Point-SCT: A Multiscale Spatial Convolution-Swin Transformer Network for Point Cloud Ground Filtering in Complex Mountainous Terrains

Jingxiang Li, Fuquan Tang, Lingfei Ma, *Member, IEEE*, Chao Zhu, Zheng Gong, Nur Intan Raihana Ruhaiyem, and Jonathan Li, *Fellow, IEEE*

***Abstract*— Deep learning (DL)-based point cloud segmentation methods have been extensively explored, but the majority focus either on local or global feature learning, with few integrating both. These integrated approaches have not been sufficiently explored in complex mountainous scenes with low feature heterogeneity. To address this gap, we propose a novel point-based multiscale spatial Convolution-Swin Transformer network (Point-SCT). Point-SCT combines convolutional local geometric detail capture with global relationship modeling via dynamic window interactions in Transformer, enhancing ground filtering accuracy in challenging mountainous scenes. The encoder incorporates a convolution-based multiscale local feature aggregation (MLFA) approach, integrating local geometric feature encoding local geometric structure encoding (LGSE) and diluent pooling (DP) strategies to effectively aggregate local detailed geometric features while suppressing irrelevant feature vectors and enhancing the representation of low-heterogeneity feature. Additionally, the dynamic spatial window strategy within the Transformer facilitates the capture of long-range feature dependencies. To mitigate noise introduced by RGB in point cloud overlays and sharpen geometric distinctions between the ground and low-lying vegetation, we introduce boundary detector, curvature, and average elevation (BCE) as prior inputs, replacing RGB. Finally, quantitative and qualitative analyses of Point-SCT are conducted on an airborne laser scanning (ALS) dataset from a mountainous area, with ablation studies validating the effectiveness of LGSE, DP, and BCE. The comprehensive experiments demonstrate that Point-SCT robustly segments ground points in complex mountainous scenes, achieving state-of-the-art levels of accuracy and generalization.**

Received 17 April 2025; accepted 17 May 2025. Date of publication 23 May 2025; date of current version 2 June 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 52274168 and Grant 42101451 and in part by Shaanxi Natural Science Foundation of China under Grant 2016JM5048. *(Corresponding authors: Fuquan Tang; Lingfei Ma.)*

Jingxiang Li, Fuquan Tang, and Chao Zhu are with the Department of Geomatics, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: 1072412273@qq.com; 2504557922@qq.com; zhuchao@stu.ccit.edu.cn).

Lingfei Ma is with the School of Geospatial Artificial Intelligence, East China Normal University, Shanghai 200241, China (e-mail: l53ma@cufe.edu.cn).

Zheng Gong is with the School of Computer Engineering, Jimei University, Xiamen 361021, China (e-mail: zheng.gong@jmu.edu.cn).

Nur Intan Raihana Ruhaiyem is with the School of Computer Sciences, Universiti Sains Malaysia, Penang 11800, Malaysia (e-mail: intanraihana@usm.my).

Jonathan Li is with the Department of Geography and Environmental Environment and the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: junli@uwaterloo.ca).

Digital Object Identifier 10.1109/TGRS.2025.3573023

***Index Terms*— Convolution, deep learning (DL), ground filtering, point clouds, spatial Swin Transformer.**

I. Introduction

With the advancement of 3-D laser scanning technology, the density of laser points from light detection and ranging (LiDAR) sensors has progressively increased, positioning airborne LiDAR as an indispensable tool for obtaining large-scale 3-D topographic maps. As an advanced active remote sensing technology, airborne laser scanning (ALS) offers dynamic flexibility, high resolution, and high accuracy, enabling 3-D observations of large-scale complex terrains and geological structures for geospatial applications such as disaster monitoring [1], subsidence [2], landslide detection [3], and the quantification of soil erosion and deposition [4]. To implement such ALS applications, large-scale and discrete ground point clouds are first transformed into precise digital models, with ground point cloud filtering serving as a crucial prerequisite for accurately extracting ground points and eliminating random noise, speckles, or anomalies at various scales. Subsequently, multitemporal digital models are constructed across the same spatial scale but different temporal dimensions to analyze terrain change patterns under the current spatial framework, conduct natural resource surveys and management, and prevent geological disasters [5]. Hence, the demand for advanced automation, high efficiency, and improved accuracy in ground point filtering methods in mountainous area has been steadily increasing.

Traditional ground point filtering methods are typically divided into three categories based on the geometric attributes of point clouds and filtering rules: slope-based [6], surface-based [7], and morphological-based [8] methods. However, these algorithms often fail in complex terrains characterized by steep slopes, gullies, and dense vegetation as they struggle to accurately capture the inherent features of objects [9]. Over the past years, deep learning (DL) has experienced exponential growth, particularly with significant advancements in the fields of computer vision and remote sensing [10]. Various effective DL-based point cloud segmentation methods have been developed, typically categorized into two types at the preprocessing stage: point compression-based and point convolution-based methods [11].

Point compression-based methods compress point cloud data into multiview images or multidimensional voxels, leveraging intensity and density features to achieve significant progress in several tasks [12], [13], [14], [15]. However, they often neglect semantic information of point cloud elevations and spatial structural features, inevitably introducing quantization errors and failing to exploit the sparse spatial characteristics of raw point clouds. Moreover, unfilled and unused voxels or grids can cause remarkable memory redundancy and computational cost, limiting the scalability in large-scale scenarios [16]. In contrast, point convolution-based methods [19], [20], [21], [22], [23] employ multiple or shared point convolution operators directly on points, followed by object classification or segmentation. Due to their end-to-end simplicity and scalable network architectures, these methods show great potential in engineering applications [22]. However, most CNN-based models focus primarily on aggregating local features and struggle to capture global dependencies across multilevel features, resulting in suboptimal performance for complex mountainous remote sensing tasks [23].

The Transformer [24] has demonstrated exceptional performance in point cloud processing tasks, even surpassing CNN models in certain scenarios. Point clouds, which are inherently point sets embedded in space, contain positional information. Vision transformer (ViT) [25] processes input using self-attention, which provides invariance to sequence and cardinality. This mechanism calculates the similarity between any pair of embedded vectors and computes a weighted sum, thereby establishing relationships among all low-level input vectors and subsequent processing modules. This process allows the model to effectively capture long-range dependencies between positional and feature information, which is crucial for large-scale point cloud tasks [17]. However, modeling these long-range dependencies in ViT necessitates stacking numerous linear projection modules to extract high-level nonlinear features, which increases computational complexity and resource demand while reducing inference efficiency compared to CNNs. To address these limitations, the Swin Transformer [26] introduces a hierarchical feature mapping approach, which reduces computational complexity from quadratic to linear. Combined with multilayer feature fusion designs, such as the feature pyramid network (FPN) [27], it optimizes the network structure, balancing semantic information and positional information across different feature levels while alleviating the computational burden of high-resolution tokens. To further balance local and global feature learning, as well as accuracy and efficiency, the convolutional ViT (CvT) [28] incorporates convolutional modules prior to the ViT input projection layer to aggregate local features from raw point clouds. This method has been shown to achieve more robust model performance and improved computational efficiency.

However, current DL-based point cloud segmentation approaches primarily focus on learning either local or global features, and they have not been sufficiently explored in large-scale mountainous scenes. These environments are characterized by complex terrain changes, geological structures, and dense vegetation coverage, where the ground and non-ground (NG) points exhibit insignificant interclass differences, resulting in complex topological structures that make the fine separation of ground and NG points highly challenging [29]. Furthermore, the existing approaches rely on RGB as auxiliary input features to capture spatial consistency through color coherence [30]. Nevertheless, in these challenging mountainous environments, the quality of optical sensor imagery is often degraded by uneven shading caused by lighting and complex vegetation occlusion, leading to inaccurate color information. Also, the registration process between point clouds and imagery inevitably introduces noise and errors. Therefore, efficiently and accurately extracting local geometric details and global relational features from discrete point clouds using DL techniques remains a highly challenging and promising research topic.

Motivated by CvT and Swin Transformer, we hypothesize that combining the robust local spatial feature perception capabilities of CNNs and global feature relationship modeling of multiscale shifted spatial windows Transformer can significantly improve the performance and accuracy of ground filtering tasks in complex mountainous scenes. To this end, we propose the Point-SCT model, an end-to-end ground filtering framework tailored for large-scale remote sensing scenes characterized by challenging terrains, geological structures, and vegetation environments. The model architecture incorporates a multilevel residual structure to capture multilevel and multiscale feature representations of unstructured point clouds. In the encoder, we employ a random sampling (RS) strategy [21] and introduce a novel CNN-based multiscale local feature aggregation (MLFA) module to efficiently aggregate multiscale local geometric features and prevent the loss of critical features during RS. The MLFA consists of two innovative submodules: local geometric structure encoding (LGSE) and diluent pooling (DP). Furthermore, we design the spatial Swin Transformer (SSwin Transformer), leveraging shifted spatial windows to enable hierarchical global geometric feature interactions. The logic of synergistic interaction between the MLFA and spatial Swin Transformer modules is illustrated in Fig. 1. Additionally, to mitigate the noise introduced when acquiring RGB data, caused by the overlay effect in point clouds, and to enhance the distinction of surface geometric features and interclass boundaries between the ground and NG, we propose a feature combination derived from the geometric relationships of point neighborhoods, including boundary detector, curvature, and average elevation (BCE) as input features.

Our main contributions are summarized as follows.

1) We develop a novel ground filtering model, Point-SCT, which integrates CNNs with spatial Swin Transformers, focusing on the aggregation of local spatial geometric information, expanding the scope of information interaction, and establishing long-range and cross-domain information dependencies.
2) In the encoder, we propose a novel MLFA module, incorporating LGSE and DP submodules that enable high-density aggregation and adaptive preservation of point neighborhood detail geometric features.

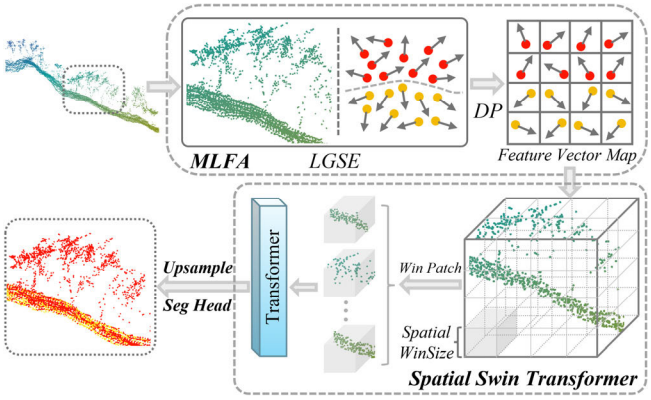

Fig. 1. Synergistic interaction between the MLFA and spatial Swin Transformer modules is demonstrated on a small sample of raw point cloud data from mountainous terrain. Red points represent NG points, while yellow points represent ground points.

3) We implement the prior BCE feature combination tailored to the structural design of Point-SCT, replacing traditional RGB inputs to mitigate overlay noise and enhance interclass feature distinction.

II. Related Work

This section reviews and summarizes previous DL-based intelligent point cloud processing methods, with a particular focus on point convolution and Transformer-based methods relevant to this article.

A. Point Compression-Based Methods

Point compression-based methods involve point voxelization and projection regularization. The voxelization method converts irregular 3-D points into a regular voxel grid for 3-D convolution. VoxNet [15] was an early example, but its voxel count grows exponentially with point resolution, leading to high computational and memory redundancy. To address this, Riegler et al. [31] introduced OctNet, which uses a multilevel octree structure to store high-resolution point clouds efficiently. Graham et al. [32] proposed sparse convolutions that focused only on valid voxels to improve data processing efficiency. Projection regularization methods project 3-D points onto 2-D planes [33], [34], extract features via 2-D convolutions, and integrate multiview features with max pooling (MP). Accordingly, TangentConv [35] addressed global projection-induced point loss by first projecting local neighborhoods onto tangent planes before applying 2-D convolutions. However, it relies on tangent estimation, inevitably leading to information folding and systematic errors. These methods are vulnerable to point occlusion, overlap, and density variations, resulting in the loss of spatial semantics and geometric relationships during projection.

