Grid Point Serialized Transformer for LiDAR Point Cloud Semantic Segmentation in Various Densities and Heights Scenes

Huchen Li, Wubiao Huang, Jiacheng Liu, Ke Chen, and Fei Deng

***Abstract*—Point cloud semantic segmentation is among the important tasks to achieve comprehensive perception of 3-D environments. However, current segmentation methods suffer from limited local receptive fields, poor extraction of global information, and insufficient scene generalizability. To alleviate these problems, we propose the grid point serialized transformer (GridPSFormer), a semantic segmentation method based on space-filling curves (SFCs). GridPSFormer maps gridded points to 1-D serialized orders via the 3-D spatial serialized order (SSO) module, which shows superior locality-preserving and provides a comprehensive understanding of the 3-D space. Then, by combining the local serialized attention (LSA) mechanism and the global serialized mamba (GSM) module, GridPSFormer effectively captures the local and global features of serialized orders and improves the modeling ability of points at long distances. In addition, the classes-refined serialization (CRS) module complements the semantic context information to enhance the generalizability in various densities and heights scenes. GridPSFormer achieved SOTA performance on three datasets, HRHD_HK, WHU_ALS, and DALES, with 64.89%, 68.17%, and 85.50% of mean Intersection over Union (mIoU) as well as 91.41%, 82.99%, and 98.24% of overall accuracy (OA), respectively. Experimental results demonstrated that the various SSOs can efficiently explore 3-D spatial information and achieve accurate semantic segmentation of large-scale scenes with lower computational complexity.**

***Index Terms*—3-D spatial serialized order (SSO), classes-refined serialization (CRS), global serialized mamba (GSM), point cloud, semantic segmentation.**

I. Introduction

THREE-DIMENSIONAL LiDAR point cloud semantic segmentation plays an important role in the fields of remote sensing [1], automated driving [2], and robotic navigation [3], [4], while significantly improving the performance of downstream tasks such as occupancy complementation [5] and 3-D reconstruction [6], [7]. Especially in urban scenes, point cloud semantic segmentation enhance the spatial analysis and urban planning [8], optimize the urban 3-D modeling process, promote low-altitude intelligent transportation systems [9], and support critical tasks of digital twin cities [10]. Due to the complexity of urban scenes, the network requires multiscale feature extraction capabilities to capture geometric information at various densities and heights and to improve the accuracy of semantic segmentation of point clouds.

For accurate segmentation of the point clouds, projection-based methods [11], [12] and voxel-based methods [13], [14], [15], [16] were proposed to leverage 2-D or 3-D convolutional neural networks (CNNs) for feature extraction. However, the projection inevitably changes the spatial topology and local geometric information, while the voxelization insufficient to fully extract the semantic information in complex scenes. PointNet [17] and its successor, PointNet++ [18], avoid the projection by directly working on point clouds and driven the evolution of point-based methods [19], [20], [21], [22]. The limited receptive field of K-nearest neighbors (KNNs) of these methods restricts the extraction of global context information and semantic relations in large-scale scenes. PCT [23] and point transformer [24] used attention mechanism to capture long-distance context information. Yet, the computational complexity grows quadratically with the increase of the number of points.

To reduce the computational complexity while expanding the receptive field, stratified transformer [25] constructed window grouping and hierarchical structure, but this localized attention obstructs the extraction of global features. Meanwhile, mamba [26], as a new selective state-space model (SSM), provided excellent performance in global context modeling with linear complexity. The lack of specific local modeling in the mamba block leads to insufficient performance when capturing fine-grained information. Recently, space-filling curves (SFCs) [27], [28] maps 3-D point clouds to 1-D serialized orders, which show excellent performance in local preservation and global modeling. The transformer-based [29], [30] and mamba-based methods [31] encoded SFCs to enhance the performance of point cloud understanding. However, the discrepancy of locality-preserving and traversal orders among

Received 1 June 2025; revised 14 August 2025; accepted 30 September 2025. Date of publication 3 October 2025; date of current version 4 November 2025. This work was supported in part by Ningbo Science and Technology Innovation Yongjiang 2035 Key Technology Project under Grant 2024Z298 and in part by the Key R&D Program of Ningxia under Grant 2024BEG02042. *(Corresponding author: Fei Deng.)*

Huchen Li and Wubiao Huang are with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: lihuchen@whu.edu.cn; huangwubiao@whu.edu.cn).

Jiacheng Liu is with the Geospatial Engineering, School of Civil and Environmental Engineering, UNSW, Sydney, NSW 2033, Australia (e-mail: jiacheng.liu4@unsw.edu.au).

Ke Chen is with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: c_ke@nuist.edu.cn).

Fei Deng is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China, and also with Hubei Luojia Laboratory, Wuhan 430079, China (e-mail: fdeng@sgg.whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3617326

SFCs limits the semantic segmentation performance of these methods in large-scale complex scenes.

Therefore, to explore the locality-preserving of different SFCs and the performance of transformer and mamba modules for local and global information extraction, a grid point serialized transformer (GridPSFormer) is proposed for semantic segmentation of point clouds in various densities and heights scenes. First, points are gridded in GridPSFormer to reduce scene redundancy and to facilitate sparse convolution to process serialized features efficiently. Subsequently, based on *Z*-order [32] and Hilbert [33] curves, GridPSFormer constructs 3-D spatial serialized order (SSO) module to fully perceive 3-D space. Meanwhile, the local features of the serialized orders are aggregated by the local serialized attention (LSA) mechanism, while the mamba module extracts the global serialized information. Finally, classes-refined serialization (CRS) module reduces regional class imbalance through semantic context information and enhance the generatability in diverse height and density scenes. The contributions of this article are as follows.

1) We innovatively design a 3-D SSO module that can completely perceive the 3-D spatial structure through varying traversal priorities.
2) The global serialized mamba (GSM) module is proposed to extract global serialized features efficiently with linear complexity. Furthermore, the combination of mamba and attention mechanism ensures that the network can focus on local detail features and global semantic and spatial relations.
3) The CRS module is introduced to effectively complement the semantic context information in the serialized orders and enhance the generalizability of the GridPSFormer.
4) GridPSFormer was trained and inferenced on three datasets with various densities and heights, and the experiments showed that the proposed method achieves the best performance.

The remainder of this article is organized as follows. Section II reviews the semantic segmentation methods and the application of SFCs in 3-D point cloud understanding tasks. Section III describes the GridPSFormer network in detail. Section IV demonstrates comparison and ablation experiments. Section V summarizes the results and gives an outlook for future research.

II. Related Work

A. Point Cloud Semantic Segmentation

The semantic segmentation of point clouds divides each point into sets with the same labels based on spatial geometry and shape information. Projection-based methods project the point clouds into 2-D views such as range view or bird-eye view [11], [12]. However, single view only learns view-specific representations and thus do not extract information on occluded areas. Voxel-based methods regularize point clouds in 3-D space. Models like MinkowskiNet [13], SPUNet [16], and SPVCNN [14] efficiently processed high-dimensional sparse data, driving the application of sparse convolution in large-scale modeling and multimodal fusion systems. Based on these sparse networks, Cylinder3D [15] proposed a cylindrical convolution network to solve the uneven distribution of point clouds density in outdoor scenes. RPVNet [34] fused points, voxels, and range views with a gating mechanism to adaptively select feature information. MVPNet [10] aggregated point and voxel features through a multiscale receptive field module to fully extract context information. Although sparse convolution improves the inference efficiency, it is still challenging to extract semantic information of all objects in various densities and heights scenes. Our method leverages sparse convolution features as conditional positional encoding for the attention and mamba modules, enhancing local and global geometric modeling of point clouds.

PointNet [17] pioneered a new paradigm of directly processing 3-D coordinates of point clouds. The successor, PointNet++ [18], extracted local features with sampling and grouping strategies, showing that interactive learning among neighbor points helps to improve the network performance. KPConv [19] then enhanced the performance of dense regions by learning the offset weights between kernel points and neighbor points. RandLA-Net [20] used a local feature extraction module with predefined KNNs clustering and attention pooling to enhance the extraction of local features. Based on these, SCF-Net [35], BAF-LACNet [36], and LACV-Net [37] further enhanced the efficiency of local feature aggregation and global feature extraction to optimize the semantic segmentation of point clouds in large-scale scenes. Point-based methods can accurately acquire fine-grained information, but the predefined kernel points and the limited receptive field of KNNs still limit their semantic segmentation performance in complex scenes. Unlike the KNN methods, our method fully extracts the 3-D spatial structure information through various SFCs.

