Point Cloud Semantic Segmentation Enhanced by Self-Training Feature Fusion

Xue Wang, Yubo Zhou, Xinyu Wang, Kun Tan, *Senior Member, IEEE*, and Yong Mei

***Abstract*—LiDAR remote sensing technology, owing to its high-precision laser pulse measurement characteristics, is capable of high-accuracy dynamic perception and modeling of 3-D spatial targets. The significant improvement in the accuracy of models based on the PointNet++ architecture has been primarily attributed to the expansion of the model size, the enhancement of the receptive fields, and the optimization of the training strategies. Inspired by the PointNeXt model, this article revisits the traditional local feature aggregation process, enhancing the receptive field of the network through an implicit decoupling approach, and jointly exploring multidimensional vector feature weighting to delve deeper into the discriminative features of the imagery. We introduce a pseudo-label generation strategy and integrate a multiclass weighted loss to achieve a comprehensive understanding of point cloud semantics and spatial information through a novel point cloud segmentation network, receptive field extension, and vector feature enhancement network (RV-Net). Compared to recent advanced fully supervised networks, RV-Net demonstrates competitive results on multiple benchmark datasets. Furthermore, experiments conducted on multiple benchmark datasets with 1% and 0.1% labeling ratios demonstrate that the proposed model outperforms various fully and weakly supervised segmentation algorithms. The results indicate that the network model proposed in this article exhibits a superior performance in the semantic segmentation of point clouds in large-scale indoor and outdoor scenarios.**

***Index Terms*—Deep learning, laser point cloud semantic, point cloud application, weakly supervised learning.**

Received 25 November 2025; revised 23 February 2026 and 11 April 2026; accepted 10 May 2026. Date of publication 18 May 2026; date of current version 9 June 2026. This work was supported in part by the Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project under Grant 2024CSJZN01300, in part by Shanghai Municipal Education Commission Science and Technology Project under Grant 2024AI02002, in part by the National Natural Science Foundation of China under Grant 42171335, in part by the National Civil Aerospace Project of China under Grant D040102, and in part by Shanghai Municipal Science and Technology Natural Science Foundation Project under Grant 25ZR1401105. *(Corresponding author: Kun Tan.)*

Xue Wang is with the Key Laboratory of Geographic Information Science, Ministry of Education, the School of Geographic Sciences, the School of GeoAI and the Hindon STAI Institute, East China Normal University, Shanghai 200241, China (e-mail: wx_ecnu@yeah.net).

Yubo Zhou is with the Key Laboratory of Geographic Information Science, Ministry of Education, and the School of Geographic Sciences, East China Normal University, Shanghai 200241, China (e-mail: 1540114108@qq.com).

Xinyu Wang and Kun Tan are with the Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), the School of Geospatial Artificial Intelligence, the School of GeoAI, and the Hindon STAI Institute, East China Normal University, Shanghai 200241, China (e-mail xy1163sl@163.com; tankuncu@gmail.com).

Yong Mei is with the Institute of Defense Engineering, AMS, Beijing 100036, China (e-mail: meiyong1990@126.com).

Digital Object Identifier 10.1109/TGRS.2026.3694418

I. Introduction

LiDAR remote sensing can accurately determine the spatial coordinates of targets through laser pulses and has the ability to dynamically perceive and model the 3-D space of targets with high precision [1], providing key data support in fields such as urban modeling and autonomous driving. Point cloud semantic segmentation, based on the theories and methods of machine vision and pattern analysis, intelligently assigns a semantic label to the unordered spatial point cloud and segments the spatial point cloud with the same attributes. Point cloud semantic segmentation operates under both fully and semisupervised conditions and is achieved by combining local and global features [4]. Due to the nonstructured nature of point clouds, they cannot be directly processed by 2-D convolutional neural networks (CNNs), leading to the development of point cloud segmentation algorithms based on voxel representation and multiview projection representation [3]. However, these methods inevitably result in information loss, making it a significant challenge for neural networks to obtain local and global information. Currently, point cloud semantic segmentation is mainly based on the representation of raw point clouds.

To address the irregularity of the data, the existing segmentation methods based on raw point clouds mainly construct local neighborhoods through spherical queries or k-nearest neighbor, and then extract local and global information by superimposing multiple relationship vectors [4]. PointNet++ [5] achieves feature extraction of point clouds through a multilayer perceptron (MLP), pioneering the research paradigm of semantic segmentation based on raw point clouds. A series of networks derived from PointNet++ constructs relationship vectors artificially, uses convolution with shared weights to capture local features from neighboring points, and gradually expands the receptive field by uniformly downsampling the point cloud using farthest point sampling (FPS), obtaining neural networks for analyzing 3-D point clouds. Inspired by CNNs, subsequent studies implicitly and adaptively constructed relationship vectors to flexibly handle different local morphologies and applied symmetric aggregation operations to obtain local morphologies.

Whether the relationship vectors are artificially constructed or self-adaptively constructed, this process can be understood as a neighborhood aggregation operation. Similar to a CNN, neighborhood aggregation encounters challenges in handling long-range dependencies and requires the stacking of multilevel modules to achieve long-range pattern perception.

Algorithms represented by the transformer model [6], [7] have excellent long-range capture capabilities, but due to the extremely large number of point clouds, the computational load increases sharply. Therefore, point cloud transformer models are still limited to local regions [8] and need to be combined with local aggregation to achieve point cloud segmentation. Subsequent research mainly evolved along two technical routes: First, MLP enhancement methods represented by PointNeXt [9], [10], which expand the network depth and optimize the data augmentation strategies, verify the decisive role of the model size and training techniques in performance improvement. Second, methods based on the transformer model attempt to capture long-range dependencies through improved self-attention mechanisms, while reducing the computational load. One such example is PointNAT [11], which fuses a point proxy transformer with multi-neighborhood RandLA-Net modules [12] to seek a balance between global context modeling and local feature extraction.

Through the research on MLP and transformer networks, it can be found that the receptive field of the network has a significant impact on the performance of point cloud segmentation. The perception of global information by a single point cloud can be achieved by increasing the depth of the network. However, there is information loss in the implicit capture of local structures in high-dimensional space, and the network has difficulty in extracting discriminative information. In this article, we divide the above problems that affect the performance improvement of network segmentation into three subproblems: 1) how to effectively model the global information of point clouds; 2) how to fully mine the discriminative information of local regions; and 3) how to efficiently fuse global and local information.

Motivation: To address the aforementioned issues, the proposed approach starts from the expansion of the network receptive field and, by stacking multiple deep local aggregation modules, acquires both local and abstract global information while extracting local information. Meanwhile, a vector-based local feature extraction module is constructed to reduce the information loss during the implicit capture of local structures by the network and to extract discriminative information with differences. Ultimately, an adaptive weighting module is utilized to allocate mixed weights to alleviate the limitations of each fixed weight. We propose a novel neural network for point cloud semantic segmentation—RV-Net—which consists of three modules. Specifically, we introduce a decoupled local graph operation module, the graph-edge-coupled feature encoding (GEC) module, which enhances the network's receptive field by stacking multiple instances and using a larger KNN range, thereby capturing global information with a relatively low computational cost. Subsequently, we propose a vector-based fine feature mining vectorized feature fusion (VFF) module, which deeply separates significant discriminative information through vector-scalar conversion, thereby improving the network's ability to extract local discriminative information. To fully leverage the advantages of the two modules, an adaptive module is proposed to assign mixed weights to the local and global information to mitigate their respective limitations. Finally, a feature projection (FP) sample generation module is constructed to utilize noisy samples to promote the network's learning of more robust feature boundaries, thereby reducing the probability of overfitting and the problem of sample imbalance. We conducted experiments on two public datasets, the Stanford large-scale 3-D indoor spaces dataset (S3DIS) [13] and the Toronto 3-D dataset (Toronto-3D) [14], demonstrating the effectiveness of the proposed method, compared to the existing approaches. We also conducted ablation experiments on both datasets to evaluate the impact of each module and the rationality of the network's detailed settings.

Point cloud data annotation is difficult. To reduce annotation costs, scholars have proposed a series of weakly supervised learning methods. Weak supervision uses only a small number of labels in all the training point clouds to achieve a segmentation accuracy that is similar to that of fully supervised learning. However, weakly supervised learning puts forward higher requirements for the generalization ability and training efficiency of the network. Hu et al. [15] proposed the semantic query network (SQN), Tran et al. [16] proposed the PointCT network, and Zhang and Bi [17] proposed the DR-Net network. All of these methods start from feature transfer between the labeled points and neighboring points, and achieve efficient point cloud segmentation by expanding the local receptive fields of multilevel neighboring point clouds. Wang et al. [18] proposed the one-class one-click (OCOC) model, and Li et al. [19] proposed the Hybrid-CR model, which generates pseudo-labels for unlabeled points based on a small amount of labeled data for the subsequent training processes, and also achieves a high point cloud segmentation accuracy. In the proposed approach, RV-Net achieves feature transfer between neighboring point clouds by expanding the receptive fields. At the same time, the introduced noisy labels can be regarded as pseudo-labels in the weakly labeled situation. As a result, the RV-Net model proposed in this article has a high segmentation efficiency in weakly supervised scenarios. In addition, two pseudo-label generation modules are added, namely two-stage learning and deep feature matching, on the basis of RV-Net, further improving the segmentation efficiency of the backbone network under the condition of sparse samples. To verify the effectiveness of the network, we conducted experiments on two open-source datasets, namely the S3DIS and the Dayton annotated LiDAR dataset (DALES) [20], and verified the effectiveness of the algorithm in segmentation performance under 1% and 0.1% annotation ratios.

II. Related Work

A. Fully Supervised Point Cloud Segmentation

1) Voxelization-Based Methods: The original point cloud data are characterized by disorder and nonstructured features. To apply a CNN to laser point cloud data, researchers have adopted voxelization to regularize the arrangement of the data. Le and Duan [21] and Peng et al. [22] conducted preliminary explorations of the semantic segmentation of voxelized laser point cloud data. However, compared with the original point cloud data, the voxelization process inevitably leads to the loss of local detail information. Moreover, due to the limitation of

computing performance, it is difficult to increase the resolution of the voxel block, which restricts the improvement of the segmentation accuracy of this type of method.

2) Projection-Based Methods: The multiview projection approach can effectively transform unstructured and disordered data into structured and ordered data. It can quickly extract features from 3-D point clouds using 2-D convolution. Kalogerakis et al. [23], Luo et al. [24], and others have successively carried out point cloud segmentation based on multiview projection. Similar to the voxelization-based methods, multiview algorithms can encounter problems such as local information loss, spatial structure loss, and foreground occlusion during the projection process. They cannot fully utilize the 3-D spatial information of point clouds, and the segmentation results are greatly affected by the projection viewpoints. These methods are mainly applied in the multisensor data fusion of autonomous driving, to supplement the long-distance depth information missing from image sensors.

3) Point-Based Methods: The original point cloud approach bypasses the geometric distortion of multiview projection and avoids the spatial detail degradation caused by the voxelization (discretization) process, demonstrating unique advantages in fine-grained semantic segmentation of large-scale scenes. The development of this segmentation approach can be traced back to the PointNet network proposed by Qi et al. [5], which achieves point cloud feature extraction through an MLP and constructs a rotation-invariant transformation matrix to address the unordered nature of point clouds. Subsequently, Qi's team proposed the PointNet++ network [25], which significantly improves the segmentation accuracy by adopting a hierarchical encoder–decoder structure and a multiresolution feature fusion strategy. On this basis, scholars have successively proposed various improved methods. Li et al. [26] designed the PointCNN network, which introduces a χ transformation mechanism to solve the unordered nature of point clouds and achieve the adaptive generation of local convolution kernels. Wu et al. [27] constructed a nonlinear feature extractor with probabilistic statistical characteristics and designed the PointConv network. The KPConv method proposed by Thomas et al. [28] achieves geometric adaptive adjustment of the convolution kernel through the design of learnable kernel points. With the increase in model complexity, researchers have begun to reflect on the efficiency of feature extraction. Hu et al. [12] constructed the RandLA-Net network, which achieves efficient processing of large-scale point clouds through random downsampling. Ma et al. [29] proposed the PointMLP network, which expands the receptive field through a deep residual network to replace fine local feature operations, verifying the effectiveness of deep networks. Qian et al. [9] verified the relationship between model complexity and performance through systematic experiments and developed PointNeXt, which balances accuracy and efficiency by optimizing the residual connections, separable MLPs, and training strategies, while maintaining the basic architecture of PointNet++. Zhao et al. [8] were the first to introduce the transformer structure into point clouds, and their proposed Point-Transformer model significantly enhances long-range dependency and modeling capabilities. Subsequent versions of point cloud transformers [30], [31] have continuously optimized the geometric feature extraction and computational efficiency. In recent years, research on large 3-D models has begun to emerge. The large Ponder model of Zhu et al. [32] constructs multimodal 3-D representations through neural rendering technology and has set new performance records in multiple downstream tasks. Despite the continuous evolution of algorithm architectures, the original point cloud-based semantic segmentation approach remains a research focus in fine-grained 3-D scene understanding, due to its ability to retain complete geometric information.

B. Weakly Supervised Point Cloud Segmentation

1) Inaccurate Weakly Supervised Learning: Inaccurate weakly supervised learning achieves this by continuously training and refining through the annotation of subclouds. The network proposed by Wei et al. [33] is a representative work of inaccurate weakly supervised learning. This network builds a multipath regional information mining module, generating pseudo-labels based on the activation maps of various categories, combined with an attention mechanism, and ultimately achieves weakly supervised point cloud semantic segmentation by combining the pseudo-labels and using a fully supervised training mode. Liu et al. [34] proposed the one-thing one-click (OTOC) method, which divides instance objects into multiple parts through an oversegmentation strategy and then annotates each instance part with a single sample. However, since this method generates point-wise pseudo-labels from subcloud annotations, the generated pseudo-labels lack local detail information, and the segmentation accuracy is somewhat limited, making this method only suitable for point cloud segmentation in extremely sparse scenes.