B. Point Convolution-Based Methods

To address the limitations of point compression-based methods, recent studies have introduced approaches that directly use 3-D point coordinates and associated features. PointNet [19], a milestone in point cloud DL, primarily aggregated single-point features using shared multilayer perceptrons (MLPs) and global pooling, but it downplayed the influence of point density and overlooked local geometric relationships. To enhance the model's sensitivity to local geometric features, PointNet++ [20] incorporated an FPN structure, sampling the farthest points to obtain centroids and using a query ball search for multiscale local feature extraction. PointCNN [36] improved spatial and local information understanding by linearly transforming feature weights and potential sequences of input points using MLPs. Furthermore, PointConv [37] trained MLPs on local points and applied the weights and density functions of local points to convolutional kernels, introducing 3-D continuous convolution that enhanced local feature aggregation, though it increased kernel complexity and hindered network convergence. KPConv [18] introduced kernel point convolutions in Euclidean space, adjusting kernel focus areas with points within a neighborhood radius, tackling the limitations of previous 3-D convolutions kernel numbers, and improving adaptability to local detailed features.

Another method for point neighborhood feature aggregation uses graph-based CNNs to model the geometric relationships of local points through graphs. DGCNN [14] constructed dynamic neighborhoods and aggregated points using symmetric functions before extracting features through graph convolution. Similarly, 3DGCN [38] combined vector and scalar features of points, aggregating neighborhood points by transforming relative position vectors. While graph-based methods improve the accuracy in small-scale tasks, they often neglect broader spatial relationships, limiting scalability to large scenes.

C. ViT-Based Methods

In the realm of 3-D vision, the inherent permutation invariance and rich semantic representation capabilities of the Transformer have led to the widespread adoption of point-based Transformer methods for 3-D classification and segmentation, resulting in significant achievements. These methods are generally categorized into two types: global point Transformers and local point Transformers [39].

Global point Transformers [40], [41], [42], [43], [44], [45], [46] directly model global relationships. The point cloud Transformer (PCT) [42] integrated spatial positioning of point clouds with sequential positioning of tokens, enabling global feature learning through multilevel Transformers. Zhong and Han [44] introduced a pyramidal structure to capture features at varying resolutions through farthest point sampling (FPS), constructing interrelations to facilitate cross-scale interactions, and modeling global features. However, these methods face challenges with large-scale outdoor scenes due to high computational cost and memory usage. Robert et al. [43] addressed this by using superpoints and incorporating sparse self-attention, but it requires presetting the number of superpoints, as improper tuning can lead to feature loss and misclassification. In contrast, Hui et al. [41] integrated the pyramid PCT (PPT) network, which employed an FPN architecture and integrated the EdgeConv method for multiscale feature aggregation using k-nearest neighbor (KNN) graphs.

Local point Transformers [46], [47] divided the global point cloud into subsets and focused on extracting local neighborhood features. Point Transformer [47] incorporated vector attention mechanisms with trainable positional encodings and employed a KNN graph to amalgamate point subsets with preceding layer features, enhancing information transfer across Transformer blocks and ensuring the network's efficacy in large-scale scenes. Lai et al. [17] proposed the stratified Transformer method, which dynamically divided the point cloud into spatial windows, allowing distant sparse points to aggregate long-range spatial information. While this method improves the task accuracy, it increases computational load due to the substantial linear transformations in multistage Transformers. Moreover, Wu et al. [28] proposed the CvT, combining CNNs for local feature modeling with Transformers for long-sequence relationship modeling, achieving significantly greater efficiency than traditional ViT and DeiT in 2-D visual tasks. Lu et al. [48] extended this approach with the multiscale 3DCTN framework, showing high performance on datasets such as ModelNet40 and ScanObjectNN.

D. DL-Based Ground Point Filtering Methods

Recent DL-based ground point filtering methods primarily focus on point compression and point convolution. However, Transformer-based methods for ground point filtering remain unexplored. Yotsumata et al. [49] utilized a voxel-based CNN approach for filtering ground points in ALS data from forested areas, but dense voxel compression lost geometric structure information and the network architecture struggled to capture point relations. Hu and Yuan [50] applied a windowing strategy with elevation differences and projected points onto 2-D images for binary classification, while Rizaldy et al. [9] accelerated computation by projecting points onto multidimensional images in a fully convolutional network (FCN). Though these methods extract features at multiple scales, they neglect the spatial structure relationships of the point clouds. Janssens-Coron and Guilbert [51] applied PointNet to ALS point clouds from the forest, but its performance in large-scale feature learning was limited. Furthermore, Krisanski et al. [52] employed an improved PointNet++ to segment ground points in densely vegetated forest, and Kada and Kuramin [53] demonstrated KPConv's robust performance in large-scale scenes. In summary, although a few point-based DL methods show promise, their limited convolutional receptive fields pose challenges in handling the spatial resolution and construction of global feature relationships in large-scale ALS point clouds, particularly in discontinuous complex terrains with covering dense vegetation.

E. Shallow Neighborhood-Based Point Geometric Features

The unordered raw point cloud data contain significant spatial structure features. Both traditional and DL-based methods rely on spatial geomatic semantic attributes. For example, Guerrero et al. [54] proposed a data-driven method using shallow point neighborhoods with pre-quantified local shape and structural relationships serving as input features, showing that learning local normal vectors and surface curvature improves classification accuracy. Hao et al. [55] incorporated boundary detection within the model's shallow layers to guide supervisory signals and reduce interclass semantic ambiguity. Cai et al. [10] replaced RGB input with a combination of point cloud intensity, enhanced elevation, and range images, showing that pre-quantified local geometric features can improve segmentation accuracy. Thus, providing rich shallow local spatial and topological information is essential for enhancing the model's ability to capture fine details.

III. Method

In this section, we provide a detailed explanation of the proposed Point-SCT model architecture, which offers a novel solution for large-scale 3-D point cloud ground filtering through a multilevel hierarchical structure. Specifically, we focus on the core components of the model, i.e., the convolution-based MLFA module, the SSwin Transformer (SST) module, and the upsamling module. Additionally, we introduce a feature combination tailored for this model framework.

A. Overview

Fig. 2 shows the workflow of the proposed Point-SCT network. This framework directly processes raw point clouds containing coordinates (X, Y, and Z) and derived geometric features (i.e., boundary detector, curvature, and average elevation). Point-SCT architecture adopts an encoder–decoder structure, integrating multiple layers of local feature aggregation mechanisms, including MLFA modules, RS layers, and SST modules.

Initially, the stem projection module maps the input point cloud to a unified high-dimensional feature space using nonlinear transformations to enhance feature expressiveness. Subsequently, the projected features are fed into the encoder, where MLFA and SST asynchronously model the local and global spatial features of the point cloud within the same stage. In the MLFA mechanism, high-dimensional neighborhood geometric features of shallow input points are encoded and embedded to aggregate shallow high-dimensional local semantic information. Specifically, LGSE encodes and embeds high-dimensional spatial vectors of the input points' neighborhood geometric features, enhancing the capture and aggregation of local spatial features through relative position encoding, while the DP module conducts unsupervised learning on the encoded features, dynamically allocating weights and adaptively optimizing and filtering significant geometric details of the neighborhood points. The RS layer rapidly reduces the density of high-dimensional feature vectors, allowing the encoder to effectively balance sampling efficiency with the preservation of local feature details. LGSE and DP modules are connected via skip connections, gradually integrating high-dimensional semantic features while enhancing the perception of local spatial geometric information, and effectively mitigating the loss of key features caused by RS. Following the integration of powerful local modeling capabilities from CNNs, the SST module builds long-term context associations

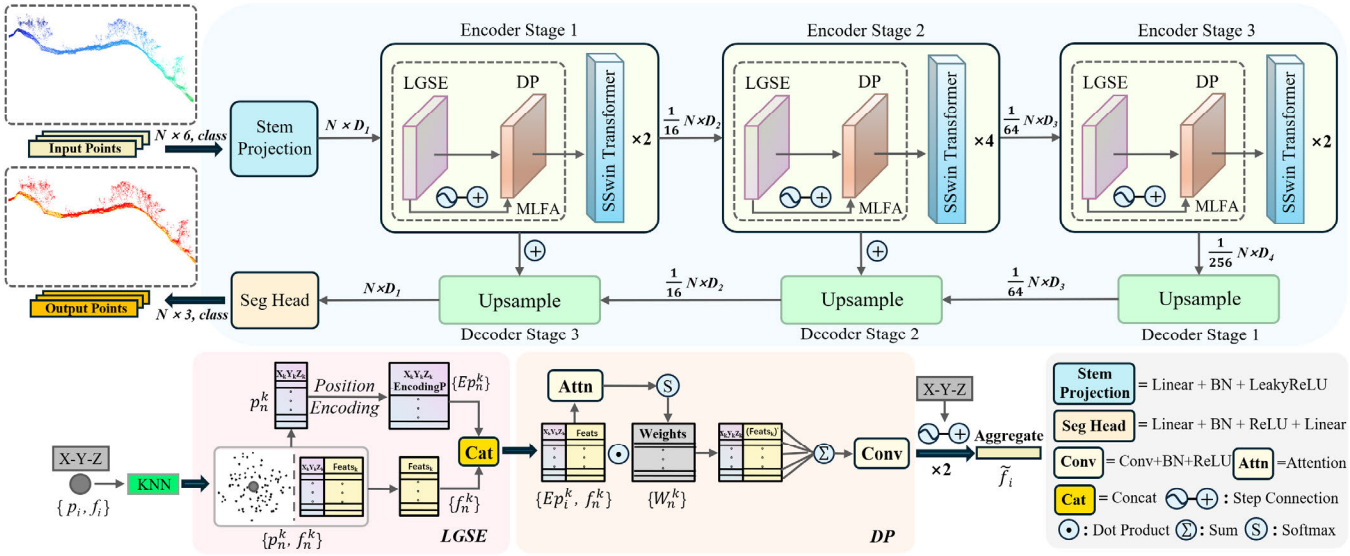

Fig. 2. Point-SCT network architecture.

among multiscale spatial feature vectors, achieving a comprehensive capture of both local and global feature information. The decoder adopts a cross-layer linking architecture similar to FPN, concatenating feature output points corresponding to the encoder. Layer normalization (LN), linear, and interpolation operations are applied to aggregate the multiscale feature vectors extracted by the above modules, capturing extensive and detailed contextual spatial semantic information. Finally, the Seg Head layer performs classification prediction on the upsampled spatial vectors.

In our proposed network, the backbone's RS has a time complexity of $O(1)$ with multilevel sampling rates of 1/4, producing point feature vector sets at each level with cardinalities of $[N, 1/16N, 1/64N, 1/256N]$, where N is the number of input points. The model achieves a balance among sampling efficiency, local detail preservation, and effective global long-range feature association, making it highly suitable for extracting complex geometric structural features.

B. Multiscale Local Feature Aggregation

The multiscale feature aggregation mechanism primarily incorporates the LGSE and DP submodules to achieve high aggregation and optimized retention of local features. Our sampling mechanism employs efficient RS. However, as point density becomes increasingly sparse during single-stage encoding, there is a risk of critical feature loss. To address this, the stacked design of LGSE and DP in the MLFA module incorporates two skip connections, as illustrated in Fig. 2. These residual connections gradually integrate surface spatial information with multiscale local features, thereby expanding the perception of local features and enhancing information propagation. Additionally, this design allows the original surface spatial information and neighborhood geometric relationships to be transmitted into the features after sampling, mitigating the potential feature loss caused by the sampling process. Within this mechanism, LGSE focuses on aggregating the local geometric structure features from the input high-dimensional spatial vectors, while DP dynamically retains the feature vector outputs by LGSE.