B. Point Transformers and Mambas

Transformer has a widespread application in 3-D point cloud understanding tasks. PCT [23] and point transformer [24] pioneered point cloud semantic segmentation with transformer. With the increase of the point numbers, the computational complexity of the model grows quadratically. Stratified transformer [25] then proposed a window attention mechanism to group point clouds and learn multiscale features. However, the window patches vary significantly due to the sparsity of the points, which further increases the memory cost. To alleviate the effects of sparsity, ASGFormer [38] combined graph and transformer to facilitate the understanding of global correlations. PVCFormer [39] performed point and voxel cross-perception to enhance the perception and modeling of complex scenes.

Compared with transformer, mamba [26], as a new selective SSM, provides excellent performance in global context modeling with linear complexity. Point Mamba [40] and Point Cloud Mamba [41] introduced mamba for point cloud semantic segmentation, but the lack of an explicit local geometric extraction structure in the mamba block leads to weak performance in segmentation. Mamba3D [42] extended the local geometry extraction capability by the bidirectional SSM module combined with K-norm and K-pooling blocks. Voxel

Mamba [43] proposed a hierarchical SSM structure to extract voxel features at larger scales. Recently, LGMamba [31] has effectively improved the performance of mamba in large-scale scene by local mamba and global mamba blocks. In contrast, our method explores the performance of the LSA mechanism combined with GSM in the semantic segmentation of point clouds.

C. Serialization-Based Methods

SFCs [28], like scanning curves, gray codes [44], *Z*-order [32], and Hilbert [27], [33], are continuous and nonderivable fractal curves that map elements in high-dimension space to 1-D information while preserving high-dimension properties. Xiang et al. [45] serialized the disordered points in the neighborhood sphere by z-order to effectively extract the complete geometric information. Chen et al. [46] proved that Hilbert curves have better spatial preservation compared with other curves. HilbertNet [47] combined Hilbert curves and transformer block to improve segmentation performance. OctFormer [29] divided the point clouds into some groups based on sorted shuffled keys of octrees, which improves the parallelization and scalability of the window attention. Point Transformer v3 [30] combined *Z*-order and Hilbert curves to improve semantic segmentation performance while reducing computational complexity significantly. Point Mamba [40] introduced Hilbert and trans-Hilbert curves to provide different perspectives on spatial locality. Pamba [48] adopted hybrid serialization to complement local geometric modeling in mamba block. Recently, grid mamba [49] mapped point clouds into serialized orders from multiple spatial views and confirmed that multiview scanning can effectively alleviate the loss of local proximity in space. UniMamba [50] investigated *Z*-order serialization in both vertical and horizontal directions. In contrast, our method further considers the traversal priority of all three coordinate axes in different serialized orders and its impact on semantic segmentation performance in various densities and heights scenes.

III. Methodology

A. Overview

The overall structure of GridPSFormer is shown in Fig. 1. The input points are mapped to discretized grids based on 3-D spatial coordinates, and a random point from each grid is retained to reduce redundancy while expanding the receptive field. Based on the *Z*-order [32] and Hilbert [33] curves, the 3-D SSO module provides six random serialized strategies to achieve a complete perception of the 3-D space. The GSM module is designed to learn the spatial distribution of large-scale scenes and the long-distance dependencies between objects with linear complexity. The CRS module further enhances the context-awareness of serialized orders and improves the semantic segmentation performance in multi-density and multi-height scenes.

As show in Fig. 2, the backbone is a U-Net structure with four encoder and decoder layers. Specifically, the serialized orders are embedded in the initialization block by sparse convolution to enhance the local feature representation. Each

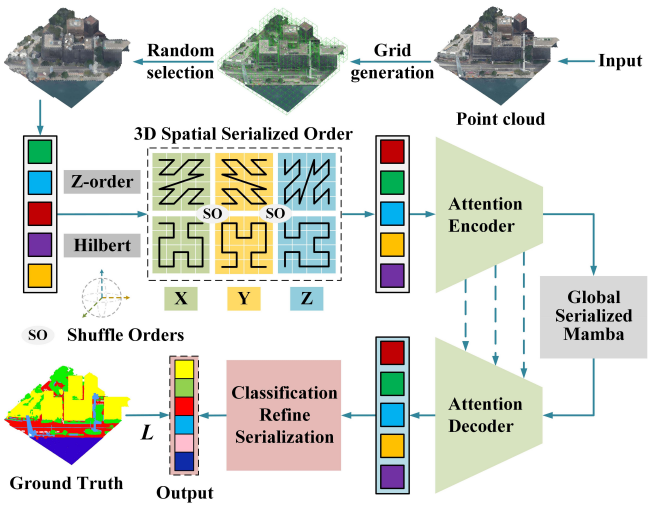

Fig. 1. Overview of GridPSFormer.

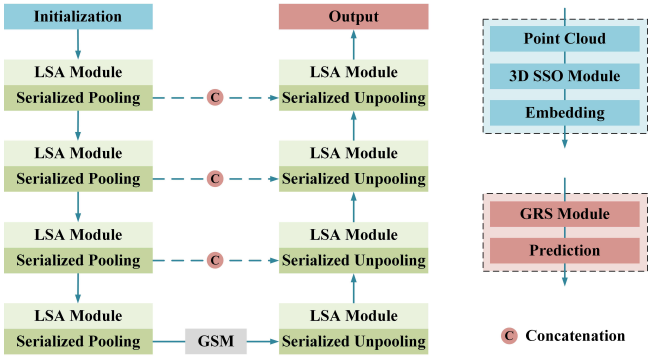

Fig. 2. Encoder and decoder structure of GridPSFormer.

encoder or decoder layer contains the LSA module and the serialized pooling or unpooling block, both based on Point Transformer v3 [30]. The LSA module groups and interacts with the serialized patches and extracts local features. Both synergize LSA and GSM modules to obtain local and global semantic and spatial information. Sections III-B–III-E describe the composition of each module of GridPSFromer.

B. Grid Points

Based on the performance of sparse voxelization [16] and SFCs [28] in 3-D point cloud comprehension tasks, we build the grid points for efficiently mapping 3-D point clouds to 1-D serialized orders.

1) Grid Points Generation: Due to the disorder, irregularity, and uneven density of the point clouds, direct serialization introduces lots of repetitive and irrelevant points, which increases the computational burden and memory cost. Point clouds are gridded to effectively reduce the unnecessary information and memory and better retain the spatial relations of the local geometric structure. The calculation formula is as follows:

$$\mathbf{GP} = \left\lfloor \frac{\mathbf{P}}{g} \right\rfloor - \mathbf{Min}\left(\left\lfloor \frac{\mathbf{P}}{g} \right\rfloor \right) \tag{1}$$

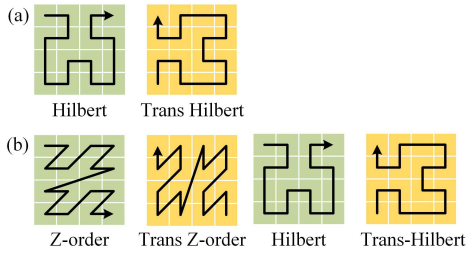

Fig. 3. Point cloud serialization. (a) Point Mamba. (b) Point Transformer v3.

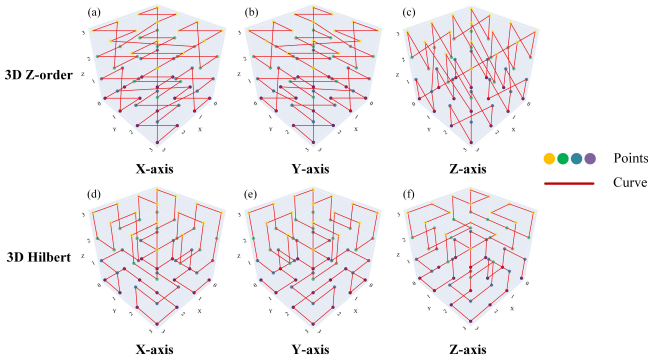

Fig. 4. 3-D SSO. (a)–(c) Z-order with different axes traversals and (d)–(f) Hilbert curve with different axes traversals.

where $\mathbf{P} = (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in \mathbf{R}^3$ donates point cloud coordinates, g donates grid size, $\lfloor \cdot \rfloor$ donates floor operation, and $\mathbf{GP}$ donates gridded points.

2) Random Selection: We randomly choose a point within each grid as the serialized feature for that grid. This approach not only improves computational parallelism but also reduces the complexity of processing irregular point clouds. Then, the gridded points are encoded into the serialized orders, which effectively expands the receptive field and learns enriched spatial context information.