2) Incomplete Weak Supervision: Incomplete weakly supervised learning achieves effective improvement of the overall model capability using partial samples. There are three main types of methods: consistency regularization, pseudo-labeling, and contrastive learning. Consistency regularization bridges the gap between dual-branch networks by constructing a loss function, deeply exploring the additional information brought by perturbations to the segmentation network, and enhancing the learning efficiency of the network when a small number of samples are introduced. Zhang et al. [35], Li et al. [19], Wang and Yao [36], and others have adopted consistency strategies to enhance the segmentation efficiency of the network. Consistency regularization methods rely on the added perturbations to expand the supervision signal, and are thus significantly influenced by the perturbation strategy and parameters, which limits their generalization performance. The basic idea of the pseudo-label self-training methods is to train an initial model with sparse labeled data and then generate pseudo-labels for the unlabeled data based on the trained model. Li et al. [19] proposed the HybridCR network, which generates pseudo-labels through a confidence threshold. Lan et al. [37] proposed the receptive-driven pseudo-label consistency and structural consistency (RPSC) framework, which constructs a hierarchical consistency verification strategy to optimize the pseudo-label generation process and improve the confidence of the pseudo-labels. Su et al. [38] used adaptive reweighting

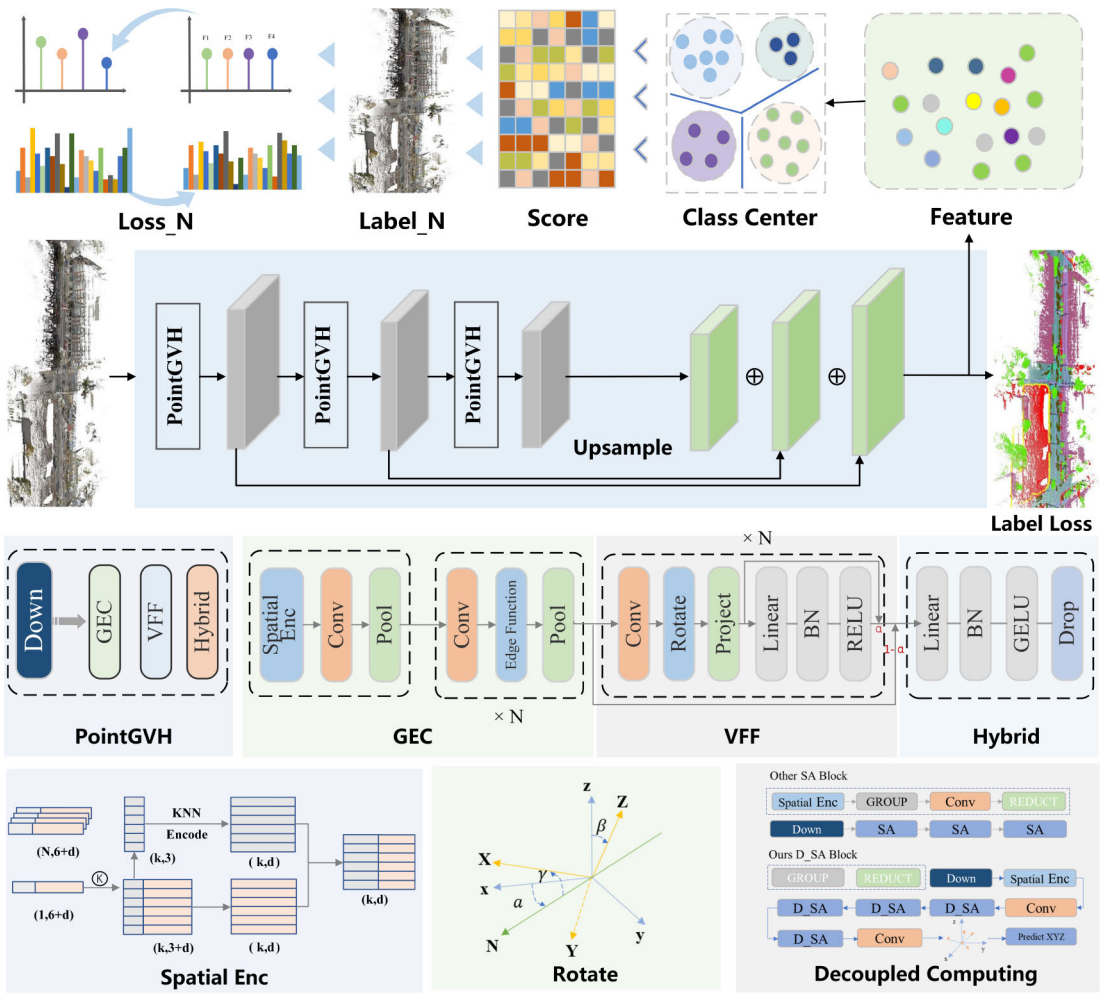

Fig. 1. Architecture of the proposed RV-Net.

for the pseudo-labels derived from unlabeled data to reduce the negative impact of incorrect pseudo-labels on the model learning process. Overall, pseudo-label-based algorithms can achieve higher-precision fine-grained point cloud segmentation through sample expansion and have a strong adaptability to various types of data. However, their drawbacks include the threshold sensitivity, the limited quantity, and the error accumulation of the introduced pseudo-labels. Contrastive learning brings samples of the same category closer in the feature space and pushes samples of different categories apart, prompting the network to learn more robust and clear high-dimensional category boundaries. Liu et al. [39], Yang et al. [40], and others have successively utilized this approach to enhance the efficiency of point cloud segmentation. Overall, contrastive learning forces the network to generate more distinct category boundaries and can extract more supervision information, to a certain extent. However, the self-supervised pretraining methods are still in the development stage, and their segmentation accuracy remains relatively low.

III. Proposed Method

To enhance feature generalization and improve the performance of point cloud semantic segmentation, from the perspectives of expanding the network depth and enhancing the local feature details, a point cloud segmentation network based on receptive field expansion and vector feature enhancement (RV-Net) is proposed. In the following, we first introduce the overall architecture of RV-Net, which consists of three main modules: the GEC module, the VFF module, and the sample boundary expansion (SBE) module. The GEC module separates the process of spatial relationship modeling and local feature aggregation, extending the scale of the local features and promoting the deep construction of point cloud networks. The VFF module achieves the transformation of point cloud feature vectors by constructing independent angle information, and the optimization process of the network deeply separates significant difference information. The SBE module expands the label boundaries by introducing duplicate labels of marked samples, thereby enhancing the network's ability to distinguish noise and reducing the probability of overfitting.

A. Overall Architecture

RV-Net adopts an encoder–decoder architecture. Different layers within the encoder can aggregate features from different receptive fields of the point cloud. Specifically, the upper layers of the encoder retain more fine-grained local features of the point cloud, while the lower layers contain more long-range dependency features. Each encoding layer consists of a

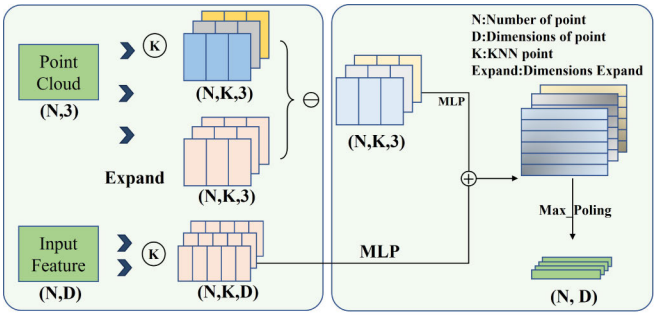

Fig. 2. Spatial information feature fusion.

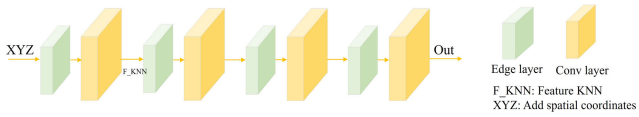

Fig. 3. Dynamic graph convolution algorithms.

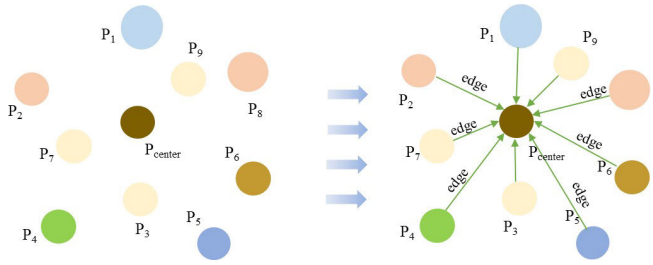

Fig. 4. Static graph construction.

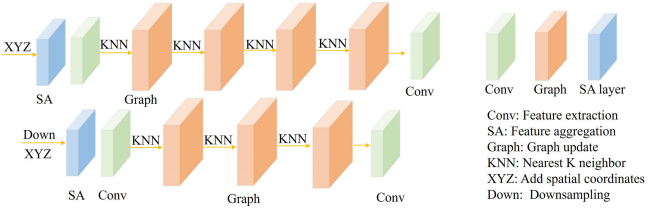

Fig. 5. Decoupled graph operations.

downsampling operation and multiple basic blocks, and each decoding layer is composed of an upsampling operation and an MLP. The basic block is composed of three types of blocks: GEC module, VFF module, hybrid block, and SBE module. As illustrated in Fig. 1, the GEC module is responsible for fully representing multiscale context and local details, the VFF module deeply separates significant differential information, the hybrid block acts as a high- and low-frequency signal mixer, adaptively bridging the GEC and VFF modules, and the SBE module introduces more samples based on the deep features of the network. Features are transmitted through skip connections between the encoder and decoder. Finally, a fully connected layer with softmax activation is used to predict the classification score of each point, and the segmentation result is determined by the category with the highest score. FPS is used to downsample the point cloud. Given a point cloud set P consisting of N points, four different features are extracted through the four layers of the encoder: (N/4, 32), (N/64, 128), (N/256, 256), and (N/512, 512).

B. GEC Module

A series of networks represented by RandLA-Net has been used to build complex local feature extraction modules. However, the complex patterns specified by humans are difficult to fully reproduce in local regions, showing limited generalization ability. Meanwhile, the high computational cost restricts the expansion of the model's receptive field. In the proposed approach, a GEC module is constructed. By fusing space and information in advance and combining a simple edge function, efficient local information fusion is achieved. Through the superposition of multilevel modules and the combination of a larger KNN range, the point cloud information flows in the graph, efficiently obtaining local–global information.

1) Spatial Information Feature Fusion: During the multilevel aggregation process, a large amount of spatial coordinate encoding is also required, which brings high computational complexity. In this stage, to reduce the computational load and promote the in-depth integration of spatial information and feature information, the spatial information feature fusion operation is conducted. Once in each layer of the multilevel network, the spatial information and feature information are fused to achieve an integrated expression of spatial and feature information, avoiding the complex spatial coordinate encoding process in the subsequent steps. This process is shown in Fig. 2.

The joint spatial-attribute information is shown as follows:

$$\hat{P}_i = MLP\left(P_i - P_i^k\right) \tag{1}$$

where $\hat{P}_i$ is the result of spatial coordinate encoding, MLP is a multilayer perceptron, P_i represents the coordinates of point cloud i, and P_i^k represents the coordinates of the k nearest point clouds to point cloud i

$$\bar{F}_i = \max\left(MLP\left(\hat{P}_i \oplus \hat{F}_i\right)\right) \tag{2}$$

where $\bar{F}_i$ represents the fusion result of point cloud i, $\hat{P}_i$ is the encoded result of spatial coordinate information, and $\hat{F}_i$ is the encoded result of features.

2) Decoupled Edge Graph Operations: Dynamic graph convolution algorithms (as shown in Fig. 3) represented by the dynamic graph CNN (DGCNN) [41] construct edge convolution functions, achieving the updating of nodes and edges in the graph through layer-by-layer edge convolution and neighborhood construction in the feature space. However, as the network iterates and deepens, it loses corresponding spatial information and the feature space becomes unstable. The subsequent RandLA-Net [12] model builds a static graph, as illustrated in Fig. 4, avoiding the search for neighboring points in the feature space and introducing a relative position encoding layer by layer, further increasing the complexity of the network. Considering the above problems, and to reduce the complexity of the algorithm and clarify the spatial information of the point cloud data, RV-Net adopts a decoupled graph operation method, as illustrated in Fig. 5. RV-Net cancels the coordinate encoding in the graph operation process and places the intermediate convolution before the graph operation.

The aforementioned spatial information feature fusion has the effect of deeply integrating the features of points and spatial information. To reduce the model complexity, RV-Net uses a simplified edge function in the graph operation, enabling the network to increase the receptive field by expanding the KNN number and the number of graph operation layers, and enhancing the network's understanding ability for global information.

Graph operations consist of edge functions and feature transformations, with the edge function defined as follows:

$$\mathbf{y}_i = \mathrm{h}\left(f\left(\mathbf{x}_j - \mathbf{x}_i\right), \mathbf{x}_i\right) \tag{3}$$

where $\mathbf{x}_i$ represents the coordinates or features of a certain point cloud in the point cloud set N, and $\mathbf{x}_j$ represents the coordinates or features of the k-nearest neighbor point cloud of $\mathbf{x}_i$. f is the feature transformation function, which can be Conv or a shared parameter MLP. $\mathbf{y}_i$ represents the output feature obtained for the ith point in the point cloud after the graph operation. In the proposed approach, the feature transformation function is placed in advance, and h is the graph operation.

To further control the complexity of the algorithm, graph operations are still implemented with max pooling and combined with skip connections to retain the original feature information. The process is shown in the following equation:

$$\mathbf{y}_i = \max\left(\mathbf{x}_j - \mathbf{x}_i\right) + \mathbf{x}_i. \tag{4}$$

C. VFF Module

In recent years, significant progress has been made in 3-D semantic segmentation methods based on point clouds. The representative work—PointNeXt [9]—by deepening the depth of the point cloud local feature module, has demonstrated a performance that is comparable to that of the mainstream transformer architectures [6], [8], [30]. RV-Net, based on an MLP architecture with parameter sharing, combines a large receptive field feature aggregation strategy to achieve hierarchical extraction of global context features. To further enhance the discriminative representation ability of the features, inspired by capsule networks based on vector feature transformation, RV-Net introduces a detailed feature mining (VFF) module based on vector space transformation.

This module builds a feature representation space with geometric perception through multiaxis vector rotation transformation of the local features. Capsule networks learn hierarchical relationships between entities through a complex dynamic routing mechanism, which has heavy computational overhead and is difficult to scale for large-scale point cloud processing. In contrast, the VFF module utilizes a lightweight yet effective rotational projection. This reduces computational complexity and simultaneously enhances the model's ability to represent local anisotropic features.