1) LGSE: Given a set of input points in high-dimensional space $P = \{p_1, p_2, \ldots, p_i\}_{i=N}$, where $N \in \mathbb{R}$ and $p_i \in \mathbb{R}^{N\times 3}$, each point carries its original features, i.e., boundary detector, curvature, and average elevation, forming a feature set $F = \{f_1, f_2, \ldots, f_i\}_{i=N}$, with $f_i \in \mathbb{R}^{N\times D_{\text{in}}}$ and D_{in} denoting the dimension of input features. To better highlight the relative spatial positions and geometric relationships between the central point and its neighboring points, this module uses a spatial relative position encoding strategy to aggregate the neighborhood point set of the central point p_i, thereby facilitating the model's deep analysis of complex local feature structures in shallow spatial layers. Specifically, before applying LGSE, we first define the neighborhood radius based on Euclidean distance by using KNN to search for neighboring points for each point in P, obtaining the spatial indices of k neighboring points corresponding to the central point p_i. Next, we capture the coordinates and features of the neighboring points p_n^k and f_n^k using the neighborhood aggregation function g, as shown in the following equation:

$$\begin{cases} p_n^k = g(P, \text{neighbor } Idx), & \text{if neigbor } Idx \le N \\ f_n^k = g(F, \text{neighbor } Idx), & \text{otherwise.} \end{cases} \tag{1}$$

Subsequently, we calculate the relative spatial coordinates and Euclidean distance between the central point p_i^k and its neighboring points p_n^k, constructing a position encoding vector. This vector contains the spatial geometric reference information between the central point and its neighbors and is then embedded as a feature vector into the central point, yielding Ep_i^k, as shown in the following equation:

$$Ep_i^k = \text{Concat}\left(\begin{array}{c} p_i, p_i^k, (p_i - p_i^k) \\ \sum_{i=0}^{n} \sqrt{(p_i - p_i^k)^2}, \text{dim} = 1 \end{array}\right). \tag{2}$$

On this basis, we embed the feature vector f_n^k, which carries strong geometric shape information, into the central

point, resulting in $\vec{F} = \vec{f}_i = (Ep_i^k,\ f_n^k)\}$. LGSE prevents the gradual weakening or loss of features in sparse point clouds during multistage encoding. Particularly in spatial point clouds with locally complex structures, LGSE effectively integrates multidimensional spatial feature vectors with different structural information, significantly enhancing the nonlinear mapping capacity of the feature space. This enhancement enables neurons to precisely capture subtle local geometric variations and their complex spatial relationships.

2) DP: In most existing point cloud processing methods [20], [47], a fixed pooling mechanism (i.e., MP) is typically used to actively retain significant sampled point features, often requiring specific sampling methods (i.e., FPS). However, for sparse point clouds and point clouds with complex geometric structures, fixed pooling often proves ineffective, leading to information loss and wasted computational resources. To address this issue, this study proposes an efficient DP strategy, which integrates an attention neuron module based on unsupervised learning. This module automatically learns potential local geometric features to allocate vector weights, achieving effective dilution of high-density features through weight-based constraints on the feature vectors. Specifically, the DP module receives the coordinates of the central points from the previous layer and the aggregated neighborhood geometric feature vector set $\vec{F}$ as the input, with the shape defined as (i, j). The DP module defines a neuron attention function Attn$(\cdot)$, which does not include nonlinear activation, mapping the central point feature vectors to attention weights. Using this attention function and Softmax operation, the attention distribution matrix W_{ij} for the feature vectors is computed from the vector set $\vec{F}$, where each element of the matrix represents the assigned positive weight w_{ij}, as shown in the following equation:

$$\begin{aligned} w'_{ij} &= \text{Attn}\left(\vec{f}_i\right) + \vec{f}_i \\ w_{ij} &= \frac{e^{w'_{ij}}}{\sum_{i,j} e^{w'_{ij}}} \\ W_{ij} &= \begin{bmatrix} w_{11} & \cdots & w_{1j} \\ \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{ij} \end{bmatrix}, \quad i \le N \text{ and } j = D \end{aligned} \tag{3}$$

where w'_{ij} represents the preliminary measure of the closeness of the geometric relationship between the central point i and the neighboring point j in the feature map, and w_{ij} is normalized using Softmax applied to w'_{ij}, with D denoting the sum of the central point's coordinates $(X, Y,$ and $Z)$ and the embedded feature vectors. Furthermore, to propagate the constraints imposed by the current weights on the features to the neurons in the next encoding stage, the feature dilution matrix is processed through the Conv function, yielding the diluted feature set $\tilde{f} = \{\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_i\}_{i=N}$, with $\tilde{f}_i \in N \times D_{\text{out}}$ and D_{out} denoting the dimension of output features, and this is passed as the input to the next processing unit. The specific (4) is as follows:

$$\tilde{f}_i = \text{Conv}\left(\sum_i^N \frac{F_{ij} \cdot W_{ij}^T}{\sqrt{D}}\right) \tag{4}$$

where W_{ij}^T represents the transpose of the attention distribution matrix W_{ij}, and the function assigns higher scores to feature units with strong geometric expressions while ignoring low-scoring feature vectors. Additionally, the Conv layer incorporates ReLU activation units to correct and amplify the diluted feature attention map, enhancing the expressiveness of high-scoring features after dilution. This strategy allows for a more effective perception of the geometric relationships among neighboring points in subsequent stages while also reducing the impact of low-scoring noise features on the feature extraction of downstream spatial self-attention units.

Point-SCT considers the scenario where the input set P contains any number of necessary feature vector elements. Thus, the target output $\tilde{f}$ dynamically adjusts D_{out} based on the nonfixed D_{in}, enabling the module to handle input data of varying quantities and scales. More importantly, the developed model ensures permutation invariance for the input points p_i. Specifically, for any permutation α, the following equation holds in the above equations' calculations:

$$\begin{aligned} &\text{Funcs}\left(\{p_1, p_2, \ldots, p_i\}, W_{ij}\right) \\ &\quad = \text{Funcs}\left(\{p_{\alpha 1}, p_{\alpha 2}, \ldots, p_{\alpha i}\}, W_{ij}\right) \end{aligned} \tag{5}$$

where Funcs performs the operations of the LGSE and DP modules. This ensures that the output vectors remain consistent and effective, regardless of the order of input points.

C. Spatial Swin Transformer

Building on the previous module that aggregates local features, the spatial Swin Transformer further constructs long-range spatial information links among locally rich geometric features, enabling the learning of global point geometric features. This module consists of a multihead self-attention module with an offset operator and an LBR [56] module (e.g., Linear, BatchNorm, and ReLU units), as shown in Fig. 3. This method addresses large-scale and complex point clouds with a space complexity below $O(N^2)$, surpassing traditional point-wise or channel-wise learning approaches, where N denotes the number of input points. The core idea is to adopt dynamic spatial patches to partition the input spatial feature vectors into nonoverlapping spatial windows. Each window has a unique query point that attends to all points within that window, and every window executes a shared multihead self-attention module independently. The outputs from different windows are subsequently integrated via interactive operations, thereby effectively capturing global spatial feature correlations.

Specifically, given a set of sampled feature vectors $\tilde{f}$ as input, p_i^w is defined as the number of feature points within the ith window, W_{size} is the spatial window size, and S_{size} is the window's sliding offset. The spatial coordinates $p_{\tilde{f}}$ of the feature points are updated with their discrete mapping relationships within the window, represented by the matrix G. Then, the cooperative mapping tensor and feature vector relations are used to filter the valid mapping indices of spatially adjacent relationships, which are used to compute the relative positional bias within the window's attention. The

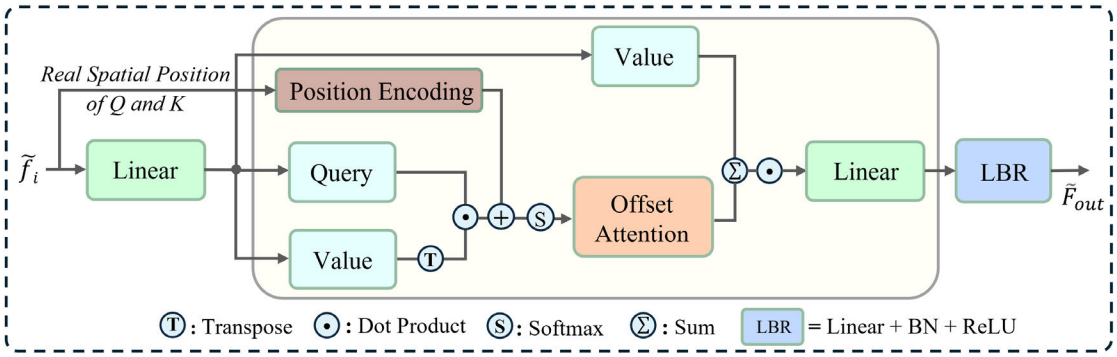

Fig. 3. Structure of the spatial Swin Transformer module.

computation is as follows:

$$G(\tilde{f}) = \left\lfloor \frac{p_{\tilde{f}} + S_{\text{size}}}{W_{\text{size}}} \right\rfloor$$
$$I_m, I_n = \left\{ G\left(\tilde{f}\right), p_{\tilde{f}}, p_i^w, W_{\text{size}} \right\} \tag{6}$$

where $\lfloor \cdot \rfloor$ represents the floor operation to ensure that elements in the mapping tensor remain integers.

In the ith spatial window, the self-attention proceeds as follows. First, given the number of heads H_n in the self-attention mechanism, each with dimension H_d, the query, key, and value matrices are obtained through learnable linear projection modules

$$Q = \text{Linear}_Q\left(\tilde{f}\right) + b_Q$$
$$K = \text{Linear}_K\left(\tilde{f}\right) + b_K$$
$$V = \text{Linear}_V\left(\tilde{f}\right) + b_V \tag{7}$$

where $Q, K \in \mathbb{R}^{N \times d_a}$, $V \in \mathbb{R}^{N \times d_b}$, and d_a and d_b are the dimensions of the query, key, and value vectors, with b representing the bias terms. Simultaneously, Q_m and K_n are passed into the positional encoding module, where their actual spatial positions are retrieved by indexing $Q_m(I_m)$ and $K_n(I_n)$. Then, the relative offset vectors Δpos_{mn} between them are computed based on the subtraction relationship, quantized into discrete plane indices, and ultimately mapped into the initial bias tensor B_{QK} using a lookup attention bias table. The specific computation is as follows:

$$\Delta\text{pos}_{mn} = Q_m(I_m) - K_n(I_n)$$
$$B_{QK}(m,n) = \left\lfloor \frac{\Delta\text{pos}_{mn} + W_{\text{size}}}{\delta} \right\rfloor \tag{8}$$

where W_{size} restricts the quantization range of Q and K, and δ is the quantization scale that ensures that the offset vector is mapped into a discrete quantized space.

Next, after computing the dot product between Q and K to obtain the similarity, the relative positional bias B_{QK} is fused. This bias is converted into a learnable parameter during training, enabling attention to compute similarity based on the relative geometric spatial relationships between Q and K. Through gradient optimization, the distribution of elements in B_{QK} is iteratively refined. In addition, by specifying the initial position of Q within K, B_{QK} quickly pinpoints the range of K that each Q must attend to, avoiding exhaustive calculations for all Q and K pairs within a window and thus accelerating the process. Subsequently, a sparse self-attention module, referred to as offset-attention (OA), is applied to compute the attention weight distribution for the query points. It then uses V to reweight and adjust the contributions of the input features. The specific operation in the ith spatial window is given by the formula

$$\tilde{w}_{i,j} = Q \cdot K^T$$
$$\overline{w}_{i,j} = \frac{\tilde{w}_{i,j}}{\sqrt{d_a}} + B_{QK}$$
$$\tilde{F}_{i,h} = \sum_{j=1}^{p_i^w} \text{OA}\left(\frac{e^{\overline{w}_{i,j}}}{\sum_h e^{\overline{w}_{i,h}}} \right) \times V_{j,h} \tag{9}$$

where $\tilde{w}_{i,j}$ represents the attention weight matrix derived from the dot product between the query and key, and $\overline{w}_{i,j}$ is formed by normalizing and adding the bias B_{QK}. In our implementation, we perform Softmax and OA operations on $\overline{w}_{i,j,h}$ across the multiple heads, aggregate along the column dimension, and finally compute a dot product with the value matrix V to obtain the aggregated feature $\tilde{F} \in \mathbb{R}^{p_i^w \times H_n \times H_d}$. Compared to conventional scalar dot-product attention, this vector-based approach enhances the model's capacity to regulate attention weights on a per-channel basis, making it particularly suitable for large-scale and complex spatial point cloud structures [42].