C. 3-D SSO Module

Point Mamba [40] combines the Hilbert curve with its transposed form (trans-Hilbert) to enhance the spatial modeling capabilities. However, it essentially relies solely on the Hilbert space-filling pattern [see Fig. 3(a)], which limits its ability to capture the diverse and complex geometric variations in 3-D scenes. In comparison, Point Transformer v3 [30] integrates both Z-order and Hilbert curves and utilizes their original and transposed versions to provide dual spatial mappings [see Fig. 3(b)]. Nevertheless, due to the strong directionality and structural variability in 3-D space, local geometric relationships in key directions may still be overlooked with limited transformation strategies.

To address these limitations, we propose the 3-D SSO module, which flexibly captures spatial proximity from multiple directions by reordering the coordinate traversal. As illustrated in Fig. 4, the colored points represent gridded points at different heights, while the solid red lines denote SFCs. The different SFCs capture varying spatial proximity even when traversed in the same order, while a single curve can reveal different spatial relationships when its traversal order is altered. This flexibility enables 3-D SSO to adaptively model spatial dependencies from multiple perspectives. The detailed formulation is shown as follows:

$$\mathbf{GP}_\sigma = \tau(\mathbf{GP}, \sigma) \tag{2}$$

$$\mathrm{SSO}(\mathbf{GP}_\sigma, b, s) = (b \ll k) \,|\, \varphi_s^{-1}(\mathbf{GP}_\sigma) \tag{3}$$

where σ represents arbitrarily permuted coordinates, τ denotes the coordinate permutation function, $\mathbf{GP}_\sigma$ denotes gridded points at σ order, b denotes the batch index, k represents the number of bits to shift, $\ll$ denotes the left shift operator, φ_s^{-1} is the inverse mapping, s refers to the Z-order or Hilbert curve, and $|$ denotes the bitwise OR.

Building on this flexibility, the Z-order and Hilbert curves in 3-D SSO achieve superior spatial locality preservation for point clouds serialization. By altering the traversal priority of the coordinate axes, we obtain six distinct spatial perspectives, which comprehensively extract the relative position and geometric shapes of the objects, and enhance the complete perception of the 3-D space. As a result, 3-D SSO can accurately capture high-level buildings, low-level roads, and infrastructures in various density and height scenes.

Furthermore, in the LSA and GSM blocks of the encoder–decoder architecture (see Fig. 2), we use a shuffle order (SO) strategy [30] to select an order from the multiple spatial serialization modalities pregenerated by the 3-D SSO module. This selected order is then used to serialize the embedded features $\mathbf{F_e}$, and later to revert them via deserialization, as shown as follows:

$$\mathbf{F_s} = \mathrm{SO}_{\sigma,s}(\mathbf{F_e}) \tag{4}$$

$$\mathbf{F_e} = \mathrm{SO}^{-1}_{\sigma,s}(\mathbf{F_s}) \tag{5}$$

where $\mathbf{F_e}$ denotes the embedded features extracted by sparse convolution in the initialization stage (see Fig. 2) and propagated through different network layers, $\mathbf{F_s}$ denotes serialized features, and $\mathrm{SO}_{\sigma,s}$ denotes a randomly selected order from either the Z-order or the Hilbert curve at σ order. Specifically, we randomly shuffle serialized orders at different stages, enabling the network to learn multiple structured representations at varying spatial scales and reducing serialization dependency, thereby improving its robustness and generalizability.

D. GSM Module

Mamba is a selective SSM with comparable modeling ability to the transformer. Its computational complexity scales linearly with the input length, thus effectively capturing long-distance dependencies and improving the efficiency of training and inference [26]. Through the zero-order holding (ZOH), discretized SSM can be directly applied to deep learning tasks. Moreover, mamba is able to dynamically adjust its information propagation or filtering mechanism according to different serialized orders [42]. When combined with the sparse convolution, mamba can fully exploit 3-D spatial features and enhance its flexibility and expressiveness in complex 3-D scenes.

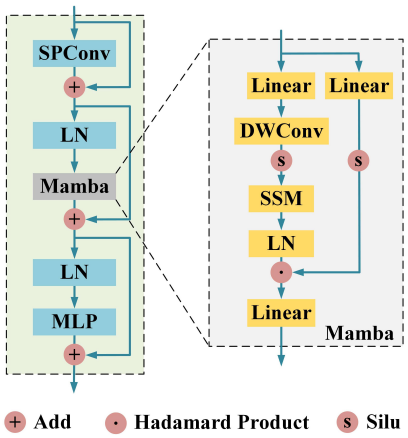

Fig. 5. GSM module.

Therefore, we constructed the GSM module to facilitate the network in extracting global context information (see Fig. 5). The detailed formula is as follows:

$$\begin{aligned} \mathbf{F}_{\mathbf{e}}' &= \mathrm{SPConv}\left(\mathbf{F}_{\mathbf{e}}\right) + \mathbf{F}_{\mathbf{e}} \\ \mathbf{F}_{\mathbf{e}}'' &= \mathrm{Mamba}\left(\mathrm{SO}\left(\mathrm{LN}\left(\mathbf{F}_{\mathbf{e}}'\right)\right)\right) + \mathbf{F}_{\mathbf{e}}' \\ \mathrm{GSM}\left(\mathbf{F}_{\mathbf{e}}\right) &= \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{F}_{\mathbf{e}}''\right)\right) + \mathbf{F}_{\mathbf{e}}'' \end{aligned} \tag{6}$$

where SPConv denotes sparse convolution [16] to enhance the representation of global spatial relations. The features learned from SPConv are fused with the $\mathbf{F_e}$ and fed into the SO module to generate a global serialization order, which ultimately leads to global modeling in mamba block. The residual connected features are normalized by layer normalization (LN) around the mamba block to maintain the stability of feature scales and enhance the training efficiency and convergence. Subsequently, the MLP further improves the global feature representation. The detailed formula for the mamba block is

$$\begin{aligned} \mathbf{M}' &= \mathrm{Silu}\left(\mathrm{DWConv}\left(\mathrm{Linear}\left(\mathbf{F}_{\mathbf{s}}'\right)\right)\right) \\ \mathbf{M}'' &= \mathrm{LN}\left(\mathrm{SSM}\left(\mathbf{M}'\right)\right) \\ \mathrm{Mamba}\left(\mathbf{F_s}\right) &= \mathrm{Linear}\left(\mathbf{M}'' \cdot \mathrm{Silu}\left(\mathrm{Linear}\left(\mathbf{M}\right)\right)\right) \end{aligned} \tag{7}$$

where $\mathbf{F}_{\mathbf{s}}'$ denotes serialized features, obtained from $\mathbf{F}_{\mathbf{e}}'$ after the RSO operations, Linear denotes the linear layer, DWConv denotes the depth separable convolution with the kernel set to 5, and Silu is an activation function.

Owing to the linear complexity of SSM, mamba not only parallelizes training like CNNs, but also efficient inference like recurrent neural networks (RNNs), which significantly improves the computation efficiency and global modeling performance. In summary, the LSA mechanism extracts local detail information, while the serialized mamba extracts the overall spatial features in the global range. The combination of the two modules drives the network to extract fine-grained features while learning the global semantic and spatial relations.

E. CRS Module

In point cloud semantic segmentation of large-scale urban scenes, the whole scene is typically cropped into many sub-scenes to accommodate the input size of the network. However, the cropping strategy increases class imbalance, causing some sparse classes to be easily overlooked in serialized orders and further reducing the segmentation accuracy. Meanwhile, this strategy breaks semantic context relations between objects (e.g., roads, buildings, and vegetation) in urban scenes, which play a crucial role in the segmentation task [51].

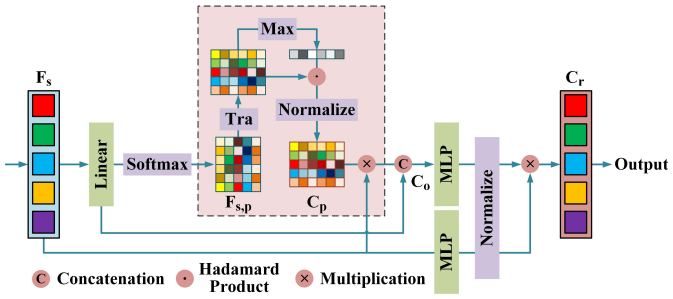

Fig. 6. CRS module.