The VFF module builds a geometric mapping model from scalar features to vector features based on the mathematical principle of 3-D space rotation transformation. Fig. 6 shows the 3-D space rotation. Specifically, by introducing independent and learnable rotation angle parameters, a nonlinear mapping relationship from the scalar feature space to the vector feature space is constructed. During the network

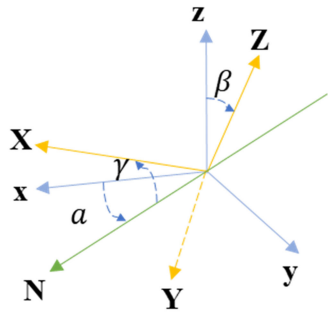

Fig. 6. 3-D space rotation.

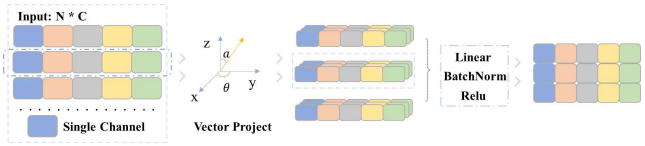

Fig. 7. Vector feature projection and aggregation.

optimization process, a channel-adaptive vector projection mechanism is adopted to project the scalar features of each channel into vector detail features with direction perception. Compared with the limitation of traditional scalar features that can only represent intensity information, vector features with direction attributes can more comprehensively describe geometric structures. The magnitude of the vector features effectively reflects the saliency of local features, while the directional information encodes the geometric distribution characteristics of local point cloud regions.

After the feature fusion in the GEC module, the model achieves an initial perception of local regions and effectively integrates spatial and attribute information. By constructing a spatial loss, it completes the decoupling of graph operations for coordinate encoding. The VFF module expands the original features into multichannel vector features. As illustrated in Fig. 7, each channel is represented as a vector along a specific coordinate axis. FP is then achieved by learning a rotation of this vector. To reduce the high optimization burden, we adopt a simplified parameterization where one axis is fixed, and the rotation is defined by only two learnable angles.

Because the elements of the rotation matrix are interdependent, directly predicting the rotation matrix using a network brings difficulties to nonlinear optimization. Therefore, the VFF module constructs the rotation matrix by predicting the rotation angles of the feature to, at most, three coordinate axes. To further reduce the parameter optimization volume of the network, if a certain vector is fixed as a certain rotation axis, two angles can be used for rotation. The rotation angles are optimized based on the network, as shown in (5). The rotation angle α_j and β_j of the feature f_j corresponding to point cloud j is obtained through optimization by the network, implemented by an MLP with shared parameters. ϕ includes batch normalization (BN) and rectified linear unit (ReLU) modules, and $\mathbf{W}$ and $\mathbf{b}$ are linear projection parameters

$$\left[\theta, \phi\right] = \phi\left(\mathbf{W} f_j + \mathbf{b}\right). \tag{5}$$

During the network training process of the VFF module, the angle information of the 3-D rotation of the point cloud is obtained. Based on the principle of the Euler angle rotation equation, the rotation matrix is constructed. For a vector $\mathbf{r}$ fixed along the y-axis [0, d, 0], the rotation matrix around the x-axis with an angle θ is $\mathbf{R}_x(\theta)$, and the rotation matrix around the z-axis with an angle ϕ is $\mathbf{R}_z(\phi)$. The final rotation result is $\mathbf{r}''$, as shown in the following equations:

$$\mathbf{R}_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \tag{6}$$

$$\mathbf{R}_z(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{7}$$

$$\mathbf{r}'' = \mathbf{R}_z(\phi)\,\mathbf{R}_x(\theta)\,\mathbf{r}' = \begin{bmatrix} -d\cos\theta\sin\phi \\ d\cos\theta\cos\phi \\ d\sin\theta \end{bmatrix} \tag{8}$$

where Vector $\mathbf{r}$ is first rotated by a θ angle around the X-axis and then by a ϕ angle around the z-axis to obtain the final vector. The three components can be simply understood as multiaxis components along the spatial coordinate system. $\mathbf{R}_z(\phi)$ and $\mathbf{R}_x(\theta)$ respectively represent the rotation matrices around the z-axis and the X-axis, and $\mathbf{r}$ represents the input feature, ϕ and θ are generated by (5).

Current multiscale feature fusion methods for point clouds typically aggregate information from different neighborhood scales through feature concatenation, followed by feature embedding or dimensionality reduction to enhance representation capability [42], [43]. To further enhance global perception, MCTNet [44] introduces a Transformer-based architecture with cross-attention, which effectively enlarges the receptive field and enables the modeling of long-range dependencies in large-scale point clouds. To extract local–global features in a multiscale range, the multihead mechanism [11], [17] is usually employed, where multiple KNN-scale (such as 8, 16, 32, etc.) neighboring feature extraction modules are stacked at the same level of the network. However, this type of method leads to an exponential increase in the computational load of the network. There is overlap among the multiple KNN scales, and the feature fusion shows a repetitive phenomenon. Moreover, due to the consistency of the feature extraction modules, the improvement in network performance is limited.

The algorithm builds the feature encoding module of the graph edge coupling and the vector feature extraction module, mining the information of the point cloud from different perspectives. To reduce the computational load during the vector projection process, RV-Net also adopts multiple KNN scales for the feature extraction. Differing from the aforementioned methods, the multiple KNN scales introduced by RV-Net are constructed in the two modules, respectively. In the graph module, due to its lower computational load, a larger KNN range is adopted to achieve the understanding of global information. In the vector projection module, a smaller KNN range is used to extract local fine features. This approach can effectively achieve multilevel and multiangle perception of the network and balance the computational load, to a certain extent.

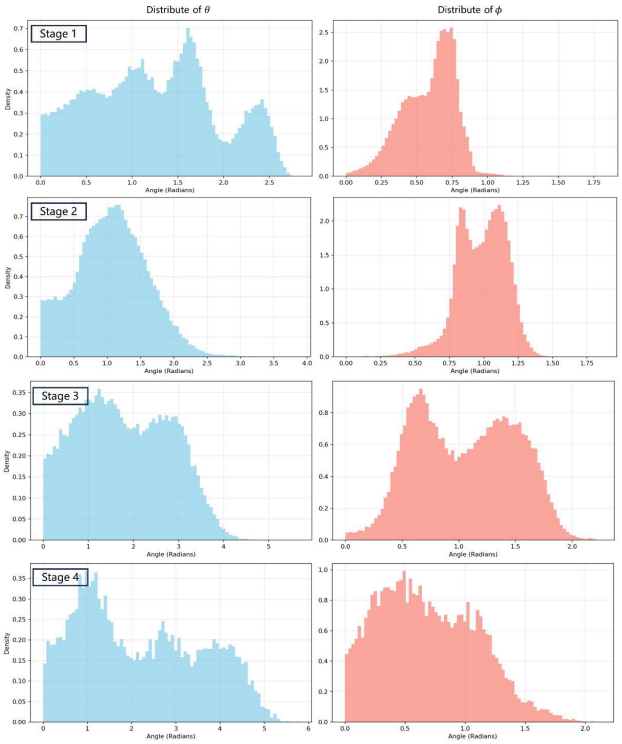

Fig. 8. Rotation angle distribution.

Existing feature fusion methods for point cloud semantic segmentation mainly aggregate information through multiscale concatenation, cross-layer propagation, and attention-based contextual interaction. Some methods concatenated geometric features extracted from different neighborhood scales and then performed embedding or dimensionality reduction to enhance representation ability [42], [43]. In encoder–decoder architectures, shallow and deep features were fused through cross-stage connections and cascade modules. More recently, Transformer-based methods emphasized contextual interaction across scales or layers via attention mechanisms [44], [45], [46]. In contrast, RV-Net incorporates an adaptive fusion strategy, introducing network parameters to optimize the fusion ratio between the two types of features. The fusion formula is shown in the following equation:

$$F_{\text{hybrid}} = \sigma(\lambda)\,F_{\text{Graph}} + [1-\sigma(\lambda)]\,F_{\text{Vep}} \tag{9}$$

where λ is a learnable parameter, σ is the activation function, F_{hybrid} represents the result of feature fusion, F_{Graph} represents the global features extracted by the graph module, and F_{Vep} represents the local features extracted by the vector module.

To verify whether the rotation angles learned by the VFF module encode meaningful geometric structures, we visualized the statistical distributions of angles θ and ϕ learned across four stages of the network. We then projected the θ angle as an attribute onto the 3-D point cloud for spatial analysis.

As shown in the histograms of Fig. 8, the learned angles do not collapse to a random or uniform distribution. Instead, they consistently converge to distinct, structured patterns across different network stages. In Stages 1 and 4, the angles exhibit clear multimodal distributions,

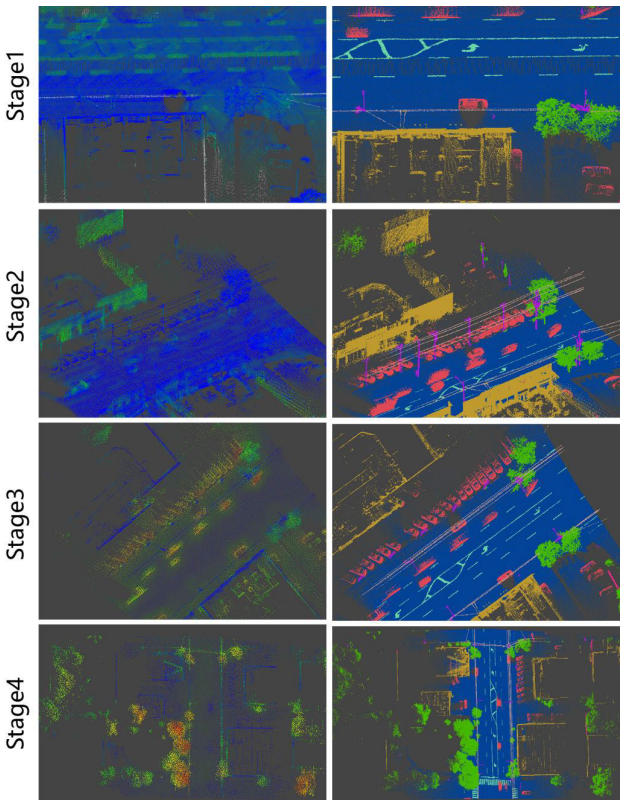

Fig. 9. Angle visualization (left) and label (right).

demonstrating a definite clustering effect. This indicates that the network has autonomously learned to decouple the 3-D space into several primary geometric orientations, which likely correspond to horizontal and vertical structures in the scene. In Stages 2 and 3, the distributions shift to a clear bimodal pattern. This may indicate that the network is also capable of capturing specific spatial angular information.

Fig. 9 illustrates the learned angles directly on the point cloud. The results reveal that the learned angles are not random but are instead highly structured and semantically correlated. Specifically, objects such as crosswalks (Fig. 9-Stage 1), building facades (Fig. 9-Stage 2), cars (Fig. 9-Stage 3), and trees (Fig. 9-Stage 4) exhibit similar angle values, respectively. This shows distinct activation patterns in the angular space, highlighting a strong correlation between the learned angles and semantic categories.

These visualization results confirm that the module has successfully avoided trivial solutions. The learned rotation angles can serve as a powerful discriminative feature, creating a direct link between the internal geometric representation and the high-level semantic identity of objects in the scene.

D. SBE Module

RV-Net effectively excavates and integrates local-global information by expanding the model's receptive field and fusing local features. However, the increase in model complexity makes it more sensitive to noise, prone to overfitting, and liable to learn incorrect information in small sample categories. In real scenarios, due to the interrelationship among point cloud

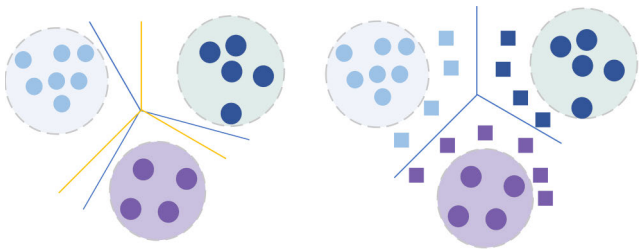

Fig. 10. SBE module.

categories, their distribution in the high-dimensional feature spaces takes on a clustered form. When the sample size is sufficient, the cluster centers of the test set and training set are theoretically merged and mutually inclusive. However, most point cloud data suffer from severe category imbalance, and the cluster centers and ranges of the training set may not fully represent and encompass those of the test set. Introducing corresponding noise can, to a certain extent, expand the feature boundaries, as shown in Fig. 10.

Han et al. [47] guided the class centers of multiclass point clouds through the class centers of single-class point clouds. The stable spatial features of single-class point clouds are obtained through the network optimization process. Adversarial learning [48] can generate artificial samples for data augmentation through adversarial training between the generator and the discriminator, promoting the backbone network to learn more discriminative features. To improve the network's performance on noisy data and reduce the probability of overfitting, the proposed approach draws on adversarial learning and the center FP strategy to generate noisy labeled samples for the network, promoting the learning of more abundant features. First, the features output by the backbone network are projected using an MLP to project the features into a more stable space and calculate the class center features. Noisy samples are generated by calculating the distance between the point cloud features and the class centers, and the cross-entropy loss between the original features and the noisy samples is constructed to promote the network to learn more abundant information. The specific process is as follows:

FP using MLP

$$F_{pr} = MLP(F). \tag{10}$$

In (10), F represents the features output by the backbone network, and F_{pr} denotes the features obtained after projection.

The class centers are calculated based on the projection results of the features. The specific operation process is shown in the following equation:

$$\mu_k = \frac{1}{\sum_{i=1}^{N} \mathbb{I}(y_i = k)} \sum_{i:y_i=k} \mathbf{f}_i \tag{11}$$

where μ_k represents the kth category center feature of the point cloud; $\mathbb{I}(y_i = k)$ is the indicator function, taking the value of 1 when the sample is k and 0 otherwise; and $\mathbf{f}_i$ is the feature of the ith point cloud in F_{pr}.

The probability of the point cloud belonging to each category is obtained by using the inner product of vectors, and

the softmax operation is performed simultaneously to set the probability between 0 and 1, as shown in the following equation:

$$s_k = \text{softmax}\left(Sim\left(F_{pr}, \mu_k\right)\right) \tag{12}$$

where s_k represents the probability of the point cloud belonging to each category, and *Sim* is the cosine similarity.

Subsequently, the expanded labels are utilized in conjunction with the cross-entropy loss to guide the training process of the backbone network.