Finally, we feed the globally sharpened self-attention feature $\tilde{F}$ into the LBR network and apply the global input feature $\tilde{f}$ point-wise. The result is then projected to obtain the output feature map $\tilde{F}_{\text{out}}$ as follows:

$$\tilde{F}_{\text{out}} = \text{LBR}\left(\tilde{F}\right) + \tilde{f}. \tag{10}$$

D. Upsampling Block

As shown in Fig. 2, to address the loss of spatial details in the input stream $\tilde{F}_{\text{out}}$ during the upsampling stage, we integrate the upsampling module with residual connections from the encoder, thereby enhancing the fusion of multilevel representations. Specifically, within the upsampling module, the feature map $\tilde{F}_{\text{out}}$ undergoes a LayerNorm followed by a linear transformation to yield $\tilde{F}'_{\text{out}}$, adjusting the output scale and preventing issues such as gradient vanishing or explosion in large-scale point cloud training. Subsequently, a KNN-based interpolation operation is employed to establish the spatial

mapping relationship between the current stage point cloud coordinates $\tilde{P}_{\text{out}}$ and the preceding encoding stage $\tilde{P}_{\text{out}-1}$. Utilizing this spatial mapping, $\tilde{F}'_{\text{out}}$ is accurately aligned and mapped to another vector space position of $\tilde{F}_{\text{out}-1}$, after which is fused with the features from the preceding encoding stage $\tilde{F}_{\text{out}-1}$.

E. Shallow Point Feature Augmentation

In most DL studies involving point clouds [23], [39], [57], RGB is typically used as the direct input feature for model processing. However, for large-scale ALS point clouds, RGB data are derived by projecting point clouds onto corresponding digital images or by integrating them with orthophotos. For point clouds in complex mountainous areas, this process inevitably introduces significant noise and calibration errors, leading to inaccuracies and biases in the RGB data, which results in unclear boundaries and imprecise predictions of objects. Furthermore, it is extremely challenging to accurately capture the semantic features of geometric spatial consistency through color consistency. To address these issues, this section introduces the geometric augmented features combination specifically designed for our model, namely BCE. These shallow geometric features replace RGB as input features, proving effective in enhancing the distinction of features between points and their adjacent neighbors [10]. Before processing the first layer of input data, each point is embedded with spatial geometric features, which enhances the aggregation of local spatial features.

1) Boundary Detector: This feature can identify the boundary areas between point cloud classes, where homogeneous points usually align in one direction [55]. We use a k-dimensional (K-D) tree to find k nearest neighbors, $P^k = \{p_1, p_2, \ldots, p_j\}_{j=k}$ for a query point, then apply a Gaussian kernel function G to compute the relationship between the average inverse density and k, obtaining the centroid C_i, and finally calculate the distance d between the centroid and the center of mass to identify the boundary region R_i as follows:

$$C_i = \frac{1}{k} \frac{\sum_{j=1}^{k} G\left(x_j y_j z_j\right)}{G(x_i y_i z_i)}$$
$$R_i = \min \| C_i - p_i \|. \tag{11}$$

2) Surface Curvature: For point clouds with complex spatial terrains that include numerous elevations, low-lying areas, and rough edges, we have utilized principal component analysis (PCAs) to capture the diverse geometric transformations within the point clouds, allowing the model to more precisely identify and segment intricate terrain details, which has been proven effective in existing research [58]. Specifically, we query k nearest neighbors p_i for each point P^k, where all neighbors are within a finite subset of Euclidean space $\mathbb{R}^N$. We compute the covariance of these neighboring points p_i as follows:

$$V_i = \frac{1}{k} \sum_{j=1}^{k} \left(\left(p_i - p_j \right) \cdot u \right) \left(\left(p_i - p_j \right) \cdot v \right) \tag{12}$$

where u and v are unit vectors in the inner product space, with V_i denoting the covariance of p_i in the directions of u and v. Subsequently, we perform eigenvalue decomposition on the covariance matrix V, obtaining eigenvalues λ_1–λ_3, (with $\lambda_1 < \lambda_2 < \lambda_3$), and calculate curvature δ using (11)

$$\delta_i = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3}. \tag{13}$$

The smaller δ indicates a smoother geometric composition of p_i with its neighborhood, while a larger δ indicates greater variability.

3) Average Elevation: Elevation is an indispensable feature in point cloud segmentation tasks. Most DL methods for point clouds typically focus on incorporating elevation information only within spatial positioning calculations, often overlooking the importance of encoding elevation features of the neighboring points [59]. To tackle this problem, we propose averaging the elevations of points within neighborhood P^k of p_i, thereby obtaining the average neighborhood elevation E_i as follows:

$$E_i = \frac{1}{k} \sum_{j=1}^{k} z_j \tag{14}$$

where z_j represents the elevation of the jth point about p_i.

IV. Experiment

In this section, we initially provided a detailed introduction to the experiment using Point-SCT for segmenting ground points from a large-scale mountainous ALS dataset, including sources, preprocessing, and implementation details. Subsequently, we selected three representative areas as the test dataset and conducted comprehensive qualitative and quantitative analyses to demonstrate the model's robustness in segmenting ground points in complex environments. Finally, we conducted ablation studies to further validate the effectiveness of each component of the proposed framework.

A. Datasets

To ensure the deployment of the Point-SCT model for large-scale point cloud data in complex terrains and geomorphic structures, we selected a mid-to-low mountainous area in the northwest of Jiangyou City, Sichuan Province, China, as the study area. This region, characterized mainly by low and middle mountains, exhibits significant topographical undulations and complex vegetation environments, as shown in Fig. 4. We employed the DJI M600 drone equipped with the SZT-R250 laser scanning system to collect point cloud data at a flight height of 120 m. During data preprocessing, we utilized LiDAR360 software for the initial processing of the point clouds, followed by meticulous manual classification and labeling via a moving profile buffer, resulting in a dataset containing 1 813 506 ground points and 1 933 801 NG points. To reduce the memory and computational resources required for each forward pass of the network and to adapt the network's receptive field for local feature learning and stability, we divided the entire terrain into 100 square blocks of 50 m each. Finally, using equations in Section III-D, we calculated the features necessary for the model's input dimensions

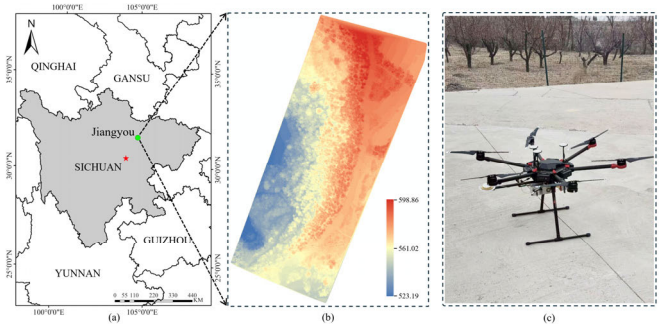

Fig. 4. Overview of the study area. (a) Geographical location. (b) DEM imagery. (c) ALS equipment.

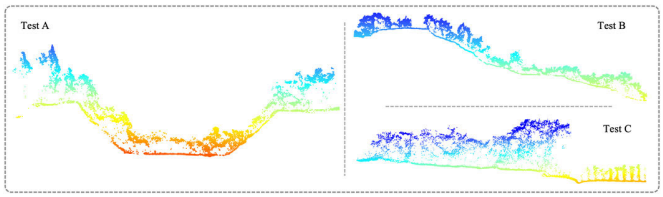

Fig. 5. Cross-sectional profile of Test data (buffer zone 0.9 m).

$(X, Y, Z, C_i, \delta_i, E_i)$ and divided the data into training, validation, and test sets in a 7:2:1 ratio.

Additionally, to analyze and evaluate the Point-SCT's performance in point cloud ground filtering in terrains with complex geometrical structures and high vegetation coverage, we selected representative areas within the study area for testing: a low-lying area with medium vegetation coverage (Test A), a foot slope (Test B), and a gradually sloping agroforestry area (Test C), as shown in Fig. 5. Test A, characterized by higher terrain surrounding a lower middle area that forms a small wetland with dense, low shrubs and trees and extensive undergrowth, exhibited uneven point density with 46 217 ground points and 53 798 NG points. Test B, a gentle slope transitioning to a steep incline and heavily vegetated with tall trees and low vegetation, included 185 914 ground points and 221 496 NG points. Test C, a gradual slope where farmland meets natural forest, covered extensively with crops and trees, consisted of 41 760 ground points and 52 565 NG points.

B. Implementation Details

1) Parameters: We implemented the Point-SCT model using the PyTorch DL framework and trained it on an NVIDIA RTX 3090 GPU for 300 epochs. The AdamW optimizer was selected for adjusting the model's weight parameters, with a weight decay of 0.01 and a momentum of 0.9 The initial learning rate was set to 0.001, and cosine annealing was employed to periodically adjust the learning rate. Given the large volume and uneven density of the outdoor point cloud data, the batch size and grid size were set to 2 and 0.02, respectively. To ensure that each spatial sampling window consistently contains enough points during the multilevel encoding stages while optimizing GPU memory usage and maximizing performance, the maximum number of input points was capped at 8000. Additionally, the k-value for nearest neighbor search was set to 32, the multilevel sampling rate was set to 0.25, and the spatial window size was set to 4. The depth of the encoding and sampling stages in this model framework was configured as [2], [4], and [2], and for a given number of input points N, the point feature vector sets generated at each encoding stage had cardinalities of $[N, 1/16N, 1/64N, 1/256N]$.

2) Comparative Study: In the current field of DL-based ALS point cloud processing, there is limited research focused on ground point filtering in mountainous areas with complex terrain structures and extensive vegetation coverage. To demonstrate the effectiveness of our method in such terrains, we selected several state-of-the-art approaches as comparison methods, including PointNet++, RandLA-Net, KPConv, Stratified Transformer, Point Transformer, and Point Transformer V3. We compared the performance of these models against our Point-SCT on the test dataset. Although these models have shown excellent performance on previous indoor and outdoor open-source point cloud dataset, their effectiveness in mountainous terrains has not yet been validated. Therefore, we followed the official guidelines for these methods during deployment and conducted fine-tuning as needed.

3) Metrics: To quantitatively assess the model performance, we employed standard metrics such as mean accuracy (mAcc), overall accuracy (OA), and intersection over union (IoU). Specifically, IoU_1 and IoU_2 represent the IoU for ground points and NG points, respectively. Additionally, we evaluated the model's efficiency using the size of network parameters and forward inference time. Furthermore, we performed qualitative analysis on three study areas with distinct characteristics, complementing the quantitative metrics.

C. Experimental Results

We comprehensively evaluated the performance of several mainstream point convolution-based models, point Transformer-based models, and Point-SCT in extracting ground points from the testing datasets A–C. We also quantified the inference results of each model on each testing set's point clouds. Specifically, we visualized the digital elevation model (DEM) of ground points from the testing set's ground truth as the base map and overlaid the misclassified ground and NG points to qualitatively analyze the interclass segmentation performance of each model in varying complex environments.