Although Z-order and Hilbert curves maintain the spatial locality-preserving of the point clouds, they ignore the semantic relations and spatial associations between objects. The serialized orders will compress or distort key geometric semantic information when the scale or spatial location of the objects changes. Therefore, the CRS module is proposed, as shown in Fig. 6, to constrain the relations between objects in the scene and enhance the perceptibility of the serialized context information. The calculation formula is as follows:

$$\mathbf{F_{s,p}} = \mathrm{Softmax}\left(\mathrm{Linear}\left(\mathbf{F_s}\right)\right) \tag{8}$$

$$\mathbf{C_p} = \mathrm{Norm}\left(\mathrm{Max}\left(\mathbf{F}_{\mathbf{s,p}}^{\mathbf{T}}\right) \cdot \mathbf{F}_{\mathbf{s,p}}^{\mathbf{T}}\right) \tag{9}$$

where $\mathbf{F_{s,p}}$ denotes probabilistic serialized features, Norm denotes the normalization function, and $\mathbf{C_p}$ denotes the prediction class information. The max function is used to select the serialized information with a confidence threshold above 0.75, which improves the confidence of the prediction classes.

Then, the $\mathbf{C_p}$ are multiplied with $\mathbf{F_s}$ and concatenated with $\mathbf{F_{s,p}}$ via the residual structure to optimize the class context information. The calculation formula is as follows:

$$\mathbf{C_o} = \delta\left(\mathrm{Concat}\left(\mathbf{C_p} * \mathbf{F_s}, \mathrm{Liner}\left(\mathbf{F}_{\mathbf{s,p}}^{\mathbf{T}}\right)\right)\right) \tag{10}$$

where δ denotes a mapping function and $\mathbf{C_O}$ denotes optimized class information. By introducing semantic context information, the network not only focuses on spatial locations but also captures semantic relations between objects, for example, roads and lights, wires, and poles, to enhance the understanding of inter-relations between objects in complex scenes. Further, to facilitate feature fusion, the class context-optimized features are multiplied with original serialized features. The calculation formula is as follows:

$$\mathbf{C_r} = \mathrm{Norm}\left(\mathrm{MLP}\left(\mathbf{F_s}\right)\right) * \mathrm{Norm}\left(\mathrm{MLP}\left(\mathbf{C_o}\right)\right) \tag{11}$$

where $\mathbf{C_r}$ denotes refined serialized prediction. The spatial relations of objects are refined through classes information,

TABLE I
3-D POINT CLOUD SEMANTIC SEGMENTATION DATASET

Dataset	Region	Height (m)	Area (km^2)	Point (Mil.)	Density (pt/m^2)	Classes
HRHD_HK	Hong Kong, China	38.50	9.38	273	29	7
WHU_ALS	Wuhan and Shanghai, China	19.11	3.20	213	67	8
DALES	Surrey, Canada	8.83	10.00	505	51	8

thus further improving the semantic segmentation performance of serialized points in complex urban scenes.

IV. EXPERIMENTS AND ANALYSIS

A. Experimental Setting and Implementation Details

In this experiment, the batch size of GridPSFormer was set to 6, the optimizer adopted AdamW with a one-cycle learning rate scheduler, the initial learning rate was set to 0.0002, the weight decay was set to 0.005, and the total epoch of training was set to 500. For all datasets, a grid size of 0.15 m was used for sampling and random spherical cropping, with a maximum of 102 400 points. The 3-D coordinates were augmented by center shift, random rotation, random scaling, and random flipping. The color values were normalized to [−1.0, 1.0], and the intensity and normal values were normalized to [0, 1.0]. The coordinates and the rest of the values were used as input features to participate in the model training, and the gradient propagation was performed using the cross-entropy loss function and the Lovász function.

In the training phase, the 3-D SSO utilized six serialized orders based on Z-order and Hilbert curves. For each block within the layer, one of these orders was randomly selected for serialization. The network contained four encoding and decoding stages, with feature dimensions of [64, 128, 256, 512]. The number of LSA modules per layer was [2, 2, 2, 6] in the encoder and [2, 2, 2, 2] in the decoder. Additionally, the patch length in the LSA was set to 1024. In the inference phase, the random scale strategy was used for data augmentation. The performance of semantic segmentation of comparative methods was evaluated by overall accuracy (OA), mean Intersection over Union (mIoU), and per-class IoU metrics, and the performance of different serialized orders was evaluated by OA, mean Accuracy (mAcc), and mIoU metrics. GridPSFormer was performed on Intel[1] Xeon[1] 6133 CPU and three NVIDIA RTX A6000 GPU servers and was trained and tested using PyTorch 2.1.0 deep learning framework in Python 3.8 environment.

B. Datasets

As shown in Table I, we selected three public datasets from different regions to fully validate the performance of GridPSFormer for semantic segmentation in various density and height scenes.

HRHD_HK [52] is a high-rise and high-density dataset collected in Hong Kong, China by photogrammetry, covering about 9.38 km^2 with 273 million points [see Fig. 7(a)]. HRHD_HK contains 150 tiles, 104 tiles for training, another 23 tiles for validation, and the last 23 tiles for testing. The classes consist of building, vegetation, road, waterbody, facility, terrain, and vehicle, in which the average height of the building is 38.50 m, and each class contains (X, Y, Z) 3-D coordinates and (R, G, B) color information.

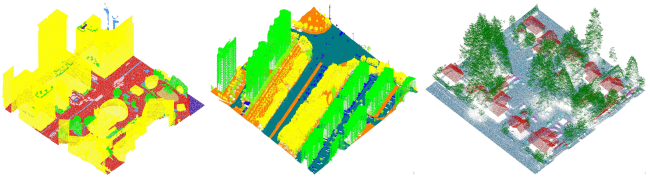

Fig. 7. Visualization of semantic segmentation datasets. (a) HRHD_HK, (b) WHU_ALS, and (c) DALES.

WHU_ALS is a part of the WHU-Urban3D [1] dataset, which was collected in Wuhan and Shanghai, China using airborne laser scanning (ALS), and covered more than 3.2 km^2 with around 213 million points [see Fig. 7(b)]. WHU_ALS contains 86 tiles, 59 tiles for training, another seven tiles for validation, and the last 20 tiles for testing. According to the preprocessing method provided by the official website of WHU-Urban3D (https://whu3d.com/tutorial.html), we utilized the pywhu3d library to map the labels and compute the normal vector with a radius of 0.8 m. The classes consist of ground, building, tree, vegetation, low vegetation, light, and wire, in which the average height of buildings is 38.50 m. Each point contains (X, Y, Z) 3-D coordinates and the normal vector information.

DALES [53] is an airborne LiDAR dataset collected in Surrey, BC, Canada, with a total of approximately 505 million points covering an area of more than 10 km^2 [see Fig. 7(c)]. DALES contains 40 tiles; 29 tiles are used for training and the remaining 11 tiles are used for testing. The original size of each tile was about 500 × 500 m, and we cropped each tile into 100 × 100 m sub-tiles to improve the computation efficiency. The classes consist of ground, vegetation, car, truck, powerline, fence, pole, and building, in which the average height of the building is 8.83 m. Each point contains (X, Y, Z) 3-D coordinates and intensity information.

C. Comparison With State-of-the-Art Methods

To validate the effectiveness and robustness of GridPSFormer on three datasets, HRHD_HK, WHU_ALS, and DALES, we compared it with various methods for

[1]Registered trademark.