Based on the aforementioned research process, three loss functions are introduced in RV-Net. The first is the cross-entropy loss function used for the segmentation of the backbone network. The second is the introduced noisy label cross-entropy loss L_{ent}. Since the introduced noisy labels are not in one-hot form, the Kullback–Leibler (KL) loss function is adopted to reduce the distance between the network output and the noisy labels. Meanwhile, information entropy loss is constructed to optimize the projection process of the class centers, and the two losses are combined into a cross-entropy loss. The decoupling operation of RV-Net only performs a single coordinate encoding. As a result, the deep network can encounter the problem of blurred spatial information. By randomly selecting the spatial coordinates of multiple point clouds and their output features from the backbone network, the mean square error function is used to measure the spatial information difference, and a spatial loss $L_{spatial}$ is constructed to clarify the spatial information. The $L_{spatial}$ is defined as shown in the following equations:

$$f_{r_{com}} = encode\left(x_k, x_i\right) \tag{13}$$

$$x_k = knn\left(x_i\right) \tag{14}$$

$$\left(x_k, x_i\right) = decode\left(f_{r_{com}}\right) \tag{15}$$

$$L_{spatial} = MSE\left(\left(x_k - x_i\right), \left(xk_{rel} - x_i\right)\right) \tag{16}$$

where $f_{r_{com}}$ denotes the result of spatial information feature fusion, (x_k, x_i) is the k nearest point clouds of point cloud i, and $xk_{rel} - x_i$ represents the coordinate difference between the real point cloud and its neighboring point clouds. *encode* and *decode* are the encoding and decoding structures of the network, respectively. *knn* is the operation for finding neighboring points. MSE stands for mean squared error

$$L_{\text{seg}} = -\sum y \cdot \log\left(P\left(x\right)\right) \tag{17}$$

$$L_{ent} = -\sum y \cdot \log\left(P\left(x\right)\right) \tag{18}$$

$$Loss_T = L_{seg} + L_{spatial} + L_{ent} \tag{19}$$

where L_{seg} and L_{ent} represent the cross-entropy losses of the backbone network and the noisy labels, x represents the network prediction result, and y represents the label.; $L_{spatial}$ is the spatial loss; and $Loss_T$ is the overall loss function.

E. SHW Module

RV-Net achieves expansion of the receptive fields with low computational cost through a decoupled approach, similar to the principles of the SQN [15] and PointCT [16] networks. By leveraging feature transfer between marked points and their neighboring points, RV-Net can expand the local receptive fields of multilevel neighboring point clouds, enabling efficient point cloud segmentation. On this basis, the introduced SBE module can generate labels, achieving a superior segmentation performance in scenarios with limited annotations, falling within the category of self-training weakly supervised segmentation. In contrast to test-time adaptation (TTA) methods [49], [50], which rely on online parameter updates during inference to adapt to unlabeled target data, the proposed self-training-based strategy aims to learn more transferable feature representations in advance through pseudo-supervision, thereby enabling the model to better preserve domain-invariant structural and semantic information before deployment.

To further enhance and evaluate the point cloud segmentation performance of RV-Net in weakly supervised scenarios, the proposed approach introduces a multilayer pseudo-label strategy based on RV-Net, achieving efficient segmentation of point clouds. This part is shown in Fig. 11.

The pseudo-label self-training algorithm achieves network-supervised training by jointly utilizing labeled and unlabeled data. Its core mechanism lies in generating pseudo-labels for unlabeled samples through a pretrained model and jointly using these pseudo-labels and labeled data to gradually improve the network's performance. Specifically, an initial network model is first trained based on the labeled dataset. This model is then used to predict the unlabeled point cloud data, outputting the probability distribution of each point cloud across the different semantic categories. Considering that instability in the network predictions during training can lead to the propagation of noisy labels, a confidence threshold screening strategy is adopted to control the quality of the pseudo-labels. For each unlabeled point cloud sample, only when the maximum predicted probability exceeds the preset threshold is the category label corresponding to this maximum probability taken as a valid pseudo-label to participate in the subsequent training, thereby ensuring the reliability of the pseudo-labels and the stability of the training process.

1) Pseudo-Label Generation in Two-Stage Learning: First, a portion of the labeled point cloud data Dl = $\{x_l, y_l\}$ is input into the basic network (RV-Net) based on the encoder-decoder architecture for supervised training, enabling the network to have an initial segmentation capability for point cloud data. At this stage, the point cloud's per-point-level predicted probability distribution is output, constructing an initial feature representation space, and the parameters of the backbone network are optimized through backpropagation. Second, based on the basic model, a teacher model with parameter sharing is constructed. Specifically, the teacher model infers on the weakly augmented (random rotation, mirroring, scaling, and other augmentation transformations) unlabeled point clouds to obtain the probability distribution of each category. The category corresponding to the maximum probability value is selected as the pseudo-label $\hat{y}_i^{\text{p}}$ for each point cloud. Meanwhile, to avoid interference from low-quality pseudo-labels and reduce the probability of introducing mislabeled samples, a confidence threshold τ_p (set to 0.7 after multiple experiments) is set for filtering.

2) Pseudo-Label Generation Based on Deep Feature Matching: Inspired by the work on point-wise weighted loss

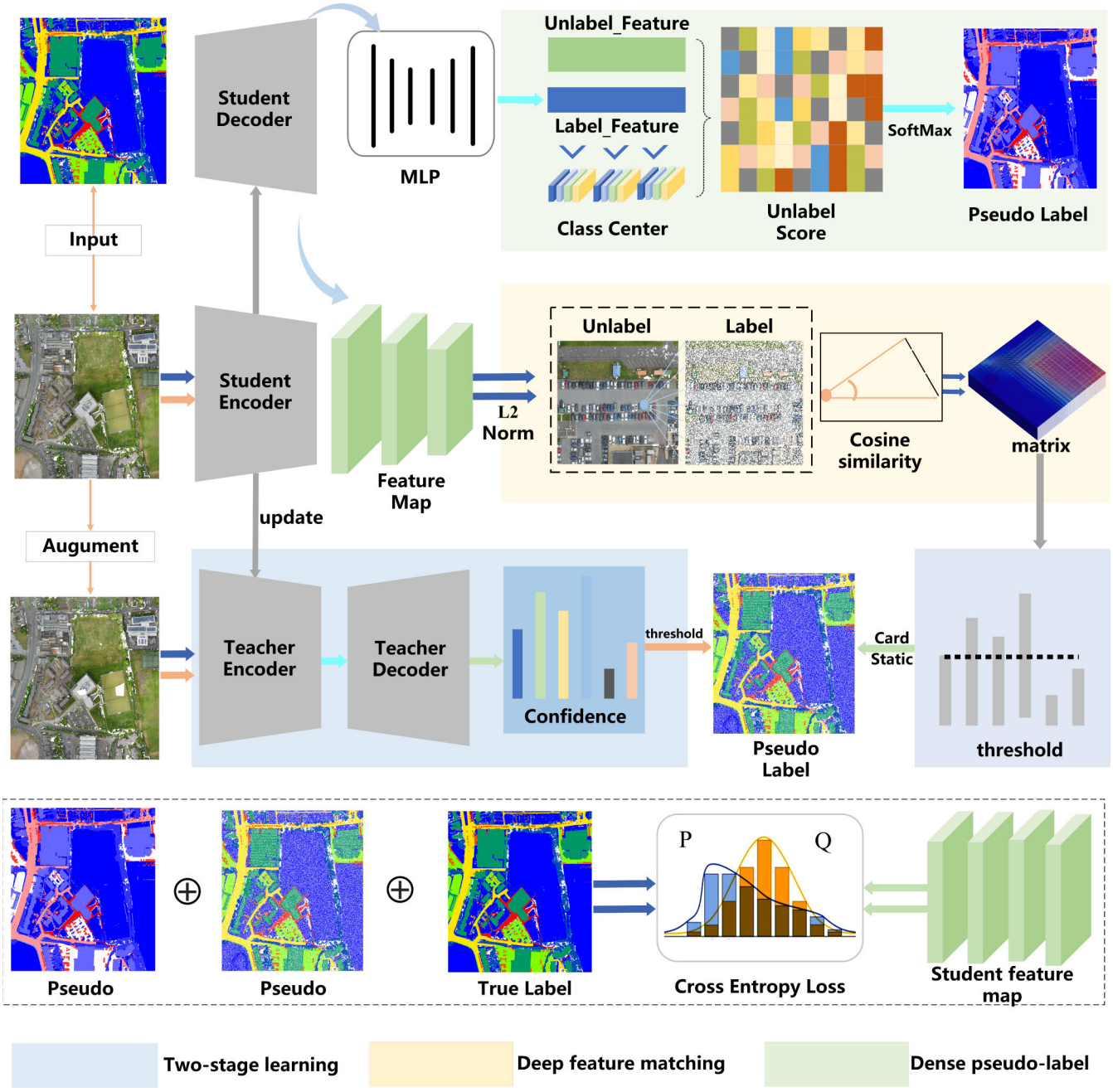

Fig. 11. Multiple pseudo-label self-training.

for imbalanced subspaces [51], we propose a pseudo-label generation mechanism based on deep feature matching, from the perspective of feature space similarity. The core theoretical basis lies in the fact that point clouds of the same category have potential feature similarity after being mapped by the network in the deep feature space. Specifically, given a point cloud dataset of size N containing L labeled samples (where $N \gg L$), the pseudo-label generation module based on deep feature matching constructs a cosine similarity matrix (MCS) between the labeled and unlabeled samples in the high-dimensional feature space, as shown in (20).

In the high-dimensional feature space, the feature differences between points are relatively small, and the cosine multiplication values are all relatively large. To enhance the discriminability of pseudo-labels, this module first uses a threshold T (set to 0.98 after multiple experiments) to perform binary processing on the MCS matrix, as shown in (21). Due to the problem of feature drift during the network training process, there will still be multicategory point clouds greater than the threshold in the binary matrix. To avoid introducing high-noise labels to the network, RV-Net calculates the cardinality statistics of the binary matrix to determine the probability of belonging to each category. Due to the significant category imbalance problem among point clouds and the random labeling of point clouds, using only the cardinality statistics of similarities higher than the threshold as the final pseudo-label category cannot avoid the influence of label category imbalance. This module also takes into account the maximum similarity, and the pseudo-label is set as the intersection of the cardinality statistics and the maximum

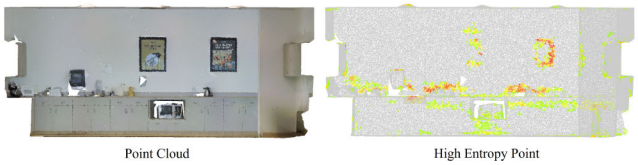

Fig. 12. High-entropy point cloud where traditional labels are not activated.

similarity, as shown in the following equation:

$$MCS\ (j,k) = \frac{\mathbf{f}_j \cdot \mathbf{f}_k}{\|\mathbf{f}_i\|_2 \cdot \|\mathbf{f}_k\|_2} \tag{20}$$

where $\mathbf{f}_j$ and $\mathbf{f}_k$ represent the depth feature vectors of the jth point cloud and the kth point cloud, respectively,

$$\hat{MCS}\ (j,k) = \begin{cases} 1 & \text{if } MCS\ (j,k) \geq T \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

$$\hat{y}_j = \max_{c \in \mathcal{C}} \left(\sum_{k \in \mathcal{D}^c_{\text{labeled}}} MCS\ (j,k) \right) \cap argmax\left[MCS\ (j,k)\right] \tag{22}$$

where $\mathcal{D}^c_{\text{labeled}}$ represents the annotated sample set of category c, and $\mathcal{C}$ is the base statistical quantity of MCS.

3) Dense Pseudo-Label Generation Module: This module is consistent with the boundary expansion strategy and mainly generates low-entropy pseudo-labels by constructing a learnable module, reducing the impact of noise in the pseudo-labels on the network training. Differing from fully supervised networks, this process is based on the momentum update operation for labeled point cloud features to obtain class centers, as shown in (23). Pseudo-labels are obtained by calculating the similarity between the unlabeled point cloud features and the class centers. Since pseudo-labels for all the unlabeled point clouds can be obtained in this process, it can be regarded as the construction of dense pseudo-labels. Fig. 12 shows the pseudo-labels generated by high-entropy points

$$\mu_c^{(t)} = \alpha \mu_c^{(t-1)} + (1-\alpha)\frac{1}{|\mathcal{B}_c|}\sum_{i \in \mathcal{B}_c} z_i \tag{23}$$

where μ_c is the prototype of category c, α is the momentum coefficient, z_i represents the characteristic of point i, and $\mathcal{B}_c$ is the sample set of category c in the current batch.

IV. Experiments and Results

A. Datasets

High-quality labeled data serve as an important benchmark for the performance evaluation and optimization of segmentation algorithms. To assess the generalization ability of the network models and meet the diverse application requirements of the mapping and remote sensing field, three benchmark datasets with significant scene differences and data characteristics were selected: the S3DIS built by Stanford University, the Toronto city vehicle-mounted point cloud dataset (Toronto-3D), and the aerial survey LiDAR point cloud dataset (DALES). The combined datasets cover three major application scenarios: indoor building analysis, vehicle-mounted road perception, and aerial surface classification.

1) Stanford Large-Scale 3-D Indoor Spaces Dataset: S3DIS is a high-quality 3-D point cloud of real indoor scenes generated by Stanford University through multiview scanning of buildings with cameras and multimodal data fusion. This dataset is a dense point cloud with RGB color information, spatial coordinates (XYZ), and 13 semantic labels. Each point cloud has been annotated at the pixel level, including typical indoor elements such as ceilings, floors, columns, tables, and chairs. The complexity of the scenes is significant, and the dataset contains real complex scenarios such as class imbalance and local feature confusion, which can effectively test the feature abstraction ability and context reasoning performance of the algorithms. In terms of dataset division, we followed the mainstream approach and selected Area 5 for testing and Areas 1–4 for training.

2) Toronto-3D: The Toronto-3D dataset was obtained through dynamic scanning by an on-board LiDAR system on urban roads. It covers approximately 1 km of continuous road, with a scanning range extending 100 m perpendicularly from the road centerline. It contains 78.3×10^6 point clouds and labels eight types of urban elements. Due to the movement characteristics of the on-board platform, this dataset shows significant density differences along the road's longitudinal direction and has relatively high point cloud distortion. At the same time, the imbalance among the different ground object categories is severe. According to the official division standard, L001, L003, and L004 were used as the training set, and L002 was used as the test set for the research.