1) Analysis of Test A: Test A is a typical small low-lying area where the slope changes drastically at the junction of high and low lands, resulting in significant undulations in the point cloud data and highlighting the prominent geometric features of this area. In the quantitative results of competitive models in this region shown in Table I, the point Transformer-based methods achieve higher overall segmentation accuracy than convolution-based methods. Additionally, as shown in Fig. 6, Transformer-based methods exhibit relatively fewer misclassified points at the junctions between high and low lands, mainly due to the Transformer's advantages in long-distance feature construction and cross-domain feature association. Moreover, the lowland area is sparse in ground points due to water accumulation and extensive weed coverage, and the

TABLE I

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE TEST A DATASET (UNIT: %). PARAMS: PARAMETERS. FLOPS: FLOATING-POINT OPERATIONS. INFER TIME: INFERENCE TIME (S)

Method	Params (MB)	FLOPs (GB)	OA	mAcc	IoU_1	IoU_2	Infer Time (s)
CSF	—	—	77.64	76.13	—	—	—
PointNet++ [20]	6.52	11.74	85.52	82.06	81.86	83.17	45
RandLA-Net [21]	**2.67**	**5.12**	90.97	88.31	86.82	87.87	**32**
KPConv [18]	15.21	21.97	89.69	86.15	86.37	87.11	85
Point Transformer [47]	29.63	11.30	91.85	90.23	89.25	88.64	55
Point Transformer V3 [43]	36.96	28.79	93.83	91.27	91.33	86.79	98
Stratified Transformer [17]	14.67	16.16	93.97	93.92	91.54	91.25	46
Point-SCT (Ours)	28.19	24.27	**95.90**	**95.58**	**93.22**	**93.83**	67

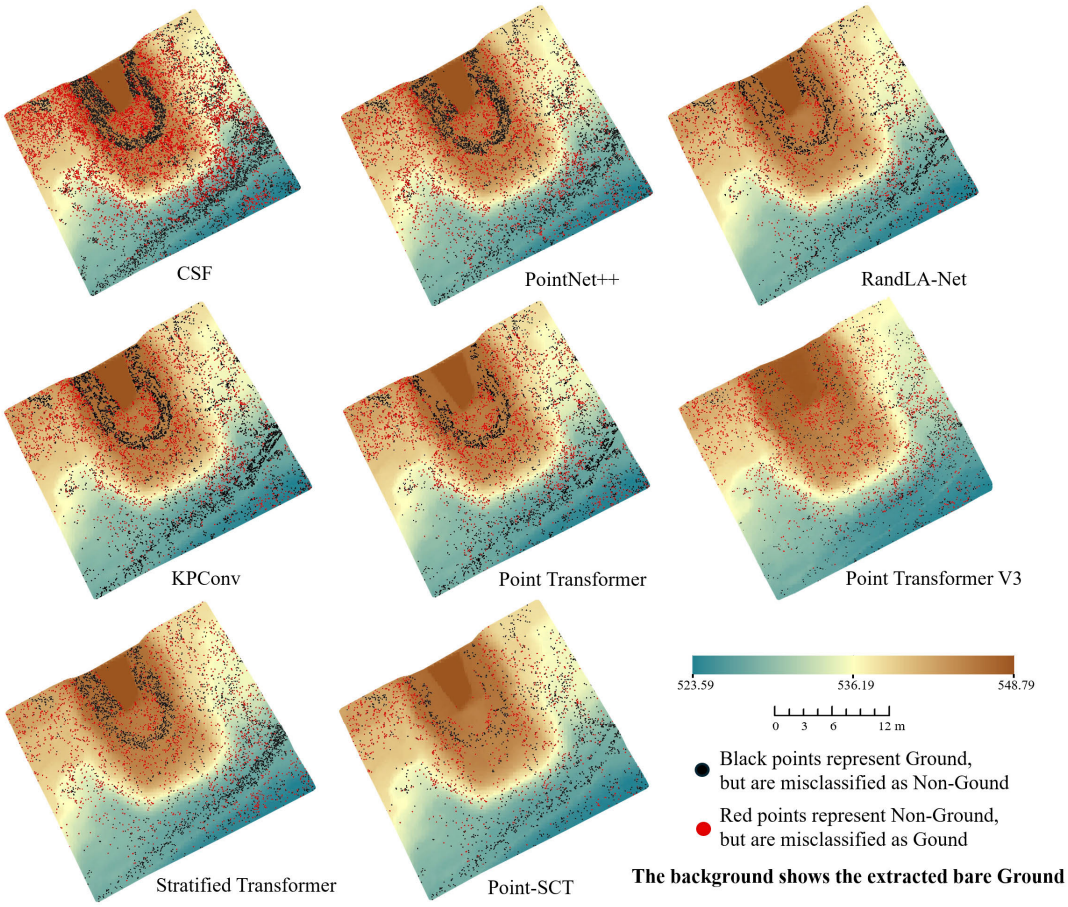

Fig. 6. Qualitative analysis of eight comparative methods on Test A (the background is DEM generated by bare ground points).

ground points bear similar characteristics to low vegetation, reducing the spatial feature differences between points and their neighbors.

However, RandLA-Net and KPConv exhibit relatively sparse distributions of misclassified points, demonstrating the convolution-based methods' strengths in capturing local details, though they are somewhat inadequate in handling long-distance features. Despite PointNet++'s lightweight network and low computational demands, its misclassification areas are relatively large and exhibit a certain regularity, mainly concentrated in the depression bottom where weeds are abundant and in areas with significant terrain changes. In contrast, Point-SCT enhances feature inputs by incorporating the original point cloud with the geometric structure of neighboring points and focuses on capturing local spatial geometric details and reconstructing long-distance features in multiple encoding stages. This enables Point-SCT to achieve the best performance in OA (95.90%), mAcc (95.58%), IoU1 (93.22%), and IoU2 (93.38%) metrics. However, Point-SCT still exhibits a relatively high number of misclassified points in waterlogged areas, mainly due to the low heterogeneity in geometric features between the ground points and the surrounding weeds in low-dimensional space, which causes the point cloud data to lack effective local contextual geometric information.

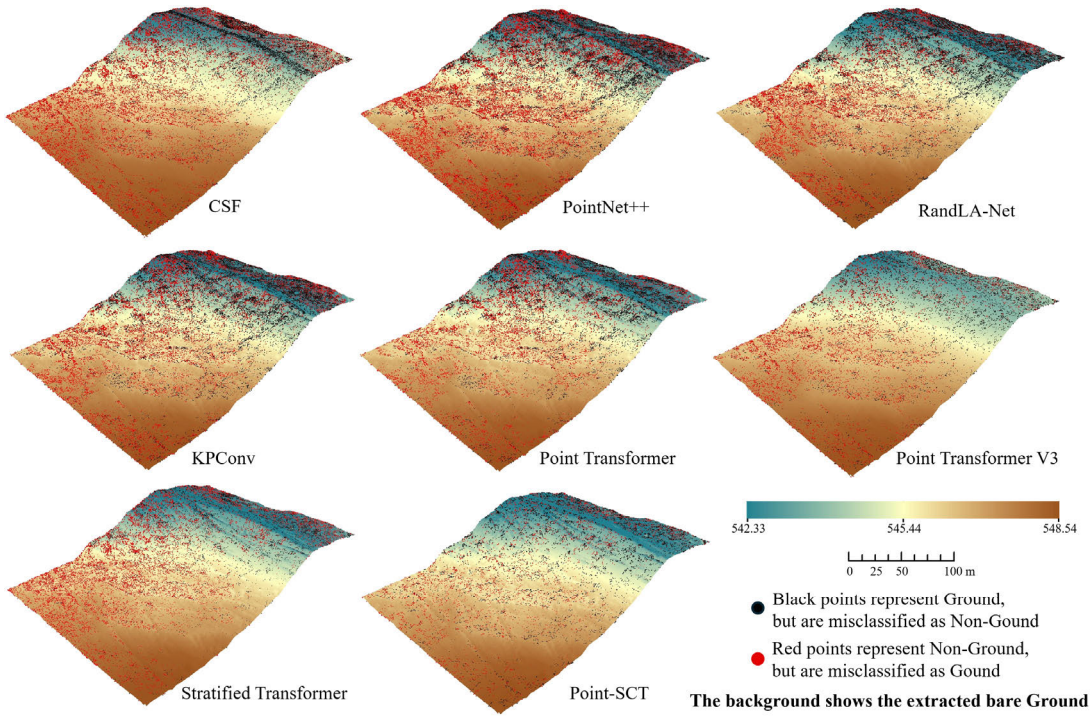

Fig. 7. Qualitative analysis of eight comparative methods on Test B (the background is DEM generated by bare ground points).

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE TEST B DATASET (UNIT: %). PARAMS: PARAMETERS. FLOPS: FLOATING-POINT OPERATIONS. INFER TIME: INFERENCE TIME (S)

Method	Params (MB)	FLOPs (GB)	OA	mAcc	IoU_1	IoU_2	Infer Time (s)
CSF	—	—	80.53	80.08	—	—	—
PointNet++ [20]	6.52	14.74	83.23	81.74	80.05	77.17	183
RandLA-Net [21]	**2.67**	**9.59**	88.97	88.12	85.23	84.82	**98**
KPConv [18]	15.21	37.70	89.35	88.28	86.65	87.93	297
Point Transformer [47]	29.63	24.29	90.01	89.83	86.25	88.12	169
Point Transformer V3 [46]	36.96	33.97	94.59	93.10	**94.09**	89.73	286
Stratified Transformer [17]	14.67	21.21	93.66	92.08	93.33	87.37	153
Point-SCT (Ours)	28.19	36.63	**95.79**	**94.32**	93.08	**94.20**	192

Despite Point-SCT's ability to reconstruct long-range features and focus on local geometric details across multiple encoding stages, in regions with small local feature differences, the model may still be limited by insufficient local information, resulting in misclassification.

Additionally, its superior performance comes at the cost of extensive network parameter storage and computational resources.

2) Analysis of Test B: The Test B scenario represents the foot slope. As shown in Fig. 7, the overall terrain exhibits a gradual incline; however, certain areas on the left experience sudden terrain subsidence, resulting in multiple step-like gaps in the point cloud data and leaving some regions blank. Sparse high trees are distributed on the gentle slope, while a long, narrow road traverses the higher terrain, surrounded by dense low vegetation. This environment makes the distinction between ground and NG points increasingly challenging, particularly in areas with significant slope changes and transitional terrain.

Table II presents the quantitative results of various methods on this terrain. The Transformer-based methods generally yield strong performance on evaluation metrics, with fewer misclassified points. Notably, Point Transformer V3, which integrates local attention modules based on space-filling curves and serialized encoding, considers the local relative positions, angles, and other geometric relationships between points and their neighborhoods. By incorporating spatial patch interaction techniques, it enhances long-range feature interactions, achieving optimal performance in ground point segmentation ($IoU_1 = 94.09\%$). However, the model performs suboptimal performance in segmenting NG points, particularly in cross-category boundary areas with dense roadside vegetation and steep terrain with significant changes, where many ground and NG points are misclassified. The former issue may result from system errors accumulated from multiple airborne scanning that cannot be manually removed, while the latter involves sparse ground point clouds in stepped areas.

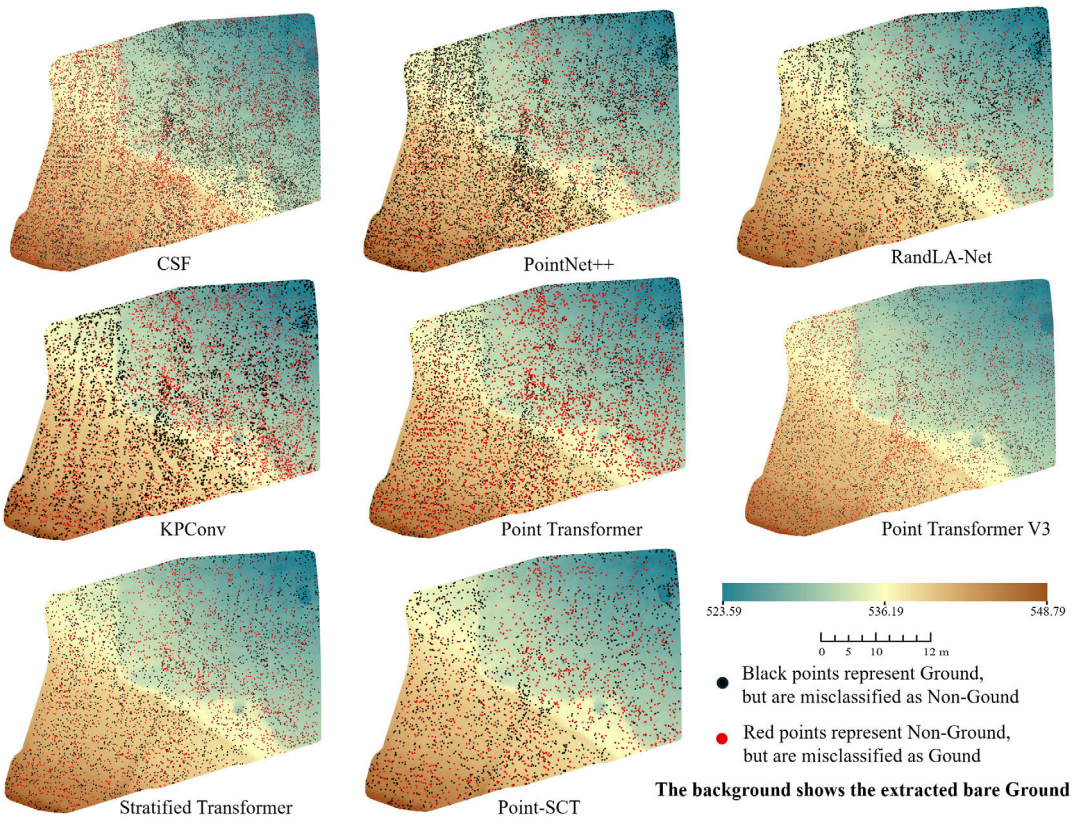

Fig. 8. Qualitative analysis of eight comparative methods on Test C (the background is DEM generated by bare ground points).