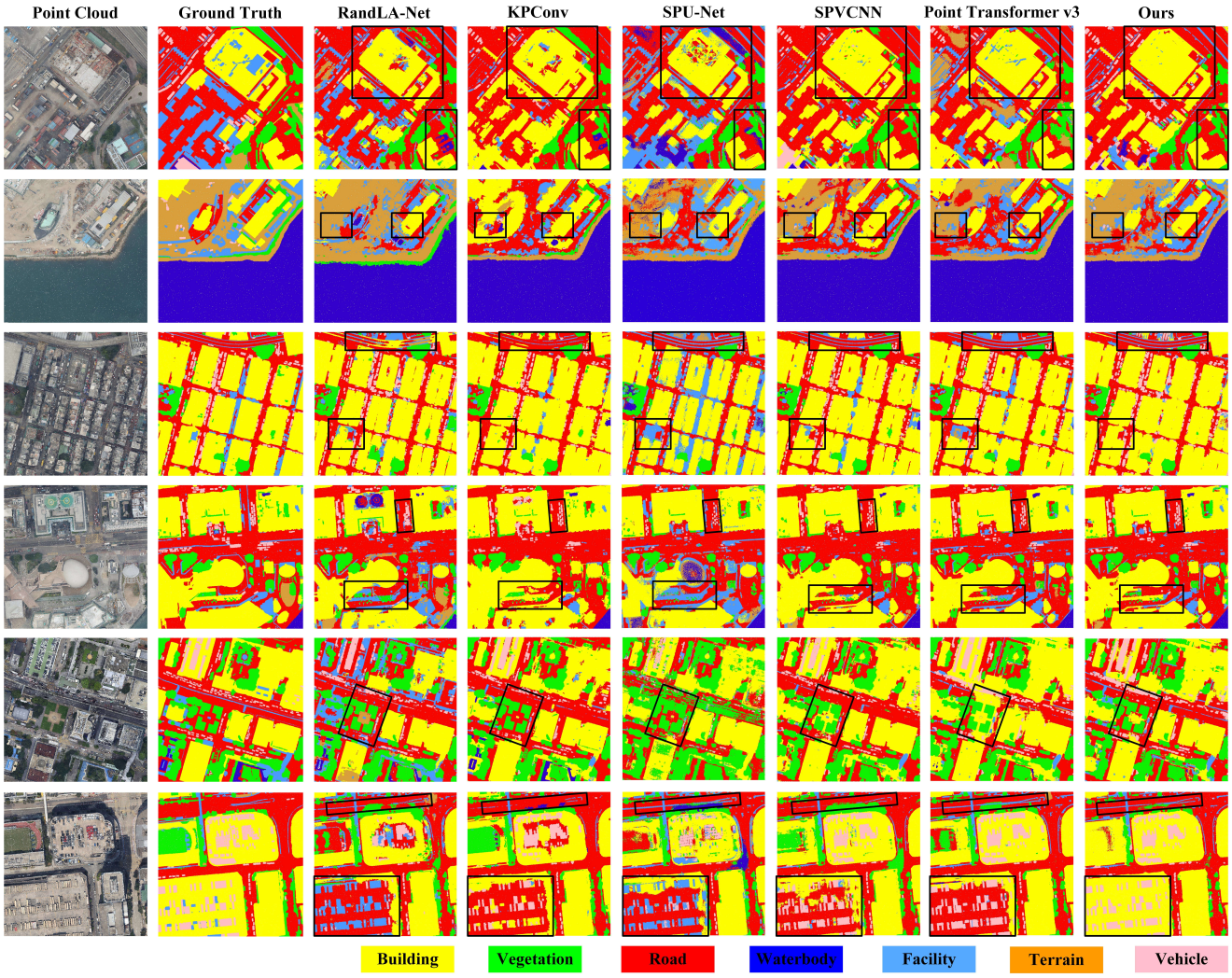

Fig. 8. Visualization of semantic segmentation results on HRHD_HK dataset.

quantitative and qualitative analysis. The comparative methods include KNNs-based methods (RandLANet [20], SCFNet [35], LACVNet [37]), graph convolution-based methods (BAF-LACNet [36]), point convolution-based methods (KPConv [19]), sparse convolution-based methods (SPUNet [16]), fusion-based methods (SPVCNN [14], MVPNet [10]), transformer-based methods (Point Transformer [24], Oct-Former [29], Point Transformer v3 [30], super Point Transformer [54]), and mamba-based methods (LGMamba [31]). This section explores the performance differences of these network architectures in real-world tasks.

1) HRHD_HK: The visualization of the segmentation results on the HRHD_HK dataset is shown in Fig. 8. RandLANet better identified facilities such as containers on the wharf but mistaken parts of the high roads as buildings. KPConv confused terrains on the ground, that is, bare land, with built-up roads and struggled to extract fenced facilities on the roads. SPUNet identified low buildings as facilities and shaded areas of buildings as water bodies. Due to the similar structure of a few classes, SPVCNN identified some vehicles as roadway facilities, and Point Transformer v3 blurred the boundary between high-rise buildings and facilities. In addition, the above methods wrongly detected large flat surfaces like high-rise building roofs as roads due to the higher similarity between large building roofs and sidewalks and roads. Compared with these methods, GridPSFromer effectively maintained the structure integrity of buildings and roads while accurately identifying vehicles on roads as well as on roofs. By randomly traversing different coordinate-prioritized serialized orders, GridPSFromer learned 3-D spatial relationships from multiple perspectives, which significantly improved semantic segmentation on high and dense building areas.

Moreover, compared with the quantitative results of other state-of-the-art methods in Table II, GridPSFromer achieved the optimal metrics in terms of OA and mIoU, which are improved by 0.58% and 2.27% over Point Transformer v3. KPConv achieved the best IoUs on vegetation and waterbodies, MVPNet achieved the best IoUs on roads, Point Mamba achieved the best IoUs on terrains, and Point Transformer v3 achieved the best IoUs on facilities. GridPSFromer achieved the optimal IoUs on buildings and vehicles, with an

TABLE II

COMPARISON OF SEMANTIC SEGMENTATION RESULTS OF DIFFERENT METHODS ON THE HRHD_HK DATASET. BOLDED SCORES INDICATE OPTIMAL PERFORMANCE OF THE METRICS, WHILE THE SCORES IN UNDERLINE ARE SECOND ONLY TO THE BEST

Methods	OA (%)	mIoU (%)	IoUs (%)						
			Build.	Veg.	Road	Water.	Facility	Terrain	Vehicle
RandLANet	87.48	57.02	89.59	85.22	58.13	86.99	25.68	30.17	23.38
SCFNet	88.11	56.78	90.31	87.16	58.98	88.37	25.31	28.40	18.97
BAF-LACNet	85.52	55.39	86.06	84.25	57.93	93.34	22.17	23.74	20.25
LACVNet	88.67	58.55	91.26	88.90	57.68	75.97	28.82	27.52	39.73
KPConv	91.08	60.69	90.98	**91.89**	62.82	**95.49**	25.23	17.84	40.57
SPUNet	82.92	47.60	85.77	70.28	53.05	67.27	24.71	17.27	14.84
SPVCNN	90.92	62.14	92.60	88.98	62.26	89.52	26.61	23.64	51.38
MVPNet	90.80	58.91	92.29	90.99	**64.40**	89.53	26.18	32.09	16.88
Point Mamba	89.91	61.75	92.23	87.76	58.09	89.86	30.83	**33.14**	40.37
Point Transformer v3	90.83	62.62	92.83	89.27	62.00	91.23	**32.42**	30.27	40.32
GridPSFromer (ours)	**91.41**	**64.89**	**92.85**	88.83	64.14	93.81	31.64	29.26	**53.72**

Fig. 9. Visualization of semantic segmentation results on WHU_ALS dataset.

improvement of 0.02% and 13.4% over Point Transformer v3. The quantitative results fully validated the effectiveness of SSOs.

2) WHU_ALS: The visualization of the segmentation results on the WHU_ALS data is shown in Fig. 9. The vegetation in this dataset was subdivided into individual trees, continuously distributed high vegetation and low vegetation. These classes have strong geometric relationships with each other and closer spatial proximity, which puts more challenges to accurate segmentation. Compared with other methods,

TABLE III
COMPARISON OF SEMANTIC SEGMENTATION RESULTS OF DIFFERENT METHODS ON THE WHU_ALS DATASET. BOLDED SCORES INDICATE OPTIMAL PERFORMANCE OF THE METRICS, WHILE THE SCORES IN UNDERLINE ARE SECOND ONLY TO THE BEST

Methods	OA (%)	mIoU (%)	IoUs (%)						
			Ground	Build.	Tree	Veg.	Low veg.	Light	Wire
RandLANet	77.85	55.88	61.83	81.00	37.25	67.94	<u>49.09</u>	28.89	65.15
SCFNet	76.05	56.58	56.20	80.00	40.19	66.33	48.60	33.86	70.85
BAF-LACNet	77.09	57.51	60.68	80.25	35.19	69.42	44.19	32.78	80.04
LACVNet	7743	59.36	63.34	79.90	42.10	69.37	34.35	42.66	83.78
KPConv	80.92	55.39	68.71	82.04	32.73	<u>71.59</u>	**51.16**	0.00	81.51
SPUNet	74.29	45.95	51.99	68.65	44.21	73.36	45.42	1.12	36.91
SPVCNN	81.29	60.62	<u>69.01</u>	83.68	<u>45.90</u>	71.56	47.99	35.36	70.84
MVPNet	79.23	59.05	64.92	81.14	37.34	70.43	47.19	36.93	75.40
Point Transformer	79.20	57.20	66.70	<u>86.20</u>	39.20	70.00	43.90	27.80	66.70
OctFormer	74.30	44.50	62.90	80.00	11.00	66.50	27.20	0.00	63.90
Point Mamba	<u>77.88</u>	<u>62.46</u>	63.73	82.85	44.12	69.79	33.98	57.08	<u>85.70</u>
Point Transformer v3	75.03	56.93	56.75	67.37	43.19	**73.68**	41.06	<u>57.63</u>	58.83
GridPSFormer (ours)	**82.99**	**68.17**	**73.31**	**90.32**	**46.51**	71.08	46.79	**58.01**	**91.19**

GridPSFormer had better structural integrity and segmentation accuracy on trees and vegetation. SPUNet wrongly segmented the whole power tower and part of the wires into buildings, RandLANet wrongly segmented the whole power tower into buildings, and SPVCNN and Point Transformer v3 wrongly segmented the bottom and top of the power towers into buildings, respectively. Moreover, SPUNet and Point Transformer v3 severely misclassified highways and buildings. In contrast, GridPSFormer accurately segmented the power wires and their appurtenances while better identified the highways and buildings. Due to the GSM module, GridPSFormer efficiently processed long-range relations between objects and avoided local deviations, thus achieving accurate semantic segmentation in complex urban scenes.