3) Dayton Annotated LiDAR Dataset: DALES is currently the largest publicly available airborne point cloud dataset in the world. This dataset contains 500 million precisely labeled point clouds, covering 40 independent spatial scenes and defining eight typical ground objects. It focuses on geometric feature modeling, retaining only XYZ coordinate attributes while discarding RGB information, providing an ideal environment for the verification of point cloud algorithms. As the first airborne benchmark dataset to exceed the square kilometer level, DALES effectively fills the data gap in urban-level aerial 3-D semantic understanding.

B. Experimental Setup

This section introduces the algorithms used in the comparison and the indicators used for comparing the network accuracy in each dataset.

1) The hardware configuration in the experiments was an AMD 5975WX CPU and an NVIDIA RTX3090Ti GPU. The software configuration included the Ubuntu 20.04 operating system, CUDA 11.7, Python 3.8, and the PyTorch 1.13 deep learning framework. The RV-Net model was trained for 100 epochs using the Adam optimizer. The initial learning rate, the number of input point clouds, and the batch size were set to 0.006, 40 000, and 8, respectively. The network had four encoding layers with depths set to [2, 2, 4, 2], and the corresponding feature dimensions were [64, 128, 256, 512].
2) The overall accuracy (OA), Intersection over Union (IoU) for individual classes, and mean IoU (mIoU)

TABLE I

QUANTITATIVE RESULTS OF THE DIFFERENT METHODS ON THE S3DIS DATASET. BOLD INDICATES THE BEST RESULT, AND UNDERLINED INDICATES THE SECOND-BEST RESULT

	PointCNN	KPConv	PTV1	Stratified	PointNeXt	PTV2	EP-Net-L [52]	RV-Net
Ceiling	57.3	92.8	94	**96.2**	94.2	94.5	94.0	95.3
Floor	98.2	97.3	98.5	**98.7**	98.5	98.6	98.0	98.6
Wall	79.4	82.4	86.3	85.6	84.4	87.2	86.1	**87.8**
Column	17.6	23.9	38	46.1	37.7	34.4	48.5	**51.5**
Window	22.8	58	63.4	60	59.3	64.7	60.2	**66.3**
Door	62.1	69	74.3	76.8	74	**77.9**	74.3	71.7
Table	74.4	81.5	82.4	92.6	83.1	**93.1**	82.6	84.1
Chair	80.6	91	89.1	84.5	91.6	84.4	90.5	**92.7**
Sofa	31.7	75.4	80.2	77.8	77.4	77.3	78.3	**85.1**
Bookcase	66.7	75.3	74.3	75.2	77.2	**86.3**	78.5	79.1
Board	62.1	66.7	76	78.1	78.8	**84.5**	78.8	81
Clutter	56.7	58.9	59.3	**64**	60.6	62.2	60.6	63.8
OA	85.9	-	90.8	91.5	90.7	91.6	91.1	**92.1**
mIoU	63.9	67.1	70.4	72	70.8	72.7	71.9	**73.6**

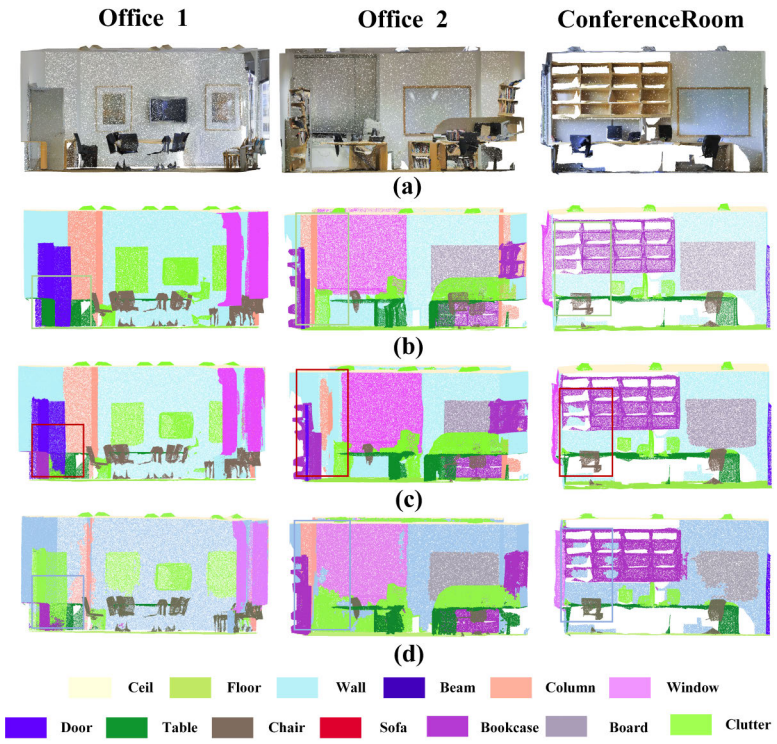

Fig. 13. S3DIS conference room segmentation results. (a) RGB. (b) GT. (c) RV-Net. (d) RandLA-Net.

for all classes are adopted here as evaluation metrics. The OA refers to the proportion of correctly classified samples by the model in the entire test set, to the total number of samples. The mIoU is achieved by calculating the ratio of the intersection to the union between the predicted results and the true labels. The calculation formulas for the OA and mIoU are shown in the following equations:

$$OA = \frac{\sum_{c=1}^{C} TP_c}{N} \tag{24}$$

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{25}$$

$$mIoU = \frac{1}{C}\sum_{c=1}^{C} IoU_c \tag{26}$$

TABLE II
QUANTITATIVE RESULTS OF THE DIFFERENT METHODS ON THE TORONTO-3D DATASET. BOLD INDICATES THE BEST RESULT, AND UNDERLINED INDICATES THE SECOND-BEST RESULT

	OA	mIoU	Road	Road Mark	Natural	Build	UtilityLine	Pole	Car	Fence
RandLA-Net	94.4	81.8	96.7	64.2	96.9	**94.2**	88.0	77.8	93.4	42.9
ResDLPS-Net [53]	96.5	80.3	95.8	59.8	96.1	90.9	86.8	79.9	89.4	43.3
BAF-LAC [54]	95.2	82.2	96.6	64.7	96.4	92.8	86.1	83.9	93.7	43.5
BAAF-Net [55]	94.2	81.2	96.8	67.2	96.8	92.2	86.8	82.3	93.1	34.0
LEARD-Net [56]	97.3	82.2	97.1	65.2	96.9	92.8	87.4	78.6	94.5	45.2
PointNeXt	97.6	83.4	97.2	68.6	**97.5**	93.8	**88.3**	**84.5**	93.5	44.0
NeiEA-NET [57]	97.0	80.9	97.1	66.9	97.3	93.0	87.3	83.4	93.4	43.1
SFL-Net [58]	97.6	81.9	97.7	70.7	95.8	91.7	87.4	78.8	92.3	40.8
LACV-Net [59]	97.4	82.7	97.1	66.9	97.3	93.0	87.3	83.4	93.4	43.1
RV-Net	**98.2**	**85.2**	**98.1**	**78.1**	97	93.1	**88.3**	82.7	**95.7**	**48.6**

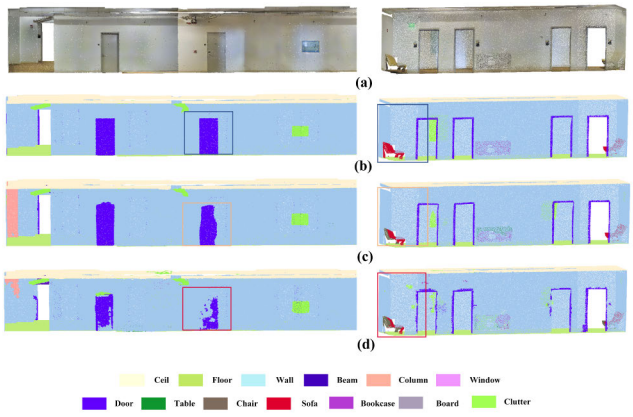

Fig. 14. S3DIS corridor segmentation results. (a) RGB. (b) GT. (c) RV-Net. (d) RandLA-Net.

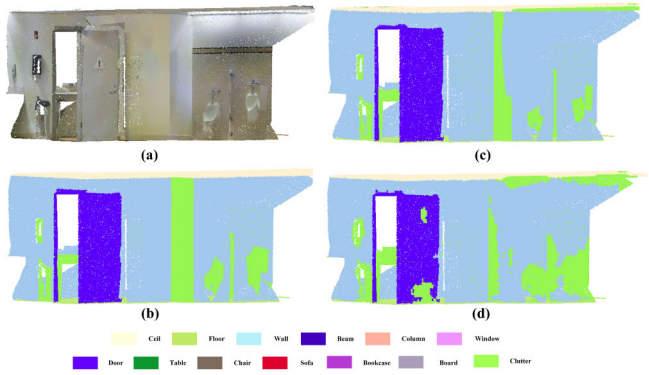

Fig. 15. S3DIS washroom segmentation results. (a) RGB. (b) GT. (c) RV-Net. (d) RandLA-Net.

where TP_c represents the true positive cases of category c (the number of points predicted as category c and actually belonging to category c), N represents the total number of points in the point cloud, FP_c represents the false positive cases of category c, and FN_c represents the false negative cases of category c.

C. Analysis of the Experimental Results

1) Results on the S3DIS Dataset: To verify the effectiveness of the proposed method, we selected the recently published point cloud segmentation algorithms for quantitative comparative analysis. The results are listed in Table I. As can be seen from Table I, the proposed RV-Net achieves an OA of 92.1% and a mIoU of 73.6%, obtaining the best OA and mIoU, compared with the recent point cloud segmentation algorithms, and achieving the best segmentation accuracy in multiple small categories. Compared with the transformer network v1 and v2 versions that adopt the self-attention mechanism, RV-Net achieves a superior segmentation performance with a lower computational complexity. By checking the corresponding categories, it can be found that the chair, sofa, floor, wall, column, and window categories all show improvements, especially column, window, and sofa, which are the most significant. This is because the decoupling of graph operations expands the receptive field and increases the network depth, achieving more complete global information capture of objects. For categories such as column and sofa that have significant differences with the surrounding objects, the recognition accuracy is improved significantly.

The different information captured in the form of vectors helps to deeply mine discriminative features, thereby improving the accuracy of categories such as floor, wall, and chair.

Figs. 13–15 show the qualitative segmentation results of the proposed method on the S3DIS dataset. Meanwhile, the mainstream RandLA-Net network is selected as the qualitative segmentation comparison algorithm. The segmentation result figures indicate that RV-Net can correctly classify the majority of the point clouds. However, some point clouds in the scene are misclassified. For instance, in Fig. 13, in Office 1, the table is misclassified as a bookcase, in Office 2, the wall is misclassified as a window, and in Conference Room, the

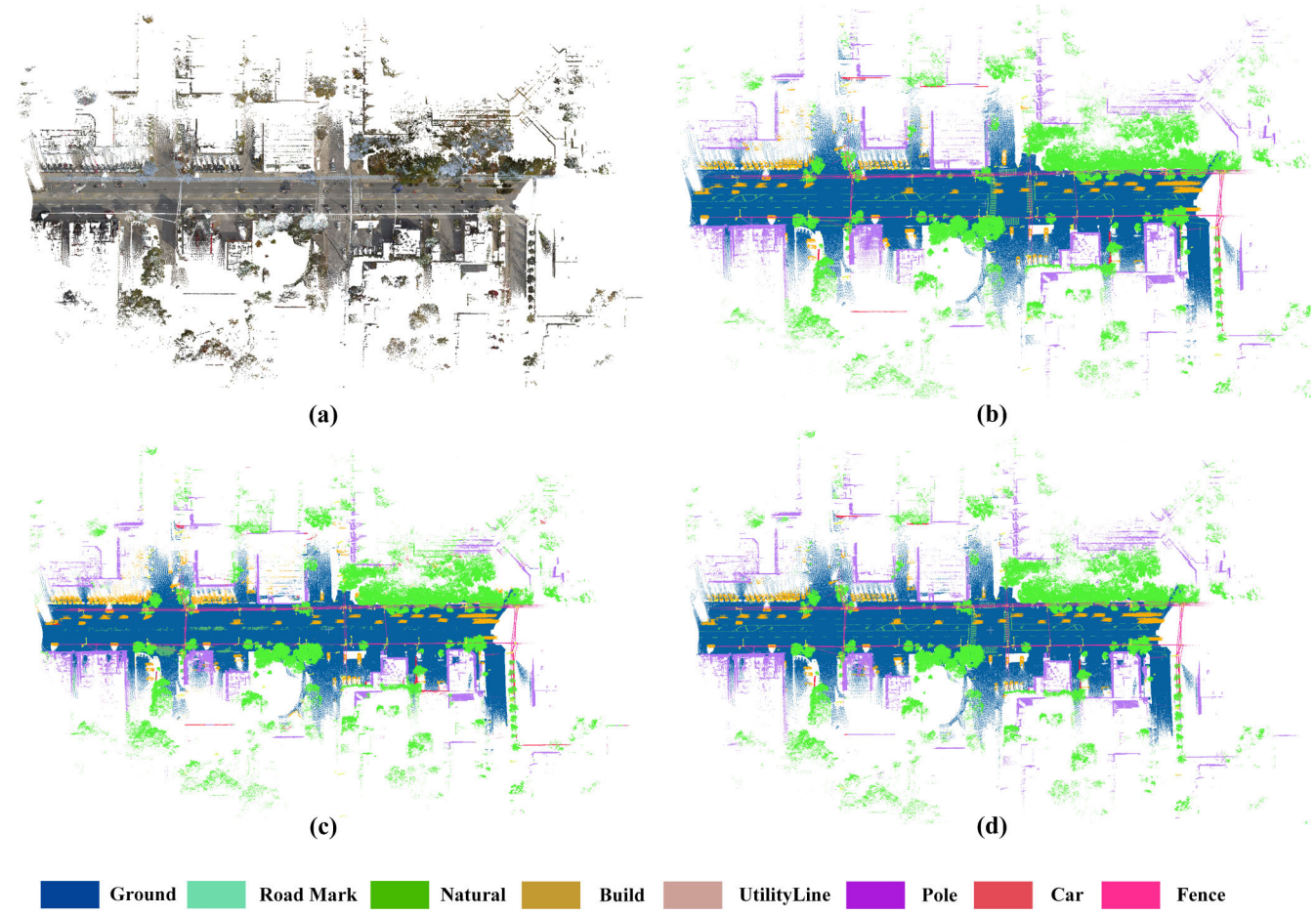

Fig. 16. Toronto-3D overall segmentation diagrams. (a) RGB. (b) GT. (c) RandLA-Net. (d) RV-Net.