Both issues are compounded by the geometric feature similarities between NG points (e.g., grass or low vegetation) and ground points. Additionally, the local attention module constructs neighborhood relationships at a shallow level, directly passing the weighted aggregated attention feature to subsequent layers without filtering and readjusting low-weight feature vectors (e.g., local outliers or points with weak heterogeneity). This leads to ineffective amplification of key feature information or inaccurate aggregation. Other Transformer-based models also struggle with capturing small local geometric differences. In contrast, Point-SCT directly constructs features in the shallow layers of the network using complex neighborhood geometric relationships as guidance, aggregates and dilutes multiscale features through MLFA, and enables cross-window information exchange via SST. This approach effectively captures fine-grained local features in complex regions, achieving the best performance in NG point segmentation ($IoU_2 = 94.20\%$). Conversely, point convolution-based methods face significant challenges, particularly in vegetated stepped terrains, due to their limitations in capturing global features, resulting in widespread misclassification.

3) Analysis of Test C: The Test C scenario represents an agroforestry area, where the terrain is sharply divided by steep slopes, creating two distinct stepped surfaces. As shown in Fig. 8, the lower left side is an agricultural area, while the upper right side consists of woodland and low vegetation, contrasting with relatively sparse bare ground points.

According to the quantitative results in Table III, RandLA-Net demonstrates impressive performance despite its lightweight network and minimal computational requirements. In the agricultural area, the density distribution of misclassified points by RandLA-Net and KPConv exhibits a distinct pattern, primarily concentrated in the uneven ground between crop rows and at the ends of crop leaves. This highlights the vulnerability of point convolution methods, which rely on local neighborhood features, to performance degradation in regions with extreme density variations. In contrast, Transformer-based methods show a more sparse and evenly distributed pattern of global misclassification points. However, both Point Transformer V3 and Point Transformer struggle in areas with extensive mixed high and low vegetation coverage, leading to a significant number of misclassified NG points, particularly in regions with marked slope changes. Point-SCT integrates the advantages of point convolution in capturing local spatial details with the Transformer's ability to reconstruct and associate global long-range features, achieving leading results across various accuracy metrics. Particularly in areas with significant slope changes, Point-SCT exhibits sparse misclassification points, showing excellent generalization capabilities. However, this comes with the tradeoff of higher network parameter storage (28.19 MB) and computational resource usage (26.42 GB).

D. Ablation Study

1) Ablation of Features Input: To explore the reliability and effectiveness of replacing the RGB input with the proposed geometric augmented feature combination (i.e., BCE) in the Point-SCT, we conducted performance validation on the

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE TEST C DATASET (UNIT: %). PARAMS: PARAMETERS. FLOPS: FLOATING-POINT OPERATIONS. INFER TIME: INFERENCE TIME

Method	Params (MB)	FLOPs (GB)	OA	mAcc	IoU_1	IoU_2	Infer Time (s)
CSF	—	—	74.51	73.91	—	—	—
PointNet++ [20]	6.52	9.25	82.31	80.09	79.93	80.76	45
RandLA-Net [21]	**2.67**	**5.92**	93.60	91.07	88.83	90.13	**32**
KPConv [18]	15.21	19.73	89.38	89.19	88.12	87.80	83
Point Transformer [47]	29.63	17.87	89.97	88.02	87.07	87.95	58
Point Transformer V3 [46]	36.96	29.05	91.95	91.86	89.43	91.60	97
Stratified Transformer [17]	14.67	18.00	93.83	91.86	90.19	90.18	43
Point-SCT (Ours)	28.19	26.42	**96.25**	**95.13**	**93.07**	**95.79**	65

TABLE IV
PERFORMANCE OF DIFFERENT FEATURE COMBINATIONS AS INPUTS TO POINT-SCT ON THE TEST A DATASET

Input Features	OA	mAcc	mIoU
XYZ	90.21	90.17	88.27
XYZ+RGB	94.73	94.00	92.09
XYZ+BCE	**95.90**	**95.58**	**93.53**

Note: $mIoU = (IoU_1 + IoU_2)/2$

Test A, characterized by relatively sparse point clouds but with significant and complex terrain variations. We evaluated the performance of Point-SCT using three different input feature combinations: 1) spatial coordinates *XYZ* only; 2) *XYZ* + RGB; and 3) *XYZ* + BCE. The experimental results are presented in Table IV.

Table IV indicates that the proposed BCE, which is based on neighborhood geometric spatial features, significantly improves the performance of point cloud ground filtering in complex terrains compared to using RGB as the feature input. This enhancement is attributed to the fact that these features capture the relative geometric relationships and physical information of points within their spatial neighborhood, directly improving the efficiency of feature learning and the model's understanding of spatial variations. Specifically, the boundary detector aids the model in identifying fault lines or edges with steep slope changes. Meanwhile, surface curvature provides crucial information about surface deformation at shallow network layers, and average elevation effectively assists the model in learning terrain features based on elevation changes.

2) Ablation of Network and Sampling Strategy: To further investigate the impact of LGSE and DP modules in the multiscale feature aggregation module on overall performance, we dynamically froze each module in the Point-SCT network. Given the MLP's similarity to arbitrary continuous functions, we replaced the LGSE and DP modules with MLPs and conventional MP and combined them with SST. Additionally, we conducted cross-sectional visualizations of the results for each module combination on Test A to qualitatively evaluate their semantic perception capabilities when capturing the complex and intricate topological structures characteristic of mountainous terrains. Finally, we selected two representative heuristic sampling strategies, FPS and inverse density importance sampling (IDIS) and compared them with the RS strategy employed in the model. We validated the effect of these module combinations and sampling strategies on the overall performance of the model using the Test A dataset.

Table V and Fig. 9. show that after freezing the LGSE and DP module layers, the performance improvement is not significant when compared to the KNN + SST configuration after introducing MLP and MP. Specifically, in areas with abundant low-lying vegetation on flat terrain, the ground point segmentation performance is not notably improved, especially in mixed high and low vegetation regions, where it is difficult to extract ground points. However, with the addition of the DP module, the mIoU improves by 2.85%, highlighting the DP module's effectiveness in selectively retaining significant spatial feature relationships. In regions with mixed high and low vegetation, ground point extraction is notably enhanced. After introducing the LGSE module, Point-SCT becomes more sensitive to intricate and geometrically similar interclass perceptions, resulting in a significant improvement in ground point segmentation performance in complex vegetation areas. While the combination of LGSE and MP is more lightweight in terms of network parameters and computational resources, the combination of LGSE and DP achieves the best results in terms of accuracy. This underscores the importance of high-density spatial geometric feature vector capture and selective dilution strategies in complex terrain scenarios.

Table VI shows that under the experimental setup with a single GPU and the Test A dataset (around 105 points), both FPS and IDIS sampling strategies introduce no additional parameters. However, due to their higher time complexities $O(N^2)$ and $O(N \log N)$, the computational overhead (in FLOPs) increases by approximately 0.05 and 0.01 GB compared to the RS strategy. Consequently, the number of points processed per second decreases significantly from 10^8 (RS) to approximately 10^4 (FPS) and 10^5 (IDIS). Thus, in large-scale training and deployment scenarios, it is essential to balance the choice of sampling strategy with both model accuracy and computational efficiency to avoid unnecessary computational overhead.

TABLE V
PERFORMANCE OF KEY MODULE COMBINATIONS IN THE POINT-SCT NETWORK ON THE TEST A DATASET

Modules	Params (MB)	FLOPs (GB)	OA	mAcc	mIoU
KNN+SST	16.33	17.15	93.17	92.60	90.12
KNN+MLP+MP+SST	19.60	20.04	93.23	92.89	91.26
KNN+MLP+**DP**+SST	32.46	27.87	94.21	93.38	92.97
LGSE+MP+SST	16.33	18.92	94.73	93.73	92.49
LGSE+**DP**+SST	28.19	24.27	**95.90**	**95.58**	**93.53**

Note: mIoU = $(IoU_1 + IoU_2)/2$

TABLE VI
PERFORMANCE OF DIFFERENT SAMPLING STRATEGIES IN THE POINT-SCT NETWORK ON THE TEST A DATASET. TC: TIME COMPLEXITY

Sampling Strategy	Params (MB)	FLOPs (GB)	OA	mAcc	mIoU	TC	Frame Per Sec.
RS	28.19	24.27	95.90	95.58	93.53	$\boldsymbol{O(1)}$	$\approx \mathbf{10^8}$
RS → IDIS	28.19	24.30	95.81	95.52	93.51	$O(N \log N)$	$\approx 10^5$
RS → FPS	28.19	24.32	95.95	95.61	93.53	$O(N^2)$	$\approx 10^4$

Note: mIoU = $(IoU_1 + IoU_2)/2$

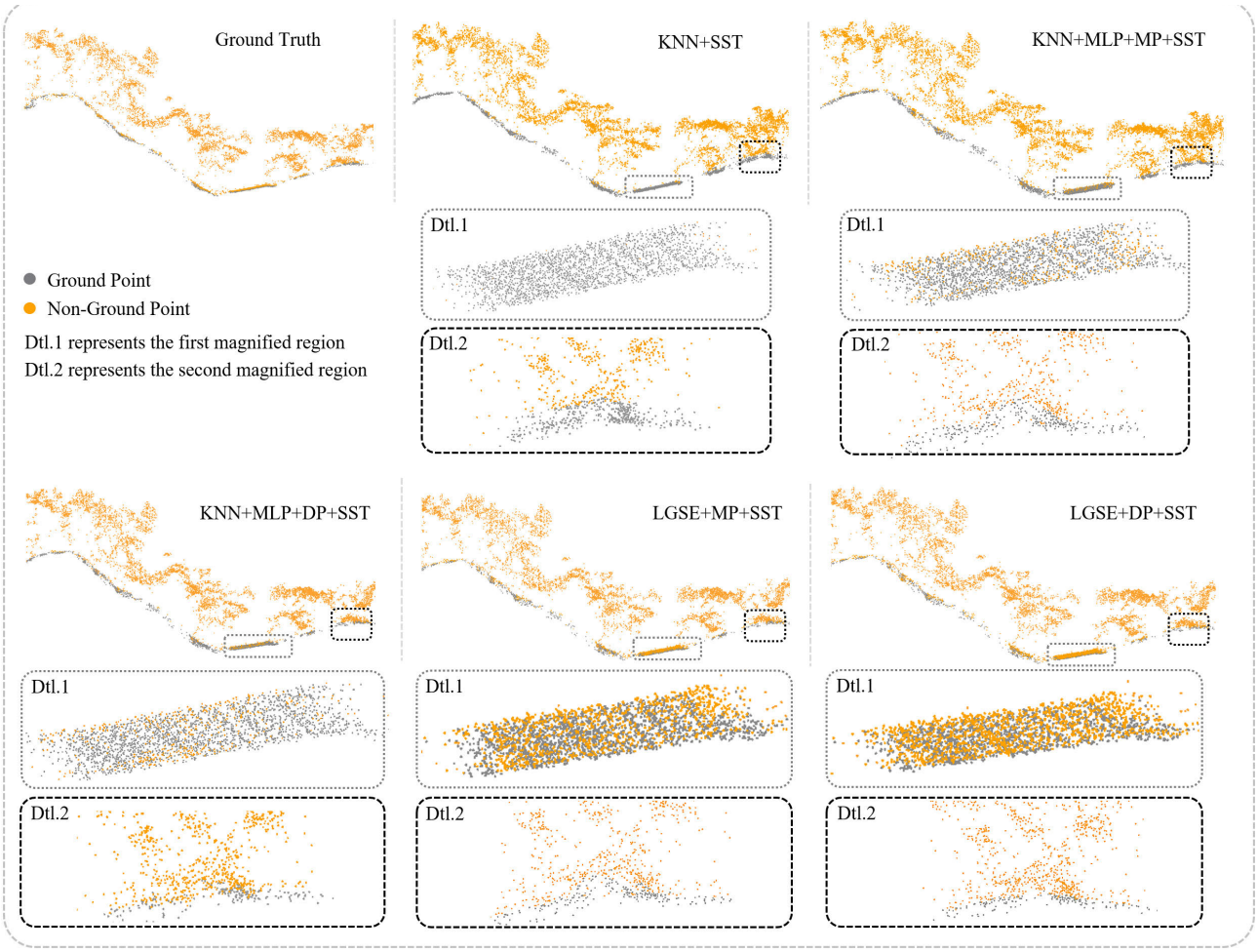

Fig. 9. Visualization of geometric detail capture performance of different module combinations in Point-SCT on Test A.

