, thereby raising both training and inference costs. Notably, although LSSCC-Net incurs higher per-batch computation time than LACV-Net, it converges in fewer training epochs. On the Toronto3D dataset, LSSCC-Net reached its optimal mIoU within 62 epochs, a 37.4% reduction in training epochs compared with LACV-Net (99 epochs). Similarly, on the S3DIS dataset, LSSCC-Net converged in 58 epochs, a 30.9% reduction relative to LACV-Net (84 epochs).

4. Discussion and Limitations

The efficacy of LSSCC-Net has been validated through experimental results, which demonstrate notable improvements in point cloud semantic segmentation, particularly for complex geometries and small-scale objects. However, certain limitations remain. Compared with the baseline network, the introduction of the Spatial–Feature Dynamic Aggregation (SFDA) module and the Adaptive Attention Fusion Mechanism (AAFM) into the encoder–decoder architecture increases the computational complexity and inference cost of the model. Moreover, experimental results indicate that the proposed approach yields more substantial advantages in outdoor urban environments with complex geometric layouts, while its overall benefits in indoor scenes characterized by regular planar structures are relatively limited. This observation highlights the model's limited adaptability to diverse scene structures. Future work will focus on simplifying the model architecture to reduce computational overhead, exploring weakly supervised learning strategies to decrease reliance on high-quality manual annotations, and further improving the model's generalization capability across different scene types.

5. Conclusions

To tackle semantic segmentation in large-scale point cloud scenes, this study enhanced the LACV-Net baseline, introducing the LSSCC-Net architecture. Key contributions and implementation details include the following: First, the Spatial-Feature Dynamic Aggregation (SFDA) module was added during the encoding phase. This module extracts offset information through symmetric interaction between spatial positions and feature channels, enhancing the local contextual representation of point clouds. Consequently, the proposed approach improves robustness to variations in spatial structures. Subsequently, an Adaptive Attention Fusion Mechanism (AAFM) was developed to improve model performance. During the encoding phase, a Spatial Attention Module (SAM) was integrated to emphasize and enhance significant regions. In the decoding phase, a Channel Attention Module (CAM) was incorporated to improve feature expression comprehensively. This design combines the strengths of both attention mechanisms to optimize feature extraction and restoration, thereby enhancing overall performance and generalization. Finally, Lovász-Softmax Loss was introduced as an auxiliary optimization objective, creating a framework focused on IoU-oriented loss function collaboration. This framework enhances the model's ability to segment small objects, delineate class boundaries, and accurately capture complete regions of large objects.

This paper conducted experiments using the outdoor point cloud dataset Toronto3D and the indoor dataset S3DIS to assess the proposed model. For Toronto3D, using Region 2 as the test set, LSSCC-Net achieved an overall accuracy (OA) of 97.7% and a

mean Intersection-over-Union (mIoU) of 83.6%. For S3DIS, with Area 5 as the test set, it achieved an OA of 88.3% and an mIoU of 65.2%. Quantitative comparisons with recent state-of-the-art methods indicate that LSSCC-Net is effective in segmenting small-scale objects with complex geometric structures and produces more accurate boundary delineation across different object categories, leading to improved overall segmentation performance.

Author Contributions: Conceptualization, W.W. and X.H.; Methodology, Validation, Formal Analysis and Writing—Original Draft, W.W. and C.L.; Data Curation, Y.W. and L.L.; Reviewing, Editing, Language Polishing and Providing Revision Suggestions, X.H., P.T. and Y.W.; Supervision, X.H.; Funding Acquisition, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC), grant numbers 42271452 and 42301486.

Data Availability Statement: All datasets utilized in this work are publicly accessible. The Toronto3D dataset [30] is available at https://github.com/WeikaiTan/Toronto-3D (accessed on 10 May 2025). The S3DIS dataset [31] is available at https://docs.google.com/forms/d/e/1FAIpQLScDimvNMCGhy_rmBA2gHfDu3naktRm6A8BPwAWWDv-Uhm6Shw/viewform?c=0&w=l (accessed on 2 April 2025).

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

References

1. Sun, M.; Rui, T.; Wang, D.; Yang, C.; Zheng, N. ELMF-Net: Semantic Segmentation of Large-Scale Point Clouds via Efficient Local Feature Learner and Multiscale Fusion. *IEEE Sens. J.* **2025**, *25*, 11392–11404. [CrossRef]
2. Liu, T.; Ma, T.; Du, P.; Li, D. Semantic Segmentation of Large-Scale Point Cloud Scenes via Dual Neighborhood Feature and Global Spatial-Aware. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103862. [CrossRef]
3. Tatarchenko, M.; Park, J.; Koltun, V.; Zhou, Q.-Y. Tangent Convolutions for Dense Prediction in 3D. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3887–3896.
4. Du, J.; Cai, G.; Wang, Z.; Huang, S.; Su, J.; Marcato Junior, J.; Smit, J.; Li, J. ResDLPS-Net: Joint Residual-Dense Optimization for Large-Scale Point Cloud Semantic Segmentation. *ISPRS J. Photogramm. Remote Sens.* **2021**, *182*, 37–51. [CrossRef]
5. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-View Convolutional Neural Networks for 3D Shape Recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 945–953.
6. Lawin, F.J.; Danelljan, M.; Tosteberg, P.; Bhat, G.; Khan, F.S.; Felsberg, M. Deep Projective 3D Semantic Segmentation. In *Computer Analysis of Images and Patterns*; Felsberg, M., Heyden, A., Krüger, N., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2017; Volume 10424, pp. 95–107, ISBN 978-3-319-64688-6.
7. Wei, X.; Yu, R.; Sun, J. View-GCN: View-Based Graph Convolutional Network for 3D Shape Analysis. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1847–1856.
8. Lyu, Y.; Huang, X.; Zhang, Z. EllipsoidNet: Ellipsoid Representation for Point Cloud Classification and Segmentation. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 256–266.
9. Maturana, D.; Scherer, S. VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 922–928.
10. Riegler, G.; Ulusoy, A.O.; Geiger, A. OctNet: Learning Deep 3D Representations at High Resolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6620–6629.
11. Cheng, R.; Agia, C.; Ren, Y.; Li, X.; Bingbing, L. S3CNet: A Sparse Semantic Scene Completion Network for LiDAR Point Clouds. In *Proceedings of Machine Learning Research, Proceedings of the Conference on Robot Learning, Virtual, 16–18 November 2020*; JMLR: Cambridge, MA, USA, 2020; Volume 155, pp. 2148–2161.
12. Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Ma, Y.; Li, W.; Li, H.; Lin, D. Cylindrical and Asymmetrical 3D Convolution Networks for LiDAR Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9934–9943.

13. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 77–85.
14. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In Proceedings of the NeurIPS, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5105–5114.
15. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11105–11114.
16. Zeng, Z.; Xu, Y.; Xie, Z.; Tang, W.; Wan, J.; Wu, W. Large-Scale Point Cloud Semantic Segmentation via Local Perception and Global Descriptor Vector. *Expert Syst. Appl.* **2024**, *246*, 123269. [CrossRef]
17. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. Graph.* **2019**, *38*, 1–12. [CrossRef]
18. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.S.; Koltun, V. Point Transformer. In Proceedings of the ICCV, Montreal, QC, Canada, 10–17 October 2021; pp. 16259–16268.
19. Wu, X.; Lao, Y.; Jiang, L.; Liu, X.; Zhao, H. Point Transformer V2: Grouped Vector Attention and Partition-Based Pooling. In Proceedings of the NeurIPS, New Orleans, LA, USA, 28 November–9 December 2022.
20. Wu, X.; Jiang, L.; Wang, P.-S.; Liu, Z.; Liu, X.; Qiao, Y.; Ouyang, W.; He, T.; Zhao, H. Point Transformer V3: Simpler, Faster, Stronger. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 4840–4851.
21. 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 14499–14508.
22. Li, Y.; Ye, Z.; Huang, X.; HeLi, Y.; Shuang, F. LCL_FDA: Local Context Learning and Full-Level Decoder Aggregation Network for Large-Scale Point Cloud Semantic Segmentation. *Neurocomputing* **2025**, *621*, 129321. [CrossRef]
23. Wen, C.; Zhang, L.; Ren, J.; Hong, R.; Li, C.; Yang, C.; Lv, Y.; Chen, H.; Yang, N. HFA-Net: Hybrid Feature-Aware Network for Large-Scale Point Cloud Semantic Segmentation. *Artif. Intell. Rev.* **2025**, *58*, 102. [CrossRef]
24. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9598–9607.
25. Zhang, C.; Wan, H.; Shen, X.; Wu, Z. PVT: Point-voxel Transformer for Point Cloud Learning. *Int. J. Intell. Syst.* **2022**, *37*, 11985–12008. [CrossRef]
26. Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; Cui, S. Sparse Single Sweep LiDAR Point Cloud Segmentation via Learning Contextual Shape Priors from Scene Completion. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 3101–3109. [CrossRef]
27. Wei, Y.; Li, Y.; Wang, Y.; Ji, C.; Jin, X.; Xu, C. Enhanced 3D Semantic Segmentation via Local Polar Encoding and Attention Fusion. In Proceedings of the 2025 Joint International Conference on Automation-Intelligence-Safety (ICAIS) & International Symposium on Autonomous Systems (ISAS), Xi'an, China, 23–25 May 2025; pp. 1–8.
28. Xu, Y.; Tang, W.; Zeng, Z.; Wu, W.; Wan, J.; Guo, H.; Xie, Z. NeiEA-NET: Semantic Segmentation of Large-Scale Point Cloud Scene via Neighbor Enhancement and Aggregation. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *119*, 103285. [CrossRef]
29. Berman, M.; Triki, A.R.; Blaschko, M.B. The Lovasz-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4413–4421.
30. Tan, W.; Qin, N.; Ma, L.; Li, Y.; Du, J.; Cai, G.; Yang, K.; Li, J. Toronto-3D: A Large-Scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 797–806.
31. Armeni, I.; Sener, O.; Zamir, A.R.; Jiang, H.; Brilakis, I.; Fischer, M.; Savarese, S. 3D Semantic Parsing of Large-Scale Indoor Spaces. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1534–1543.
32. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Learning Semantic Segmentation of Large-Scale Point Clouds with Random Sampling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *11*, 8338–8354. [CrossRef] [PubMed]
33. Ji, H.; Yang, S.; Jiang, Z.; Zhang, J.; Guo, S.; Li, G.; Zhong, S.; Liu, Z.; Xie, Z. BEMF-Net: Semantic Segmentation of Large-Scale Point Clouds via Bilateral Neighbor Enhancement and Multi-Scale Fusion. *Remote Sens.* **2023**, *15*, 5342. [CrossRef]
34. Chen, J.; Kakillioglu, B.; Velipasalar, S. Background-Aware 3-D Point Cloud Segmentation with Dynamic Point Feature Aggregation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]

35. Du, Z.; Ye, H.; Cao, F. A Novel Local–Global Graph Convolutional Method for Point Cloud Semantic Segmentation. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 4798–4812. [CrossRef] [PubMed]
36. Shuai, H.; Xu, X.; Liu, Q. Backward Attentive Fusing Network with Local Aggregation Classifier for 3D Point Cloud Semantic Segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 4973–4984. [CrossRef] [PubMed]