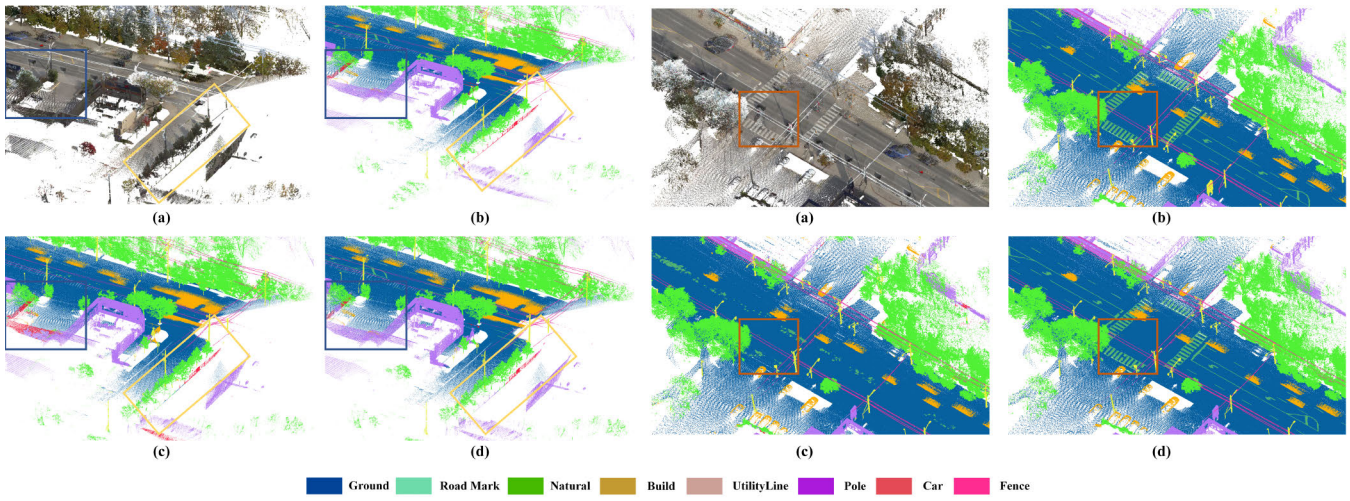

Fig. 17. Toronto-3D local segmentation results. (a) RGB. (b) GT. (c) RandLA-Net. (d) RV-Net.

bookcase is misclassified as a wall. From the misclassification results, it can be seen that this is due to the similar shapes and colors of some heterogeneous objects in certain scenes of the S3DIS dataset, which leads to segmentation errors in some areas of the network. From the segmentation results of the RandLA-Net network, it can be found that there are more severe segmentation errors at the category boundaries, such as wall and window, bookcase and wall. Fig. 14 shows the corridor area of the S3DIS dataset, and Fig. 15 shows the washroom. Compared with the real labels, both algorithms show segmentation errors for the door and sofa categories. However, RV-Net shows a better segmentation integrity and smoother edge areas.

2) Results on the Toronto-3D Dataset: Table II presents the quantitative comparison results of the different point cloud segmentation algorithms. It can be seen from the table that the

TABLE III
PERFORMANCE COMPARISON RESULTS OF THE DIFFERENT METHODS ON THE S3DIS DATASET. BOLD INDICATES THE BEST RESULT, AND UNDERLINED INDICATES THE SECOND-BEST RESULT

Method	OA(%)	mIoU(%)	Param. M	FLOPs G	Pts
PointCNN	85.9	57.3	0.6	-	-
PointNet	-	41.1	3.6	35.5	-
KPConv	72.8	67.1	15.0	-	-
RandLA-Net	88.0	70.0	1.3	5.8	-
PointNeXt-L	<u>90.7</u>	<u>70.8</u>	7.1	15.2	24k
RV-Net	**92.1**	**73.6**	6.31	10.5	30k

proposed method achieves an OA of 98.2% and an mIoU of 85.2%, with the highest overall segmentation accuracy among the compared algorithms. It also achieves the highest segmentation accuracy for some single categories, such as Road and Road Mark. RV-Net significantly improves the segmentation accuracy for the Road and Road Mark categories, as the spatial difference between Road Mark and Road is small, and a large receptive field is needed to capture the differences between the road surface and road markings. By decoupling the graph operations, the receptive field of the network is expanded, allowing the network to learn subtle spatial difference features. After the point cloud was initially downsampled through gridization, the Fence category had a relatively low proportion in the dataset and was difficult to distinguish from the Natural category. RV-Net projects the extracted multidimensional features onto multiple coordinate axes through vector projection, enabling the network to discover more discriminative features and improve the segmentation accuracy for this category. During the data collection of the Toronto-3D dataset, due to the movement of vehicles on the road, the shape of the collected Car category point cloud was significantly distorted and contained a large amount of noise. RV-Net introduces a projection-based class center noise removal module, which effectively improves the segmentation accuracy of the Car category by generating partial pseudo-labels and enhancing the network's noise resistance through the class center.

Figs. 16 and 17 show the segmentation results for the Toronto-3D dataset using RV-Net and RandLA-Net. Visually, both networks achieve good segmentation results. The segmentation results for the Natural, Build, Car, and Utility-Line categories are consistent. These categories are relatively complex in terms of spatial hierarchy and structure, and the networks can easily distinguish objects from the geometric information of the point clouds. Analyzing the detailed images, RandLA-Net achieves good results for most of the Car category point clouds, but there are segmentation errors in the edge areas where it is in contact with objects, such as the Build category. RV-Net does not have such errors. RandLA-Net has weak recognition ability for Road Mark, with almost all the zebra crossings and lane lines being misidentified. This is because the spatial difference between the ground mark point clouds and ground point clouds is not significant, and the two types of point clouds are easily confused.

TABLE IV
ABLATION RESULTS FOR RV-NET. BOLD INDICATES THE BEST RESULT

Model			S3DIS dataset		Toronto-3D dataset	
GEC	VFF	SBE	OA	mIoU	OA	mIoU
√	×	×	91.6	72.1	97.9	82.9
√	√	×	92.0	73.1	98.1	84.4
√	×	√	91.7	72.7	97.9	83.8
√	√	√	**92.1**	**73.6**	**98.2**	**85.2**

TABLE V
ABLATION RESULTS FOR SHW MODULE. BOLD INDICATES THE BEST RESULT

Module	TS	DF	FP	OA	mIoU
$Base$	×	×	×	97.6	82.1
$Base_T$	√	×	×	97.8	83.5
$Base_{TD}$	√	√	×	98.0	84.2
$Base_{TDP}$	√	√	√	**98.0**	**84.6**

The spatial geometric features extracted by the network cannot easily identify the discriminative difference information. In this case, the network can enhance its ability to mine RGB attribute information. RV-Net has a strong recognition ability for Road Mark, and the visual results are almost consistent with the ground-truth values, to some extent, demonstrating its strong adaptive ability. In the segmentation results for Pole, RandLA-Net misidentifies large tree trunks and small trees in the Natural category as Pole, while the RV-Net algorithm does not have this problem, indicating that this algorithm has a strong ability to mine abstract features.

3) Performance Comparison: Experimental results are presented in Table III. On the S3DIS dataset, the proposed method consistently outperforms the PointNeXt baseline in both accuracy and efficiency. Specifically, our model achieves

TABLE VI

QUANTITATIVE EXPERIMENTAL RESULTS OF THE VARIOUS METHODS ON THE TORONTO-3D DATASET. IN THE SETTINGS COLUMN, FULL DENOTES THE FULLY SUPERVISED METHODS, WHILE 1%, 10% SIGNIFY THE PROPORTIONS OF LABELED POINTS. THE BOLD NUMBERS HIGHLIGHT THE BEST PERFORMANCE UNDER THE SPECIFIC NUMBER OF LABELED POINTS

Set	Method	OA	mIoU	Road	Road Mark	Natural	Build	UtilityLine	Pole	Car	Fence
Full	RandLA	94.4	81.8	96.7	64.2	96.9	94.2	88	77.8	93.4	42.9
	ResDLPS-Net	96.5	80.3	95.8	59.8	96.1	90.9	86.8	79.9	89.4	43.3
	BAF-LAC	95.2	82.2	96.6	64.7	96.4	92.8	86.1	83.9	93.7	43.5
	BAAF-Net	94.2	81.2	96.8	67.2	96.8	92.2	86.8	82.3	93.1	34.0
	LEARD-Net	97.3	82.2	97.1	65.2	96.9	92.8	87.4	78.6	94.5	45.2
	PointNeXt	97.6	83.4	97.2	68.6	97.5	93.8	88.3	84.5	93.5	44.0
	NeiEA-NET	97.0	80.9	97.1	66.9	97.3	93.0	87.3	83.4	93.4	43.1
	SFL-Net	97.6	81.9	97.7	70.7	95.8	91.7	87.4	78.8	92.3	40.8
	LACV-Net	97.4	82.7	97.1	66.9	97.3	93.0	87.3	83.4	93.4	43.1
10%	RV-Net	98.1	84.8	98.2	77.2	96.9	93.0	87.7	82.4	93.1	50.0
1%	RandLA	95.9	76.6	95.2	56.7	96.0	91.2	85.9	79.6	88.4	19.8
	PSD	96.2	76.2	96.3	62.2	93.4	87.9	**85.9**	77.5	89.0	17.0
	DR-Net	96.6	79.1	96.3	60.1	95.4	91.8	85.0	78.1	88.4	37.3
	RV-Net	**98.0**	**84.6**	**98.0**	**76.1**	**97.2**	**93.6**	85.7	**82.2**	**94.9**	**49.2**

92.1% OA and 73.6% mIoU, surpassing PointNeXt by 1.4% in OA and 2.8% in mIoU, respectively.

Notably, these accuracy gains are achieved with remarkable efficiency. While processing a larger input scale (30k points versus 24k points), the proposed model maintains a more compact architecture with only 6.3M parameters, resulting in a 11.3% reduction compared to PointNext-L. The floating-point operations amount to 10.5 GFLOPs, representing a 30.1% reduction compared to PointNext-L.

These results demonstrate that the proposed method achieves a superior trade-off between segmentation accuracy, model compactness, and scalability to large-scale point clouds, making it particularly suitable for high-resolution indoor scene understanding tasks.

D. Ablation Experiments

1) Base Module: To evaluate the effectiveness of the constructed learning module and the introduced strategies, ablation experiments were conducted on the S3DIS and Toronto-3D datasets. The ablation results are listed in Table IV. In the S3DIS dataset, compared with PTV1 (with an OA of 90.8% and an mIoU of 70.4%), the GEC module, as the basic module of the network, shows a significant improvement, achieving an OA of 91.6% and an mIoU of 72.1%. The GEC module can achieve a performance surpassing that of the transformer network, with a relatively low computational cost. On this basis, the VFF module was introduced, with the OA and mIoU being increased by 0.4% and 1.0%, respectively. At the same time, the SBE module was introduced, resulting in the OA and the mIoU both being improved, with the OA increasing by 0.5% and the mIoU increasing by 1.5%, verifying the effectiveness of the introduced modules. Ablation experiments were also conducted on the Toronto-3D dataset. Compared with the SFL-Net network (with an OA of 97.6% and an mIoU of 81.9%), the GEC module increases the OA by 0.3% and the mIoU by 1%. On this basis, the VFF and SBE modules were introduced, respectively, resulting in an improvement of both the OA and mIoU, proving the effectiveness of the proposed modules.

2) SHW Module: We conducted ablation experiments on the soft-hard pseudo-label weakly (SHW) module to evaluate the impact of a single or multiple pseudo-labeling mechanisms on the training of the network model. The experiments were conducted on the Toronto3D dataset using 1% labeled data, evaluating OA and mIoU. As shown in Table V, we incrementally added three main components: two-stage (TS) learning, deep feature (DF) matching, and dense label generation via FP.

Base is a method that performs weakly supervised segmentation solely using the original network. By leveraging a large receptive field and multilayer feature fusion, the backbone network effectively propagates sparse labels and outperforms prior weakly supervised methods. Compared to networks such as DR-Net, the semantic segmentation performance of *Base* is stronger. The two-stage learning mechanism yields a significant performance boost, shown as $Base_T$. This confirms that generating high-confidence pseudo-labels based on the model's own predictions is a highly effective strategy for improving segmentation accuracy.

$Base_{TD}$ simultaneously introduces the pseudo-label generation based on feature matching and the two-stage learning. These two types of pseudo-labels are introduced from two different perspectives: network output and feature space similarity. TS leverages the output logits, while DF utilizes similarity in the learned feature space. Combining these two distinct perspectives provides the network with a richer and more robust set of high-quality pseudo-labels, leading to superior performance.

Finally, $Base_{TDP}$ integrates all modules. The dense pseudo-labels generated by FP provide comprehensive supervision across the entire point cloud. While this slightly stabilizes OA, it delivers a crucial improvement in mIoU. This indicates that the label quantity in the weakly supervised learning process of the network is limited. The process may introduce noise to the network and cause the segmentation performance of some categories to decline. Overall, the above results show that hard labels can improve the segmentation performance of large categories, while soft labels can improve the segmentation performance of small categories.