3) Ablation of K-Value and Spatial Window Size: Following the verification of the reliability and effectiveness of our proposed modules, we explored the impact of k-values and window sizes on the performance of Point-SCT due to the model's implementation of the KNN and the introduction of spatial shifted windows in SST. These adaptations primarily concern the model's learnable range of point neighborhoods.

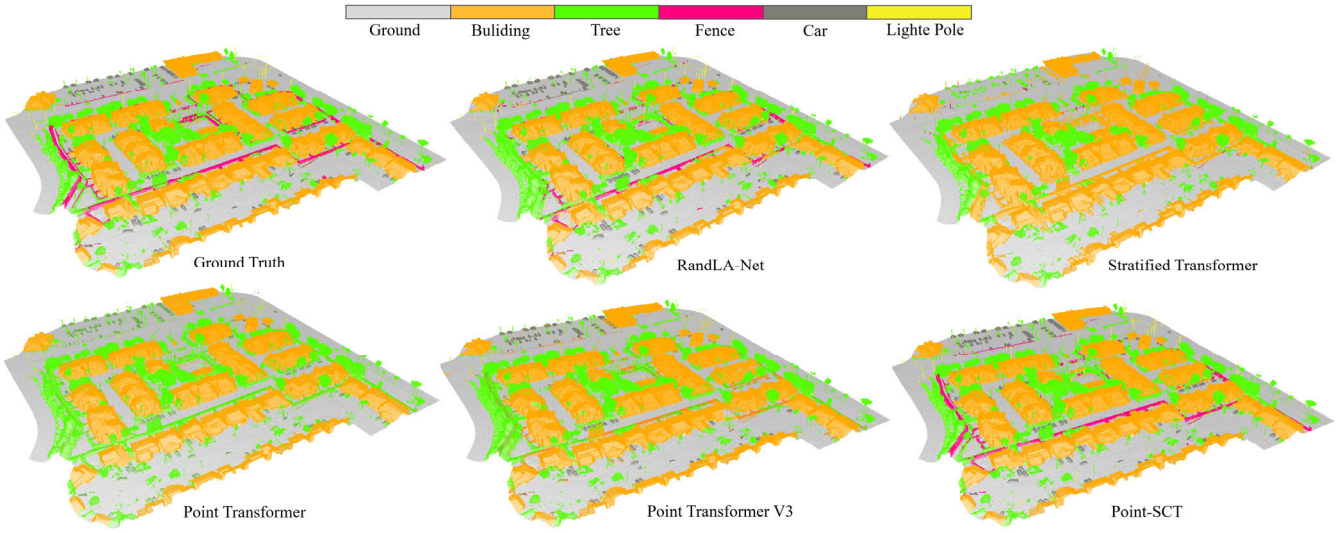

Fig. 10. Qualitative analysis of five comparative methods on the STPLs3D dataset.

TABLE VII
PERFORMANCE OF DIFFERENT K-VALUES ON POINT-SCT ON THE TEST A DATASET

k-value	OA	mAcc	mIoU
4	91.85	91.12	89.24
8	94.79	94.03	92.55
16	95.90	95.58	93.53
32	95.82	95.47	93.36

Note: mIoU = (IoU_1 + IoU_2)/2

TABLE VIII
PERFORMANCE OF DIFFERENT WINDOW SIZES ON POINT-SCT ON THE TEST A DATASET

Win Size	OA	mAcc	mIoU
2	92.13	91.81	91.09
4	95.90	95.58	93.53
8	95.77	95.13	92.86
16	N/A	N/A	N/A

Note: mIoU = (IoU1 + IoU2)/2

In Table VII, k-values of 4, 8, 16, and 32 were employed for BCE features and nearest neighboring points searches within the model. The results indicate stable performance at k-values of 8, 16, and 32, demonstrating robustness to variations within a certain neighborhood range. A decrease in performance metrics at a k-value of 4 suggests an inadequate expression of geometric features due to a smaller neighborhood search range.

In Table VIII, window sizes (Win Size) of 2, 4, 8, and 16 were adjusted to explore spatial relationships within Point-SCT. Performance metrics were similar for window sizes of 4 and 8, with optimal results at a size of 4. This indicates that appropriate adjustments in window size are sufficient for the model to construct and generalize more complex spatial relationships. However, a setting of 16 exceeded the memory capacity available on our current hardware, and thus, results could not be provided. These ablation studies were conducted solely on the Test A dataset to guide parameter settings for unknown datasets.

E. Generalization Study

To evaluate the generalization ability of Point-SCT on the cross-scenario dataset, we conducted generalization experiments on the STPLs3D dataset [60]. STPLs3D is a high-density ALS point cloud dataset consisting of four real-world large-scale outdoor scenes: USC, WMSC, OCCC, and RA. In the experimental benchmark setup, we selected RandLA-Net as the representative point convolution-based method for its superior performance, as well as three Transformer-based methods, with the RA scene as the test set.

As shown in Table IX and Fig. 10, the segmentation results for the Fence and Car classes are suboptimal across all models, primarily due to the relatively low point density for these classes compared to others. Additionally, the Fence class is extensively intertwined with the Tree class, and the Car class is sparsely distributed, further complicating segmentation. However, Point Transformer V3, which integrates neighborhood-relative geometric relationships and long-range semantic associations, achieved superior results in the Car category, with an accuracy of 79.17%. Moreover, the geometric shapes of Light Poles and Trees are highly similar, leading to segmentation errors in the pole areas. In contrast, the point convolution-based RandLA-Net yielded more stable results for these classes, demonstrating the convolution's robust ability to capture local geometric semantics. Point-SCT, which combines convolution-based local geometric feature aggregation with global long-range relational modeling, achieved well-balanced and excellent overall segmentation results, with an mIoU of 81.56% and OA of 96.89%. This performance underscores its exceptional generalization capability.

TABLE IX
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE STPLS3D DATASET (UNIT: %)

Method	Per Class IoU						mIoU	OA
	Ground	Building	Tree	Fence	Car	Light Pole		
RandLA-Net [21]	89.46	88.26	79.23	25.58	71.81	58.52	68.81	90.68
Point Transformer [47]	90.03	87.26	81.36	9.54	38.28	48.36	58.92	87.26
Point Transformer V3 [46]	92.39	90.21	85.27	19.95	79.17	55.34	70.39	92.23
Stratified Transformer [17]	90.76	87.52	81.53	12.65	44.39	49.86	61.12	88.37
Point-SCT (Ours)	94.68	93.35	84.71	56.83	80.91	78.83	**81.56**	**96.89**

V. CONCLUSION

In this article, we propose Point-SCT, a novel point cloud ground filtering model that synergistically combines point convolutions and spatial shifted window Transformer within a multiscale hierarchical architecture. Utilizing an encoder–decoder structure as its core architecture, Point-SCT integrates the powerful local spatial feature perception capabilities of convolution with the global context modeling abilities of Transformers. This dual approach renders Point-SCT exceptionally effective in handling ground filtering tasks across large-scale remote sensing scenarios characterized by complex terrain variations, geological structures, and vegetation environments. In local spatial feature learning, we have developed a convolution-based MLFA module, consisting of the LGSE and DP strategies. For global feature learning, we employ the spatial Swin Transformer to establish long-range associations and enhance the high-weight feature vectors after sampling. Additionally, we have innovated a shallow enhancement feature combination, BCE, based on the spatial relationship of neighborhoods. Finally, Point-SCT was applied to ground point filtering tasks within complex mountainous terrain scenarios, with comprehensive quantitative and qualitative analyses undertaken. Furthermore, when benchmarked against contemporary state-of-the-art models across Test A–C datasets, Point-SCT achieved top-tier OA scores of 95.9%, 95.79%, and 95.90%, demonstrating its superior segmentation accuracy and underscoring its robustness in complex remote sensing scenarios.

Future Work: We will further explore pruning, quantization, and distillation techniques for the Point-SCT to balance computational complexity with performance, thereby facilitating the lightweight deployment of the model. Simultaneously, we will apply the Point-SCT to a broader range of cross-scenario datasets to continually enhance its generalization capability. Additionally, in the visualization of results for complex mountainous point cloud filtering tasks, we aim to improve and refine the representation of 3-D effects in a flat and intuitive manner, which will be a key focus of our future research.

REFERENCES

[1] F. Tang and Q. Yang, "Progress and prospect of multi-source remote sensing monitoring technology for coal mining subsidence in western Loess Plateau mining area," *Coal Sci. Technol.*, vol. 51, no. 12, pp. 1–18, 2023.

[2] F. Tang, J. Lu, S. Wei, X. Li, K. He, and Q. Yang, "Improvement of mining subsidence modeling method based on UAV LiDAR in Yushen mining area," *J. Coal Sci.*, vol. 45, no. 7, pp. 2655–2666, 2020, doi: 10.13225/j.cnki.jccs.2020.0466.

[3] H. Zhu, Y. Liu, and H. Qin, "Rapid extraction of potential landslide variables using UAV-borne LiDAR point cloud," *Prot. Eng.*, vol. 45, no. 5, pp. 61–66, 2023.

[4] M. Bremer and O. Sass, "Combining airborne and terrestrial laser scanning for quantifying erosion and deposition by a debris flow event," *Geomorphology*, vol. 138, no. 1, pp. 49–60, Feb. 2012, doi: 10.1016/j.geomorph.2011.08.024.

[5] H. Ma, W. Zhou, and L. Zhang, "DEM refinement by low vegetation removal based on the combination of full waveform data and progressive TIN densification," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 260–271, Dec. 2018, doi: 10.1016/j.isprsjprs.2018.09.009.

[6] G. Vosselman, "Slope based filtering of laser altimetry data," in *Proc. IAPRS*, Jan. 2000, pp. 935–942.

[7] K. Kraus and N. Pfeifer, "Advanced DTM generation from LiDAR data," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. 34, pp. 23–30, Oct. 2011.

[8] T. J. Pingel, K. C. Clarke, and W. A. McBride, "An improved simple morphological filter for the terrain classification of airborne LiDAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 77, pp. 21–30, Mar. 2013, doi: 10.1016/j.isprsjprs.2012.12.002.

[9] A. Rizaldy, C. Persello, C. Gevaert, S. Elberink, and G. Vosselman, "Ground and multi-class classification of airborne laser scanner point clouds using fully convolutional networks," *Remote Sens.*, vol. 10, no. 11, p. 1723, Oct. 2018, doi: 10.3390/rs10111723.

[10] Y. Cai, L. Fan, P. M. Atkinson, and C. Zhang, "Semantic segmentation of terrestrial laser scanning point clouds using locally enhanced image-based geometric representations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, doi: 10.1109/TGRS.2022.3161982.

[11] Y. Chen et al., "EPDet: Enhancing point clouds features with effective representation for 3D object detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 127, Mar. 2024, Art. no. 103688, doi: 10.1016/j.jag.2024.103688.

[12] Y. Chen, R. Wu, C. Yang, and Y. Lin, "Urban vegetation segmentation using terrestrial LiDAR point clouds based on point non-local means network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102580, doi: 10.1016/j.jag.2021.102580.

[13] A. V. Phan, M. L. Nguyen, Y. L. H. Nguyen, and L. T. Bui, "DGCNN: A convolutional neural network over large-scale labeled graphs," *Neural Netw.*, vol. 108, pp. 533–543, Dec. 2018, doi: 10.1016/j.neunet.2018.09.001.

[14] H. Xiu, T. Shinohara, and M. Matsuoka, "Dynamic-scale graph convolutional network for semantic segmentation of 3D point cloud," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 271–2717, doi: 10.1109/ISM46123.2019.00062.

[15] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 922–928, doi: 10.1109/IROS.2015.7353481.