Table III shows the quantitative comparison of the semantic segmentation results of the different methods on the WHU_ALS dataset. GridPSFormer achieved the optimal metrics in terms of OA and mIoU, which improved by 7.96% and 11.24% over Point Transformer v3. KPConv and OctFormer were incapable of extracting the lights, and SPUNet performed relatively poorly in this class. KPConv achieved the best IoUs in low vegetation and Point Transformer v3 in vegetation. Compared to these three methods, GridPSFormer achieved the best IoUs in trees and balanced optimization of IoUs in the vegetation-similar classes, that is, trees, vegetation, and low vegetation. Moreover, GridPSFormer achieved the optimal IoUs in the ground, buildings, lights, and wires. In particular, GridPSFormer achieved significant improvement on wires compared to Point Transformer v3 by 32.36%, respectively. GridPSFormer effectively fused global and local information, balances the segmentation accuracy between all classes, and shows the strong robustness and generalizability in complex urban scenes.

3) DALES: The visualization of the segmentation results on the DALES dataset is shown in Fig. 10. Some buildings in the DALES dataset have relatively lower heights compared to the other two datasets, and some cars and trucks have higher similarities, which affects the effectiveness of point cloud segmentation. For the lower houses, KPConv had difficulty identifying their complete structure, and SPUNet, SPVCNN, and Point Transformer v3 inaccurately segmented them as ground. RandLANet confused some of the smaller buildings and trucks, and the segmentation boundaries of the buildings were more ambiguous. Compared with these methods, GridPSFormer accurately and completely segmented the buildings. In the area where vegetation and fences are closely adjacent to each other, GridPSFormer excelled at extracting fences, validating that the context information of the category-guided sequence module effectively improves scene distinguishability. Furthermore, GridPSFormer extracted power lines and poles more effectively and segmented the boundaries of them more clearly. Even in the lower height scenes, GridPSFormer still performed well and effectively handled the challenges of complex spatial relations and similar classes.

Further, compared with the other state-of-the-art methods in Table IV, GridPSFormer achieved the optimal metrics in terms of mIoU, with 0.63% improvement over Point Transformer v3. GridPSFromer significantly outperformed other transformer-based and mamba-based methods. SPVCNN achieved the best IoUs in vegetation and Point Transformer v3 in ground, cars, and powerlines. GridPSFormer achieved the best IoUs in trucks, fences, poles, and buildings. In summary, GridPSFormer effectively improved the performance of point cloud segmentation in complex scenes, showed excellent performance in dealing with diverse scenes, and achieved competitive experimental results.

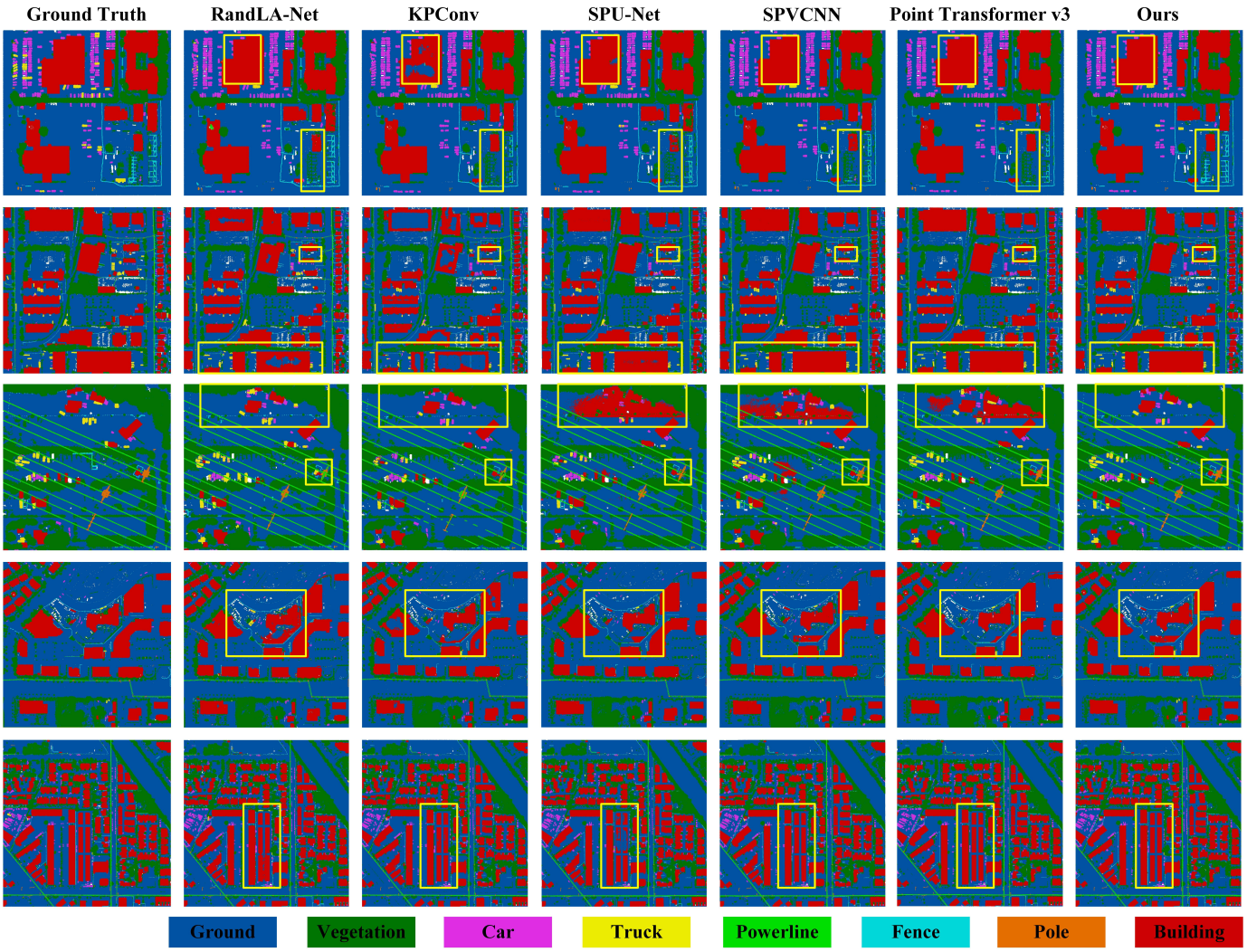

Fig. 10. Visualization of semantic segmentation results on DALES dataset.