TABLE VII

QUANTITATIVE EXPERIMENTAL RESULTS OF THE VARIOUS METHODS ON THE DALES DATASET. IN THE SETTINGS COLUMN, FULL DENOTES THE FULLY SUPERVISED METHODS, WHILE1%, 10% SIGNIFY THE PROPORTIONS OF LABELED POINTS. THE BOLD NUMBERS HIGHLIGHT THE BEST PERFORMANCE UNDER THE SPECIFIC NUMBER OF LABELED POINTS QUANTITATIVE EXPERIMENTAL RESULTS OF THE VARIOUS METHODS ON THE TORONTO-3D DATASET. IN THE SETTINGS COLUMN, FULL DENOTES THE FULLY SUPERVISED METHODS, WHILE 1%, 10% SIGNIFY THE PROPORTIONS OF LABELED POINTS. THE BOLD NUMBERS HIGHLIGHT THE BEST PERFORMANCE UNDER THE SPECIFIC NUMBER OF LABELED POINTS

Set	Method	OA	mIoU	Ground	Buildings	Cars	Trucks	Poles	Power lines	Fences	Vegetation
Full	PointCNN	97.2	58.4	97.5	95.7	40.6	4.8	57.6	26.7	52.6	91.7
	SPG [60]	95.5	60.6	94.7	93.4	62.9	18.7	28.5	65.2	33.6	87.9
	KPConv	97.8	81.1	97.1	96.6	85.3	41.9	75.0	95.5	63.5	94.1
	Pyramid [61]	98.3	83.6	97.8	97.3	88.4	47.9	77.6	96.7	67.5	95.4
	PTV1	90.8	67.2	91.1	63.4	54.4	37.8	69.4	77.3	52.0	75.9
	SPT [62]	-	79.6	96.7	96.6	86.1	52.4	65.3	95.5	52.7	93.1
	PreFormer[63]	92.9	70.9	94.5	83.8	60.3	47.7	64.8	84.4	49.5	82.1
	RandLA	97.1	80.0	97.0	93.2	83.7	43.8	59.4	94.8	71.5	96.6
10%	RV-Net	98.0	81.8	97.4	97.3	85.6	41.6	95.6	78.7	64.7	94.2
1%	DR-Net	97.0	78.0	96.8	92.8	82.5	38.1	93.5	55.1	**69.3**	96.3
	RV-Net	**97.9**	**81.1**	**97.4**	**94.0**	**85.3**	**42.5**	**95.1**	**74.5**	62.1	**97.3**

E. Analysis of Self-Training

Experiments were conducted with the training process using 1% and 0.1% labeled data, respectively. The experimental datasets were the Toronto-3D and DALES datasets. The quantitative and qualitative analysis results of the experiments are as follows.

1) Toronto-3D: Table VI presents the quantitative segmentation results for the Toronto-3D dataset. As can be seen from the table, at a 1% annotation ratio, the OA of RV-Net reaches 97.9%, which is 1.3% higher than that of the DR-Net weakly supervised network (96.6%), and even surpasses the fully supervised algorithms such as PointNeXt (97.6%) and LACV-Net (97.4%). In terms of mIoU, RV-Net achieves an mIoU of 84.2%, which is 5.1% higher than that of DR-Net (79.1%), reaching the highest value among all the comparison methods and even exceeding the recent fully supervised methods such as PointNeXt (83.4%). In categories with a high proportion of point clouds, such as Road, Road Mark, and Natural, RV-Net shows an improvement of approximately 2%, which is due to the large proportion of these categories and the introduction of more high-confidence pseudo-labels during self-training, promoting the improvement of the network's segmentation ability. For categories with a smaller proportion, such as Fence and Car, the introduction of dense labels provides supervisory signals for the extremely low-annotated Fence category, facilitating the network to learn more information. For the Car category, due to the frequent occurrence of point cloud distortion and color information loss caused by movement, the introduction of dense pseudo-labels increases the number of point clouds in the category and activates the low-confidence labels of boundary point clouds, thereby improving the segmentation accuracy to a certain extent. At a 10% annotation ratio, the segmentation performance of the RV-Net network is further improved. Since the annotated point clouds can effectively guide the network training in this process, the introduced multiple pseudo-labels only serve as auxiliary information, and thus the segmentation performance approaches the fully supervised level.

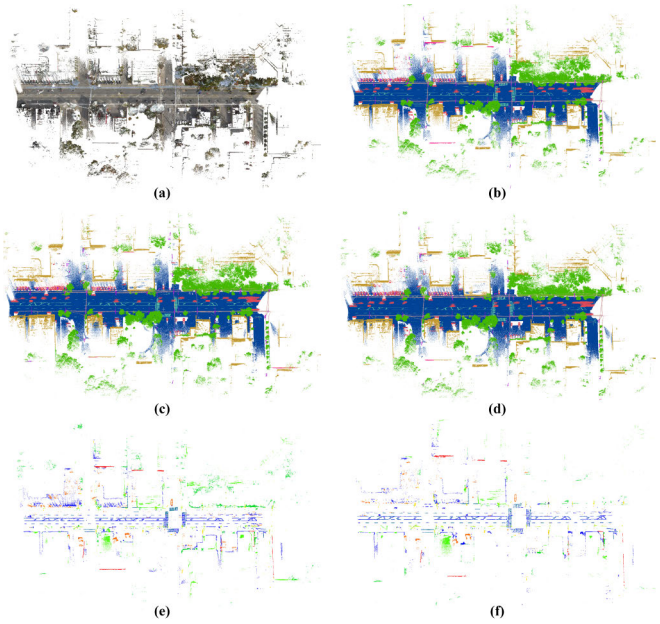

Fig. 18. Toronto-3D weakly supervised segmentation results. (a) RGB point. (b) True label. (c) DR-Net. (d) RV-Net. (e) DR-Net error. (f) RV-Net error.

Figs. 18–20 show the qualitative segmentation results of the RV-Net algorithm and the DR-Net algorithm under 1% labeled conditions. To facilitate comparison, the error maps of the labels and the true results (RV-Net-error and DR-Net-error) are also presented. Fig. 18 presents the overall segmentation results. From the perspective of the segmentation effect diagrams, both algorithms achieve good segmentation results. For categories such as Natural, Ground, and Car, due to their significant spatial hierarchy and structural differences, the errors are very small. In the Road Mark category, RV-Net performs significantly better than RandLA-Net, but there are still many errors in the boundary areas connected to the Ground

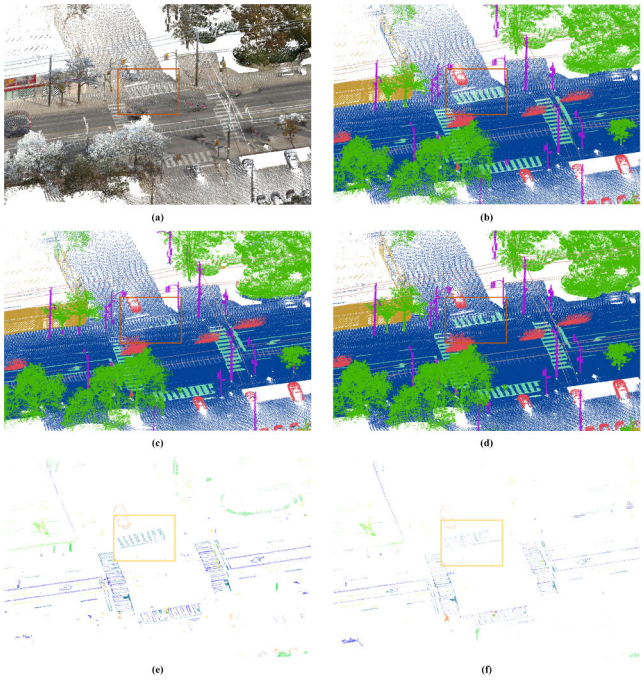

Fig. 19. Toronto-3D segmentation details (Road Mark and Natural).

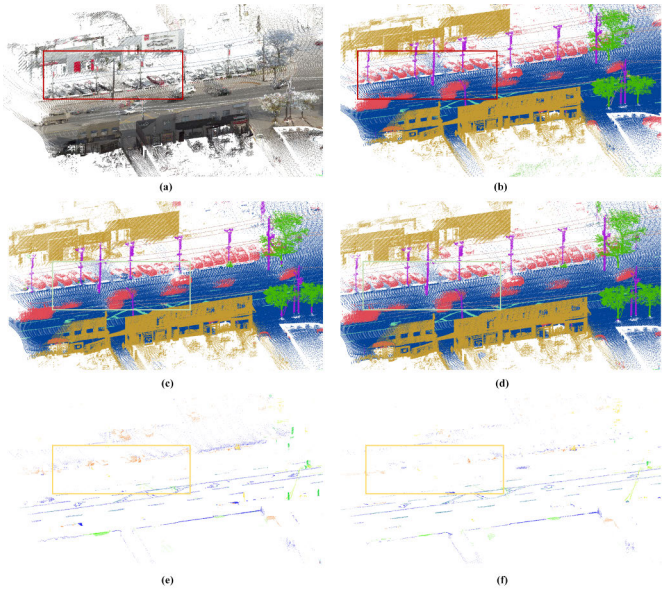

Fig. 20. Toronto-3D segmentation details (Build and Car).

category. Both algorithms show relatively large segmentation errors for the Fence category, but RV-Net obtains a slightly higher segmentation accuracy. The main reason for this is that the point cloud of the Fence category accounts for a small proportion, and the number of samples used for supervised training is even smaller after sparse labeling. Moreover, some areas are similar in shape to the point cloud of tree trunks, which affects the final segmentation accuracy. From the error maps, it can be seen that RV-Net has fewer error point clouds than DR-Net, and its overall segmentation performance is superior. Figs. 19 and 20 show the detailed segmentation results and error maps of some scenes. In Fig. 19, the road mark and the natural point clouds are better segmented by the RV-Net algorithm. In Fig. 20, the Car point clouds and the build point clouds are better segmented by the RV-Net algorithm than by the DR-Net algorithm. The qualitative visual results indicate that the segmentation performance of the RV-Net algorithm is superior to that of the DR-Net algorithm.

2) Dayton Annotated LiDAR Dataset: Table VII presents the results of the RV-Net algorithm and the recently published algorithms on the large-scale DALES dataset. RV-Net achieves an mIOU of 81.1% and an OA of 79.9% with only 1% labeled data, outperforming the fully supervised networks of Point-Transformer and RandLA-Net, and matching the performance of the fully supervised KPConv. It also outperforms the latest weakly supervised DR-Net algorithm by 3% in mIOU and 1% in OA, demonstrating a superior performance. With 10% labeled data, the segmentation performance further improves, with an OA of 98.0% and an mIOU of 81.8%. At this labeling rate, the proposed self-training strategy further exploits the supervisory information from the existing labeled data, promoting the improvement of the network's segmentation performance. Compared with the representative fully supervised networks RandLA-Net, KPConv, and Point-Transformer, RV-Net has a higher overall segmentation accuracy, with the Trucks and Fences categories having larger errors.

The results of DR-Net and RV-Net are selected for simultaneous visualization. The segmentation results are shown in Figs. 21 and 22, and the segmentation error maps are also provided for judgment. Generally speaking, due to the relatively simple types of ground objects and the significant spatial differences among the point clouds of the various ground objects, good segmentation results are achieved in the Ground, Buildings, Power lines, and Vegetation point clouds, which is consistent with the quantitative index performance. One of the categories with poor segmentation performance is the Trucks category. It is speculated that this is because the Cars category also exists in the labels, and Cars and Trucks belong to the same major category, making it difficult to capture the differences between them based solely on spatial information. The segmentation accuracy for the Fences category is also relatively low, which is due to the sparse point cloud of the Fences category. As the airborne sensor collects data from the air, it is difficult to obtain high-density point clouds of this category. Moreover, this is a large-scale airborne point cloud dataset. To reduce the number of input point clouds, preliminary grid downsampling is often performed (RV-Net uses a grid downsampling size of 0.32 m, as per the aforementioned research setting), which affects the segmentation performance for smaller categories such as Fences, Power lines, and Trucks. The error maps in Figs. 21 and 22 show that, compared to DR-Net, RV-Net has fewer errors in categories such as Poles and Buildings, achieving better segmentation results.

V. Conclusion

In this article, we have proposed a point cloud semantic segmentation algorithm—RV-Net—for both fully supervised

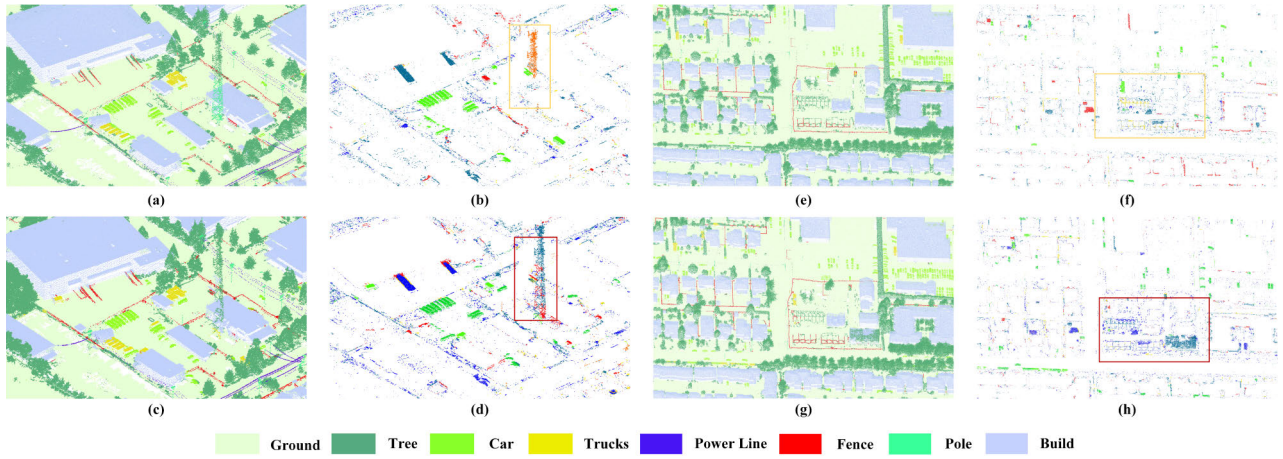

Fig. 21. Details of the weakly supervised segmentation results on the DALES dataset. (a) and (e) RV-Net. (b) RV-Net error. (c) and (g) DR-Net. (d) RV-Net. (d) and (h) DR-Net error. (f) RV-Net error.