[16] T. Jiang, J. Sun, S. Liu, X. Zhang, Q. Wu, and Y. Wang, "Hierarchical semantic segmentation of urban scene point clouds via group proposal and graph attention network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102626, doi: 10.1016/j.jag.2021.102626.

[17] X. Lai et al., "Stratified transformer for 3D point cloud segmentation," 2022, *arXiv:2203.14508*.

[18] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "KPConv: Flexible and deformable convolution for point clouds," 2019, *arXiv:1904.08889*.

[19] 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. 77–85, doi: 10.1109/CVPR.2017.16.

[20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," 2017, *arXiv:1706.02413*.
[21] Q. Hu et al., "RandLA-net: Efficient semantic segmentation of large-scale point clouds," 2019, *arXiv:1911.11236*.
[22] S. Jin, Y. Su, X. Zhao, T. Hu, and Q. Guo, "A point-based fully convolutional neural network for airborne LiDAR ground point filtering in forested environments," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3958–3974, 2020, doi: 10.1109/JSTARS.2020.3008477.
[23] A. Nurunnabi, F. N. Teferle, J. Li, R. C. Lindenbergh, and S. Parvaz, "Investigation of pointnet for semantic segmentation of large-scale outdoor point clouds," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLVI-4, pp. 397–404, Dec. 2021, doi: 10.5194/isprs-archives-xlvi-4-w5-2021-397-2021.
[24] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.
[25] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[26] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[28] H. Wu et al., "CvT: Introducing convolutions to vision transformers," 2021, *arXiv:2103.15808*.
[29] N. Qin, W. Tan, L. Ma, D. Zhang, H. Guan, and J. Li, "Deep learning for filtering the ground from ALS point clouds: A dataset, evaluations and issues," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 246–261, Aug. 2023, doi: 10.1016/j.isprsjprs.2023.06.005.
[30] Y. Chen, S. Zhang, T. Han, Y. Du, W. Zhang, and J. Li, "Chat3D: Interactive understanding 3D scene-level point clouds by chatting with foundation model for urban ecological construction," *ISPRS J. Photogramm. Remote Sens.*, vol. 212, pp. 181–192, Jun. 2024, doi: 10.1016/j.isprsjprs.2024.04.024.
[31] G. Riegler, A. O. Ulusoy, and A. Geiger, "OctNet: Learning deep 3D representations at high resolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6620–6629.
[32] B. Graham, M. Engelcke, and L. v. d. Maaten, "3D semantic segmentation with submanifold sparse convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9224–9232.
[33] Y. Li, G. Tong, X. Li, L. Zhang, and H. Peng, "MVF-CNN: Fusion of multilevel features for large-scale point cloud classification," *IEEE Access*, vol. 7, pp. 46522–46537, 2019, doi: 10.1109/ACCESS.2019.2908983.
[34] F. J. Lawin, M. Danelljan, P. Tosteberg, G. Bhat, F. S. Khan, and M. Felsberg, "Deep projective 3D semantic segmentation," in *Computer Analysis of Images and Patterns*. Cham, Switzerland: Springer, 2017, pp. 95–107, doi: 10.1007/978-3-319-64689-3_8.
[35] Y. Zhu, Y. Long, Y. Guan, S. Newsam, and L. Shao, "Towards universal representation for unseen action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9436–9445, doi: 10.1109/CVPR.2018.00983.
[36] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on $\mathcal{X}$-transformed points," 2018, *arXiv:1801.07791*.
[37] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11057–11066, doi: 10.1109/CVPR.2019.01132.
[38] H. Cho and I. S. Choi, "Enhanced deep-learning prediction of molecular properties via augmentation of bond topology," *ChemMedChem*, vol. 14, no. 17, pp. 1604–1609, Sep. 2019, doi: 10.1002/cmdc.201900458.
[39] D. Lu, J. Zhou, K. Gao, J. Du, L. Xu, and J. Li, "Dynamic clustering transformer network for point cloud segmentation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 128, Apr. 2024, Art. no. 103791, doi: 10.1016/j.jag.2024.103791.
[40] J. Sun, C. Qing, J. Tan, and X. Xu, "Superpoint transformer for 3D scene instance segmentation," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 2393–2401, doi: 10.1609/aaai.v37i2.25335.
[41] L. Hui, H. Yang, M. Cheng, J. Xie, and J. Yang, "Pyramid point cloud transformer for large-scale place recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6078–6087, doi: 10.1109/ICCV48922.2021.00604.
[42] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Comput. Vis. Media*, vol. 7, no. 2, pp. 187–199, Jun. 2021, doi: 10.1007/s41095-021-0229-5.
[43] 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. 17149–17158.
[44] Q. Zhong and X.-F. Han, "Point cloud learning with transformer," 2021, *arXiv:2104.13636*.
[45] X. Wu, Y. Lao, L. Jiang, X. Liu, and H. Zhao, "Point transformer v2: Grouped vector attention and partition-based pooling," 2022, *arXiv:2210.05666*.
[46] X. Wu et al., "Point transformer v3: Simpler, faster, stronger," 2023, *arXiv:2312.10035*.
[47] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," 2020, *arXiv:2012.09164*.
[48] D. Lu, Q. Xie, K. Gao, L. Xu, and J. Li, "3DCTN: 3D convolution-transformer network for point cloud classification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24854–24865, Dec. 2022, doi: 10.1109/TITS.2022.3198836.
[49] T. Yotsumata, M. Sakamoto, and T. Satoh, "Quality improvement for airborne LiDAR data filtering based on deep learning method," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLIII-B2, pp. 355–360, Aug. 2020, doi: 10.5194/isprs-archives-xliii-b2-2020-355-2020.
[50] X. Hu and Y. Yuan, "Deep-learning-based classification for DTM extraction from ALS point cloud," *Remote Sens.*, vol. 8, no. 9, p. 730, Sep. 2016, doi: 10.3390/rs8090730.
[51] E. Janssens-Coron and E. Guilbert, "Ground point filtering from airborne LiDAR point clouds using deep learning: A preliminary study," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLII-2, pp. 1559–1565, Jun. 2019, doi: 10.5194/isprs-archives-xlii-2-w13-1559-2019.
[52] S. Krisanski, M. S. Taskhiri, S. G. Aracil, D. Herries, and P. Turner, "Sensor agnostic semantic segmentation of structurally diverse and complex forest point clouds using deep learning," *Remote Sens.*, vol. 13, no. 8, p. 1413, Apr. 2021, doi: 10.3390/rs13081413.
[53] M. Kada and D. Kuramin, "ALS point cloud classification using PointNet++ and KPConv with prior knowledge," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLVI-4/W4-2021, pp. 91–96, Oct. 2021, doi: 10.5194/isprs-archives-xlvi-4-w4-2021-91-2021.
[54] P. Guerrero, Y. Kleiman, M. Ovsjanikov, and N. J. Mitra, "PCPNet learning local shape properties from raw point clouds," *Comput. Graph. Forum*, vol. 37, no. 2, pp. 75–85, May 2018, doi: 10.1111/cgf.13343.
[55] F. Hao, J. Li, R. Song, Y. Li, and K. Cao, "Mixed feature prediction on boundary learning for point cloud semantic segmentation," *Remote Sens.*, vol. 14, no. 19, p. 4757, Sep. 2022, doi: 10.3390/rs14194757.
[56] D. Lu, K. Gao, Q. Xie, L. Xu, and J. Li, "3DGTN: 3-D dual-attention GLocal transformer network for point cloud classification and segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.
[57] Z. Zhang, J. Ding, L. Jiang, D. Dai, and G. Xia, "FreePoint: Unsupervised point cloud instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 28254–28263.
[58] B. Yang and Z. Dong, "A shape-based segmentation method for mobile laser scanning point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 81, pp. 19–30, Jul. 2013, doi: 10.1016/j.isprsjprs.2013.04.002.
[59] S. Huang, Q. Hu, P. Zhao, J. Li, M. Ai, and S. Wang, "ALS point cloud semantic segmentation based on graph convolution and transformer with elevation attention," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2877–2889, 2024, doi: 10.1109/JSTARS.2023.3347224.
[60] M. Chen et al., "STPLS3D: A large-scale synthetic and real aerial photogrammetry 3D point cloud dataset," 2022, *arXiv:2203.09065*.

Jingxiang Li received the B.S. degree in computer science and information, and the M.Sc. degree from Xi'an University of Science and Technology, Xi'an, Shanxi, China, in 2025.

In 2024, he went to the University of Sains Malaysia, Penang, Malaysia, as a joint-trained master's student for six months. He has published papers in IEEE. His research interests include deep learning, computer vision, and 3-D point cloud processing.

Fuquan Tang received the Ph.D. degree in engineering from Xi'an University of Science and Technology, Xi'an, China, in 2009.

He is currently a Professor and a Ph.D. Supervisor with the School of Geomatics, Xi'an University of Science and Technology, where he also leads the Doctoral Program in Surveying and Mapping Science and Technology. He serves on several academic and professional committees, including the Mine Surveying and Education Committees of the Chinese Society for Geodesy, Photogrammetry and Cartography, the Mine Surveying Committee of the China Coal Society, and the Engineering Accreditation Expert Group of the Ministry of Education. He is also the Director of the Underground Engineering and Mine Surveying Committee of Shaanxi Province. He has led six provincial- and ministerial-level research projects, including those funded by the National Natural Science Foundation of China, and has received four awards for scientific and technological progress. He has published over 60 academic papers, authored five monographs and textbooks, and held 11 patents.

Lingfei Ma (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in geomatics engineering from the University of Waterloo, Waterloo, ON, Canada, in 2015, 2017, and 2020, respectively.

He is currently a Research Professor with the School of Geospatial Artificial Intelligence, East China Normal University, Shanghai, China. His research interests include autonomous driving, mobile laser scanning, intelligent processing of point clouds, 3-D scene modeling, and machine learning.

Dr. Ma was a recipient of the 2020 National Best Ph.D. Thesis Award granted by Canadian Remote Sensing Society. He served as a Guest Editor for *International Journal of Applied Earth Observation and Geoinformation*.

Chao Zhu is currently pursuing the Ph.D. degree with Xi'an University of Science and Technology, Xi'an, China.

He has published five papers in peer-reviewed journals. His research interests include airborne LiDAR surveying, deep learning, and computer vision.

Zhen Gong received the Ph.D. degree in communication and information systems from Xiamen University, Xiamen, China, in 2019.

He was a Visiting Ph.D. Scholar at the University of Waterloo, Waterloo, ON, Canada. He is currently a Research Fellow and a Lecturer with the College of Computer Engineering, Jimei University, Xiamen. His main research interests include computer vision and 3-D perception for robotics, covering areas such as multisensor state estimation and optimization, point cloud semantic analysis, online calibration of multisensor fusion, and integrated perception control networks based on reinforcement learning. He is the Principal Investigator of the National Natural Science Foundation of China (NSFC) Youth Project (2023), Fujian Provincial General Grant, and Xiamen Municipal Youth Grant.

Nur Intan Raihana Ruhaiyem received the Ph.D. degree specializing in computational cell biology from The University of Queensland, Brisbane, QLD, Australia, in 2014.

She is currently a Senior Lecturer with the School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia. Her research interests include the areas of medical image processing and analysis, computer vision, and data visualization and analysis.

Jonathan Li (Fellow, IEEE) received the Ph.D. degree in geomatics engineering from the University of Cape Town, Cape Town, South Africa, in 2000.

He is currently a Professor of geomatics and systems design engineering with the University of Waterloo, Waterloo, ON, Canada. He has published papers in flagship conferences in computer vision and AI, including CVPR, AAAI, and IJCAI. He has supervised nearly 200 master's, Ph.D. students, postdoctoral fellows, and visiting scholars to completion. He has co-authored almost 600 publications, more than 150 of which were published in top remote sensing journals, including *Remote Sensing of Environment, ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and *International Journal of Applied Earth Observation and Geoinformation.* His research interests include AI-based information extraction from Earth observation images and LiDAR point clouds, pointgrammetry and remote sensing, GeoAI, and 3-D vision for digital twin cities and autonomous driving.

Dr. Li is a fellow of Canadian Academy of Engineering, the Royal Society of Canada (Academy of Science), and the Engineering Institute of Canada. He is the President of Canadian Institute of Geomatics (CIG). He is the Editor-in-Chief of *International Journal of Applied Earth Observation and Geoinformation* and an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.