TABLE IV

COMPARISON OF SEMANTIC SEGMENTATION RESULTS OF DIFFERENT METHODS ON THE DALES DATASET. BOLDED SCORES INDICATE OPTIMAL PERFORMANCE OF THE METRICS, WHILE THE SCORES IN UNDERLINE ARE SECOND ONLY TO THE BEST

Methods	OA (%)	mIoU (%)	IoUs (%)							
			Ground	Veg.	Car	Truck	Power.	Fence	Pole	Build.
RandLANet	97.74	78.60	97.17	93.95	85.57	43.53	93.66	54.90	63.58	96.46
SCFNet	97.65	78.25	97.10	93.72	83.72	40.86	95.23	51.98	66.93	96.46
BAF-LACNet	97.65	77.92	97.02	93.85	84.11	40.30	94.56	54.20	63.01	96.28
LACVNet	97.86	79.69	97.35	94.33	87.59	45.48	94.83	56.34	65.14	96.45
KPConv	97.70	79.03	96.77	95.00	87.33	37.02	93.41	63.90	64.43	94.37
SPUNet	98.31	83.25	97.68	95.72	89.02	47.71	95.93	69.43	73.58	96.90
SPVCNN	98.38	84.93	97.77	**95.82**	89.93	53.57	96.80	70.97	77.48	97.12
MVPNet	97.92	81.14	97.19	94.69	87.97	46.97	94.55	61.93	69.65	96.19
Point Transformer	97.60	79.30	96.80	93.20	82.20	38.90	95.90	60.60	70.40	96.80
Point Transformer v2	97.30	79.60	96.00	93.30	84.60	39.60	95.30	61.50	71.50	94.90
Super Point Transformer	97.50	79.60	96.70	93.10	86.10	52.40	94.00	52.70	65.30	96.70
LGMamba	97.70	82.30	96.70	93.50	87.20	45.60	96.80	65.90	76.00	96.90
Point Mamba	98.05	79.99	97.45	94.69	87.73	41.02	94.31	63.24	64.49	97.00
Point Transformer v3	**98.43**	84.87	**97.87**	95.56	**90.40**	53.77	**97.27**	70.42	75.85	97.80
GridPSFormer (ours)	98.24	**85.50**	97.53	94.82	90.10	**55.14**	97.13	**71.67**	**79.56**	**98.07**

TABLE V
SEGMENTATION RESULTS OF DIFFERENT MODELS ON THREE DATASETS. BOLDED SCORES INDICATE OPTIMAL PERFORMANCE OF THE METRICS

Model	3D SSO	GSM	CRS	HRHD_HK		WHU_ALS		DALES	
				OA (%)	mIoU (%)	OA (%)	mIoU (%)	OA (%)	mIoU (%)
Baseline				90.83	62.62	75.03	56.93	98.43	84.87
A1	√			90.90	63.07	73.64	59.17	**98.50**	85.07
A2		√		91.21	62.81	82.15	67.37	98.45	85.22
A3			√	91.01	64.22	76.91	62.02	98.39	85.18
A4	√	√		91.23	63.55	82.85	67.54	98.46	85.30
GridPSFormer	√	√	√	**91.41**	**64.89**	**82.99**	**68.17**	98.24	**85.50**

TABLE VI
COMPARISON OF DIFFERENT SERIALIZATIONS ON THE HRHD_HK DATASET

Order	Axis-priority	OA (%)	mIoU (%)	mAcc (%)
Z-order	X-axis	91.07	63.89	72.38
	X+Y-axis	91.26	64.03	73.76
	X+Y+Z-axis	91.24	64.20	74.08
Hilbert	X-axis	91.31	63.93	72.71
	X+Y-axis	91.24	64.01	73.15
	X+Y+Z-axis	91.30	64.15	73.57
Z+Hilbert	X-axis	91.22	64.10	72.88
	X+Y-axis	91.25	64.47	74.02
	X+Z-axis	91.33	64.70	74.36
	X+Y+Z-axis	**91.41**	**64.89**	**74.40**

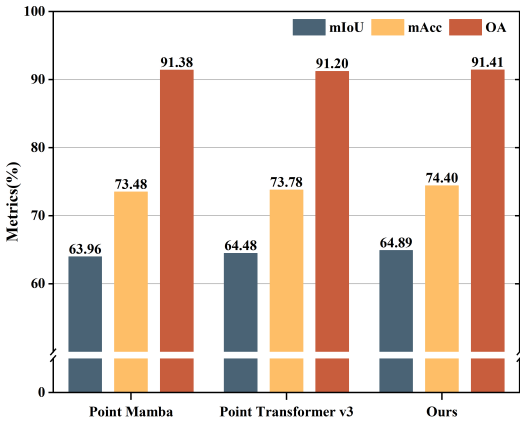

Fig. 11. Comparison of different methods on the HRHD_HK dataset.

D. Ablation Study

On the HRHD_HK, WHU_ALS, and DALES datasets, ablation studies were performed for the 3-D SSO module, the GSM module, and the CRS module, as shown in Table V. Baseline adopts the same architecture as Point Transformer v3 and is evaluated using the parameters trained in Section IV-A. Model A1 represents the addition of the 3-D SSO module to the baseline, model A2 represents the addition of the GSM module to the baseline, model A3 represents the addition of the CRS module to the baseline, and Model A4 adds the 3-D SSO, and GSM modules to the baseline to fully validate the effectiveness of each module.

1) Effect of 3-D SSO Module: Compared with the baseline, model A1 improved 0.07% of OA and 0.45% of mIoU on the HRHD_HK dataset, 2.24% of mIoU on the WHU_ALS dataset, and 0.07% of OA and 0.20% of mIoU on the DALE dataset, respectively. It is shown that the 3-D SSO module can effectively overcome the variation of height and density in complex scenes and improve the spatial perception of the network.

As shown in Table VI, we evaluated six serialized orders in 3-D SSO by permuting the coordinate axes. With the introduction of diverse 3-D SSOs, the model encoded the point clouds more accurately and then combined the local structure and global context information to better recognize and segment objects with complex shapes. The complete integration of all six orders achieved the best performance with 91.41% OA, 64.89% mIoU, and 74.40% mAcc.

Furthermore, as illustrated in Fig. 11, compared with Point Mamba, our method improved mIoU, mAcc, and OA by 0.93%, 0.92%, and 0.03%, respectively. In comparison to Point Transformer v3, the improvements were 0.41% in mIoU, 0.62% in mAcc, and 0.21% in OA. These consistent gains across all three metrics demonstrated that the proposed 3-D SSO with diversified serialized orders can more effectively preserve spatial locality and capture geometric variations, thereby enhancing both segmentation accuracy and overall classification performance.

2) Effect of GSM Module: Compared with the baseline, model A2 improved 0.38% of OA and 0.19% of mIoU on the HRHD_HK dataset, 7.12% of OA, and 10.44% of mIoU on the WHU_ALS dataset, and 0.02% of OA and 0.35% of mIoU on the DALE dataset, respectively. Compared with model A1, model A4 improved 0.33% of OA and 0.48% of mIoU on the HRHD_HK dataset, 9.21% of OA and 8.37% of mIoU on the WHU_ALS dataset, and 0.23% of mIoU on the DALE dataset, respectively. It demonstrated that the GSM module provided a more comprehensive feature representation and understood the scene from a global perspective, which improves the performance in complex scenes.

Furthermore, we verified the difference in semantic segmentation performance between the residual GSM module jointly constructed with SPConv and the vanilla mamba module in Table VII. Compared with vanilla mamba, GSM module improves 0.08% of OA, 0.89% of mIoU, and 1.02% of mAcc

TABLE VII
COMPARISON OF SEMANTIC SEGMENTATION RESULTS OF DIFFERENT MAMBA MODEL ON HRHD_HK DATASETS

Model	OA (%)	mIoU (%)	mAcc (%)
Vanilla Mamba	91.33	64.00	73.38
GSM	**91.41**	**64.89**	**74.40**

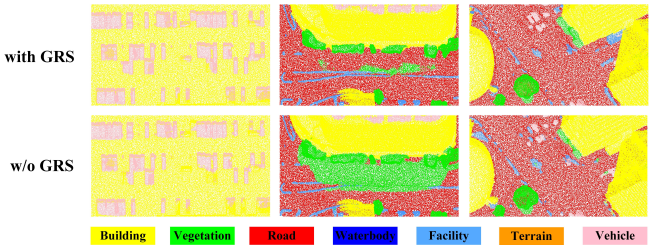

Fig. 12. Comparison of visualization results with or without CRS module on the HRHD_HK dataset.

on the HRHD_HK dataset. The experimental results showed that the joint sparse convolutional encoding and residual connection can effectively enhance the performance of mamba in global feature modeling.

3) Effect of CRS Module: Compared with the baseline, Model A3 improved 0.18% of OA and 1.60% of mIoU on the HRHD_HK dataset, 1.88% of OA and 5.09% of mIoU on the WHU_ALS dataset, and 0.31% of mIoU on the DALE dataset, respectively. The CRS module not only focuses on the spatial distribution of different classes in the scenes but also helps serialized orders understand the semantic context dependencies between objects. Compared with Model A4, GridPSFormer improves 0.18% of OA and 1.34% of mIoU on the HRHD_HK dataset, 0.14% of OA and 0.63% of mIoU on the WHU_ALS dataset, and 0.20% of mIoU on the DALE dataset, respectively. The CRS module ensures that each class is reasonably represented during serialized orders, which further improves the segmentation accuracy.