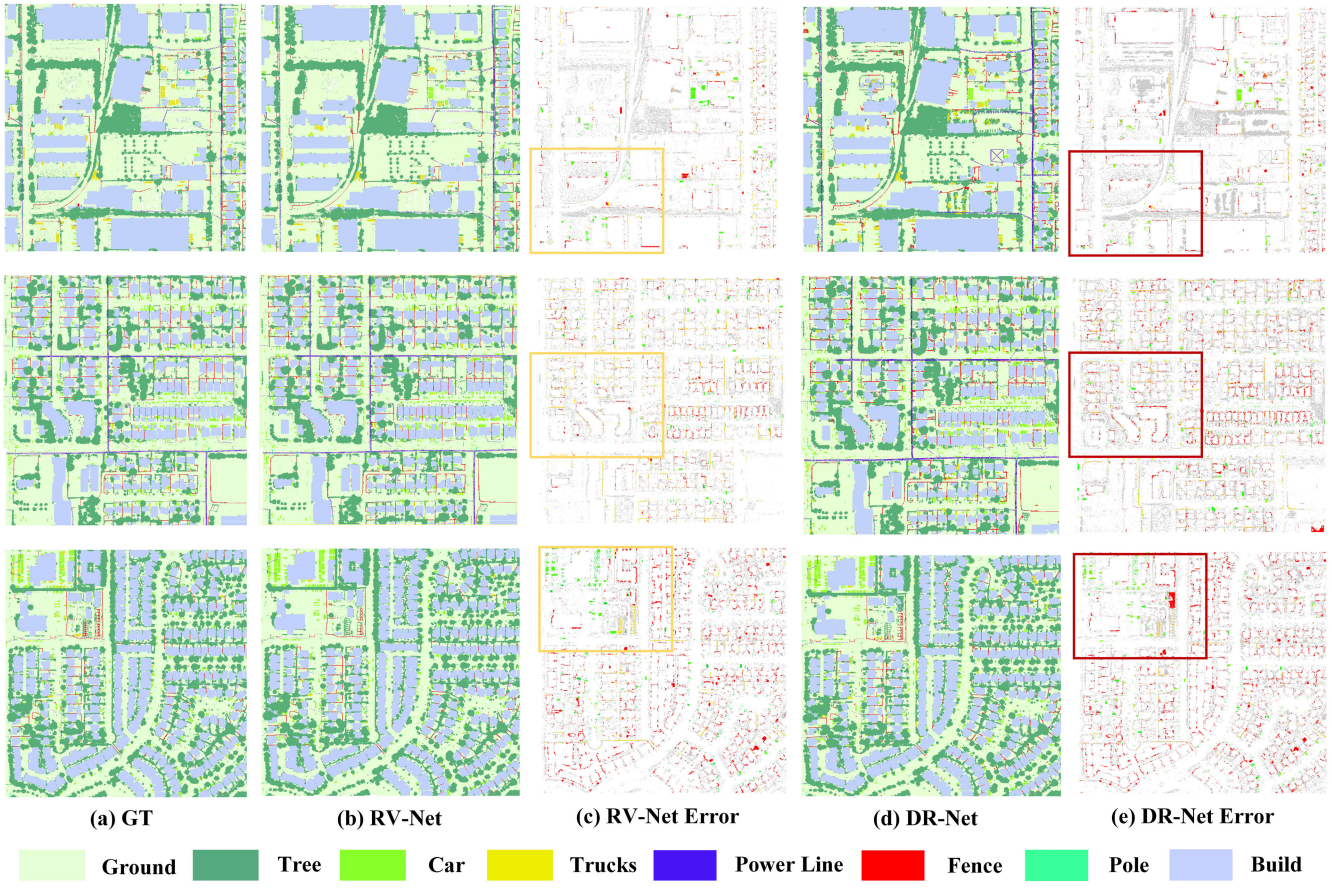

Fig. 22. Overall results of weakly supervised segmentation of the DALES dataset. (a) GT. (b) RV-Net. (c) RV-Net error. (d) DR-Net. (e) DR-Net error.

and weakly supervised scenarios. In RV-Net, from the perspective of expanding receptive fields, a graph-edge-coupled feature encoding module is constructed to enrich the network's extraction of local and global information. Meanwhile, a vector FP module is introduced to enhance the network's ability to extract local discriminative information, thereby improving the network's performance. Finally, noisy labels are utilized to promote the network to learn more robust feature boundaries, reducing the probability of overfitting and the problem of sample imbalance. RV-Net integrates the above modules and strategies and achieves significant improvements in segmentation performance on multiple public datasets. We also conducted a series of ablation experiments to verify the effectiveness of each module and strategy in RV-Net. In addition, due to the large receptive fields in local regions of this network, similar to the principles of SQN, PointCT, and DR-Net in enhancing the performance of small sample segmentation, the expansion of local receptive fields can

effectively improve the learning efficiency of the network with a small number of samples. Therefore, this network is also applicable to weakly supervised point cloud segmentation. The introduced noisy labels can be considered as pseudo-labels. Based on RV-Net, multiple pseudo-labels are simultaneously introduced from different perspectives. Through experiments on multiple public datasets with different proportions, it was verified that RV-Net demonstrates an outstanding generalization performance in scenarios with weakly labeled small-scale samples.

The algorithm proposed in this article aims to enhance the segmentation performance for point cloud data and the learning efficiency for samples, with a relatively low computational cost. By comparing the fully supervised and weakly supervised segmentation results, it can be found that the performance difference between the two is small, which indicates that the current fully supervised algorithms do not make full use of a large number of labeled samples, and there is redundancy in the training data. However, the network overfits the larger categories seriously, and it is difficult to solve the long-tail effect problem with the current deep network. Therefore, increasing the labeling ratio of small sample data and improving the network's segmentation performance for such data will be part of our future research work.

REFERENCES

[1] A. Prabhakara et al., "High resolution point clouds from mmWave radar," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 4135–4142.

[2] J. Wang, Y. Liu, H. Tan, and M. Zhang, "A survey on weakly supervised 3D point cloud semantic segmentation," *IET Comput. Vis.*, vol. 18, no. 3, pp. 329–342, Apr. 2024.

[3] J. Han, K. Liu, W. Li, G. Chen, W. Wang, and F. Zhang, "A large-scale network construction and lightweighting method for point cloud semantic segmentation," *IEEE Trans. Image Process.*, vol. 33, pp. 2004–2017, 2024.

[4] J. Li, J. Wang, and T. Xu, "PointGL: A simple global-local framework for efficient point cloud analysis," *IEEE Trans. Multimedia*, vol. 26, pp. 6931–6942, 2024.

[5] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[6] X. Lai et al., "Stratified transformer for 3D point cloud segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8500–8509.

[7] M. Kolodiazhnyi, A. Vorontsova, A. Konushin, and D. Rukhovich, "OneFormer3D: One transformer for unified point cloud segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 20943–20953.

[8] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16259–16268.

[9] G. Qian et al., "PointNeXt: Revisiting PointNet++ with improved training and scaling strategies," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 23192–23204.

[10] C. Zhou, Y. Dong, M. Hou, Y. Ji, and C. Wen, "MP-DGCNN for the semantic segmentation of Chinese ancient building point clouds," *Heritage Sci.*, vol. 12, no. 1, p. 249, Jul. 2024.

[11] Z. Zeng, H. Qiu, J. Zhou, Z. Dong, J. Xiao, and B. Li, "PointNAT: Large-scale point cloud semantic segmentation via neighbor aggregation with transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5704618.

[12] Q. Hu et al., "RandLA-Net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11108–11117.

[13] I. Armeni et al., "3D semantic parsing of large-scale indoor spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1534–1543.

[14] W. Tan et al., "Toronto-3D: A large-scale mobile LiDAR dataset for semantic segmentation of urban roadways," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 202–203.

[15] Q. Hu et al., "SQN: Weakly-supervised semantic segmentation of large-scale 3D point clouds," in *Proc. Eur. Conf. Comput. Vis.*, Jul. 2022, pp. 600–619.

[16] A.-T. Tran, H.-S. Le, S.-H. Lee, and K.-R. Kwon, "PointCT: Point central transformer network for weakly-supervised point cloud semantic segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 3556–3565.

[17] L. Zhang and Y. Bi, "Weakly-supervised point cloud semantic segmentation based on dilated region," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.

[18] P. Wang, W. Yao, and J. Shao, "One class one click: Quasi scene-level weakly supervised point cloud semantic segmentation with active learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 204, pp. 89–104, Oct. 2023.

[19] M. Li et al., "HybridCR: Weakly-supervised 3D point cloud semantic segmentation via hybrid contrastive regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14930–14939.

[20] N. Varney, V. K. Asari, and Q. Graehling, "DALES: A large-scale aerial LiDAR data set for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 186–187.

[21] T. Le and Y. Duan, "PointGrid: A deep network for 3D shape understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9204–9214.

[22] B. Peng et al., "OA-CNNs: Omni-adaptive sparse CNNs for 3D semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 21305–21315.

[23] E. Kalogerakis, M. Averkiou, S. Maji, and S. Chaudhuri, "3D shape segmentation with projective convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3779–3788.

[24] C. Luo et al., "MVP-Net: Multiple view pointwise semantic segmentation of large-scale point clouds," 2022, *arXiv:2201.12769*.

[25] C. Qi, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5105–5114.

[26] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on $\mathcal{X}$-transformed points," 2018, *arXiv:1801.07791*.

[27] W. Wu, Z. Qi, and L. Fuxin, "PointConv: Deep convolutional networks on 3D point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9621–9630.

[28] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6411–6420.

[29] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual MLP framework," 2022, *arXiv:2202.07123*.

[30] X. Wu et al., "Point transformer V2: Grouped vector attention and partition-based pooling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 33330–33342.

[31] X. Wu et al., "Point transformer V3: Simpler, faster, stronger," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 4840–4851.

[32] H. Zhu et al., "PonderV2: Pave the way for 3D foundation model with a universal pre-training paradigm," 2023, *arXiv:2310.08586*.

[33] J. Wei, G. Lin, K.-H. Yap, T.-Y. Hung, and L. Xie, "Multi-path region mining for weakly supervised 3D semantic segmentation on point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4383–4392.

[34] Z. Liu, X. Qi, and C. W. Fu, "You only need one thing one click: Self-training for weakly supervised 3D scene understanding," in *Deep Learning For 3D Vision: Algorithms and Applications*. Singapore: World Scientific, 2024, pp. 57–89.

[35] Y. Zhang, Y. Qu, Y. Xie, Z. Li, S. Zheng, and C. Li, "Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15520–15528.

[36] P. Wang and W. Yao, "A new weakly supervised approach for ALS point cloud semantic segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 237–254, Aug. 2022.

[37] Y. Lan et al., "Weakly supervised 3D segmentation via receptive-driven pseudo label consistency and structural consistency," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, pp. 1222–1230.

[38] Y. Su, M. Cheng, Z. Yuan, W. Liu, W. Zeng, and C. Wang, "Multistage scene-level constraints for large-scale point cloud weakly supervised semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5705118.

[39] K. Liu et al., "Weakly supervised 3D scene segmentation with region-level boundary awareness and instance discrimination," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 37–55.

[40] C.-K. Yang, J.-J. Wu, K.-S. Chen, Y.-Y. Chuang, and Y.-Y. Lin, "An MIL-derived transformer for weakly supervised point cloud segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11830–11839.

[41] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, Oct. 2019.

[42] R. Huang, Y. Xu, D. Hong, W. Yao, P. Ghamisi, and U. Stilla, "Deep point embedding for urban classification using ALS point clouds: A new perspective from local to global," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 62–81, May 2020.

[43] R. Huang, D. Hong, Y. Xu, W. Yao, and U. Stilla, "Multi-scale local context embedding for LiDAR point cloud classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 721–725, Apr. 2020.

[44] H. Gong, H. Wang, and D. Wang, "Multilateral cascading network for semantic segmentation of large-scale outdoor point clouds," 2024, *arXiv:2409.13983*.

[45] B. Guo, L. Deng, R. Wang, W. Guo, A. H.-M. Ng, and W. Bai, "MCTNet: Multiscale cross-attention-based transformer network for semantic segmentation of large-scale point cloud," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5704720.

[46] X. Zhang, D. Lin, and U. Soergel, "Target-aware attentional network for rare class segmentation in large-scale LiDAR point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 220, pp. 32–50, Feb. 2025.

[47] J. Han et al., "Subspace prototype guidance for mitigating class imbalance in point cloud semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 255–272.

[48] W. Hu et al., "GAN-assisted road segmentation from satellite imagery," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 21, no. 1, pp. 1–29, 2024.

[49] P. Wang, W. Yao, J. Shao, and Z. He, "Test-time adaptation for geospatial point cloud semantic segmentation with distinct domain shifts," *ISPRS J. Photogramm. Remote Sens.*, vol. 229, pp. 422–435, Nov. 2025.

[50] J. Jiang et al., "PCoTTA: Continual test-time adaptation for multi-task point cloud understanding," in *Proc. Adv. Neural Inf. Process. Syst*, vol. 37, 2024, pp. 96229–96253.

[51] S. Qiu et al., "Subclassified loss: Rethinking data imbalance from subclass perspective for semantic segmentation," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 1547–1558, Jan. 2024.

[52] H. Deng, S. Chen, X. Zhu, B. Jiang, K. Jing, and L. Wang, "EP-net: Improving point cloud learning efficiency through feature decoupling," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, 2024.

[53] J. Du et al., "ResDLPS-Net: Joint residual-dense optimization for large-scale point cloud semantic segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 182, pp. 37–51, Dec. 2021.

[54] H. Shuai, X. Xu, and Q. Liu, "Backward attentive fusing network with local aggregation classifier for 3D point cloud semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 4973–4984, 2021.

[55] S. Qiu, S. Anwar, and N. Barnes, "Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1757–1767.

[56] Z. Zeng, Y. Xu, Z. Xie, W. Tang, J. Wan, and W. Wu, "LEARD-Net: Semantic segmentation for large-scale point cloud scene," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102953.

[57] Y. Xu et al., "NeiEA-NET: Semantic segmentation of large-scale point cloud scene via neighbor enhancement and aggregation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 119, May 2023, Art. no. 103285.

[58] X. Li, Z. Zhang, Y. Li, M. Huang, and J. Zhang, "SFL-NET: Slight filter learning network for point cloud semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5703914.

[59] Z. Zeng, Y. Xu, Z. Xie, W. Tang, J. Wan, and W. Wu, "Large-scale point cloud semantic segmentation via local perception and global descriptor vector," *Expert Syst. Appl.*, vol. 246, Jul. 2024, Art. no. 123269.

[60] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4558–4567.

[61] N. Varney and V. K. Asari, "Pyramid point: A multi-level focusing network for revisiting feature layers," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 6502305.

[62] D. Robert, H. Raguet, and L. Landrieu, "Efficient 3D semantic segmentation with superpoint transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 17195–17204.

[63] P. H. Akwensi, R. Wang, and B. Guo, "PReFormer: A memory-efficient transformer for point cloud semantic segmentation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 128, Apr. 2024, Art. no. 103730.