Additionally, we qualitatively analyzed the effect of using or not using the CRS module on the segmentation results of point clouds. As shown in Fig. 12, without the CRS module, objects like vegetation, vehicles, roads, and buildings present scattered structures and ambiguous segmentation boundaries. The CRS module helps objects maintain complete geometrical structures and clear segmentation boundaries. It strengthens the spatial correlation between roads and vegetation and reduces noise interference near buildings. This demonstrates that incorporating class-aware relational information can effectively improve the performance of serialized representations.

E. Network Structure Analysis

To better compare the differences between transformer and mamba block in local and global modeling, we constructed PSMamba, a method that implements a complete mamba architecture, which replaces the attention mechanism in Point Transformer v3 with vanilla mamba. Building on PSMamba, we then developed GridPSMamba by incorporating the three modules proposed in this article, following the same architecture design as GridPSFormer. As shown in Table VIII, compared with PS Mamba, GridPSMamba improved 4.20% of OA and 6.17% of mIoU on the HRHD_HK dataset, 10.39% of OA and 11.77% of mIoU on the WHU_ALS dataset, and 0.68% of OA and 5.57% of mIoU on the DALES dataset. The experimental results demonstrate that the 3-D SSO, GSM, and CRS modules have broad applicability and improve the segmentation performance of serialized networks.

TABLE VIII
COMPARISON OF SEMANTIC SEGMENTATION RESULTS OF DIFFERENT METHODS ON THREE DATASETS

Model	HRHD_HK		WHU_ALS		DALES	
	OA (%)	mIoU (%)	OA (%)	mIoU (%)	OA (%)	mIoU (%)
PSMamba	86.49	55.64	71.40	54.96	97.59	79.81
GridPSMamba	90.69	61.81	81.79	66.73	**98.27**	85.38
GridPSFormer	**91.41**	**64.89**	**82.99**	**68.17**	98.24	**85.50**

Compared with GridPSMamba, GridPSFormer improved 0.72% of OA and 3.08% of mIoU on the HRHD_HK dataset, 1.20% of OA and 1.44% of mIoU on the WHU_ALS dataset, and 0.12% of mIoU on the DALES dataset. The experimental results fully validate the effectiveness of the local attention mechanism in mining local geometric information and highlight the advantages of the GSM module for global feature extraction. Furthermore, the synergy between the local attention mechanism and the global serialization mamba module facilitates the comprehensive exploitation of serialized features and improves the performance of the model for semantic segmentation of point clouds in various densities and various heights scenes.

F. Grid Size Analysis

The grid size plays a crucial role in balancing the feature expressiveness and computational efficiency of the model. Considering the varying density and height of large-scale urban scenes, we evaluated the semantic segmentation performance and memory consumption of GridPSFormer under five grid sizes, that is, [0.100, 0.125, 0.150, 0.175, 0.200], as shown in Fig. 13. The same grid size was used during both training and inference phases, which were performed on a single A6000 GPU with batch size set to 2. Other training settings followed those described in Section IV-A. As the grid size increased, the GPU memory footprint gradually decreased, especially at 0.150 m. In terms of model performance, the best results in OA and mIoU metrics were achieved at a grid size of 0.150 m, while the best performance in mAcc was achieved at 0.175 m.

When the grid size is too small, the spatial coverage of spherical cropping becomes limited, which restricts the effectiveness of the SFCs to capture the 3-D structural information, and at the same time, significantly increases the computational complexity. Conversely, a larger grid size improves the

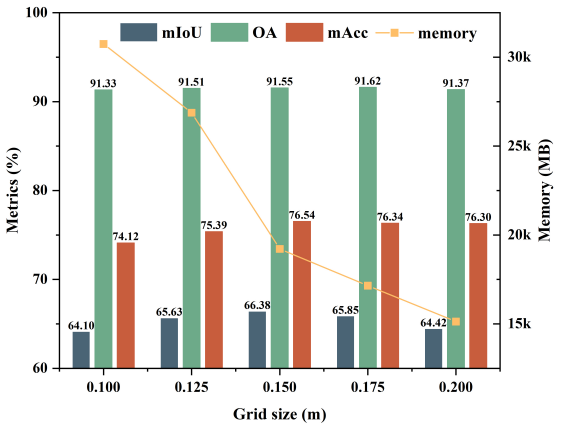

Fig. 13. Comparison of different grid size on the HRHD_HK dataset.

TABLE IX
COMPARISON RESULTS OF COMPUTATIONAL COMPLEXITY OF DIFFERENT METHODS ON HRHD_HK DATASET

Methods	Para.	Training		Inference	
		Lat. (ms)	Mem. (M)	Lat. (ms)	Mem. (M)
SPUNet	39.2M	71	1664	55	649
SPVCNN	21.8M	49	3229	38	1211
PT v3	46.2M	147	4376	138	754
PT v3*	46.2M	182	9903	246	752
PT v3**	46.2M	695	36043	612	752
PS Mamba	49.2M	251	8145	198	807
GridPSMamba	60.4M	279	9216	205	979
GridPSFormer	57.4M	184	5155	168	930

processing efficiency and expands the receptive field of spherical cropping, but it also tends to lose the key geometric details, which negatively impacts segmentation accuracy. Considering both segmentation accuracy and computational efficiency across scenes with varying density and height, we selected 0.150 m as the default grid size.

G. Complexity Analysis

For a fair comparison of the computational complexity of different methods, we performed single-epoch training and inference on a single A6000 with batch size set to 1. We used the same data augmentation to compute the parameters of the different methods, as well as the average latency and GPU memory usage for the training and inference phases.

As shown in Table IX, SPUNet and SPVCNN have fewer parameters compared to the other methods. However, the segmentation performance of these methods is poor due to the weakness of single sparse convolution block. The parameters of GridPSFormer were slightly higher than Point Transformer v3 and PS Mamba but lower than GridPSMamba due to the use of local window attention and GSM modules to extract local and global information, respectively.

In the training phase, GridPSFormer was lower than PS Mamba and GridPSMamba in terms of latency and GPU memory usage. In addition, GridPSFormer significantly reduced GPU memory consumption compared to PTv3* without the flash-accelerated attention at a patch size of 128, and PTv3** at a patch size of 512. In the inference phase, GridPSFormer was also lower than PS Mamba and GridPSMamba in latency and lower than SPVCNN and GridPSMamba in GPU memory usage. In summary, GridPSFormer significantly improved the accuracy of point cloud semantic segmentation with a lower cost of computational complexity.

V. CONCLUSION

We proposed a GridPSFormer for semantic segmentation of point clouds in density-varying and height-varying scenes. The point clouds were first gridded in GridPSFormer, which effectively reduces redundant information and improves computational efficiency. The 3-D SSO module maintained the spatial proximity of the point clouds and also provided a complete perception of the overall 3-D space by traversing the different orders. Then, GridPSFormer combined the serialized attention mechanism and GSM module to extract multiscale serialized features. Furthermore, the CRS module further provided semantic context information to improve the accuracy of semantic segmentation in various densities and heights scenes. In experiments on three datasets, HRHD_HK, WHU_ALS, and DALES, GridPSFormer achieved excellent performance in both segmentation accuracy and computational efficiency and outperforms GridPSMamba with the same structure. Compared with the baseline method, GridPSFormer improved HRHD_HK by 2.27%, WHU_ALS by 11.24%, and DALES by 0.63% in terms of mIoU, and in terms of OA by 0.58% for HRHD_HK and 7.96% for WHU_ALS. Experimental results showed that multiple 3-D SSOs enhance the comprehensive perception of 3-D space. The serialized mamba fully explored the serialized features with lower computational complexity. However, the proposed method does not perform best in all classes due to class imbalance in various densities and heights scenes. Owing to the efficiency of SFCs in 3-D point cloud scene understanding, we will further explore their potential in the field of multimodal data fusion, especially in the combination of point clouds and images, for more powerful semantic understanding and scene perception.

REFERENCES

[1] X. Han, C. Liu, Y. Zhou, K. Tan, Z. Dong, and B. Yang, "WHU-Urban3D: An urban scene LiDAR point cloud dataset for semantic instance segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 209, pp. 500–513, Mar. 2024, doi: 10.1016/j.isprsjprs.2024.02.007.

[2] Y. Li et al., "Deep learning for LiDAR point clouds in autonomous driving: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3412–3432, Aug. 2021.

[3] J. Wang et al., "AGRNav: Efficient and energy-saving autonomous navigation for air-ground robots in occlusion-prone environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 11133–11139.

[4] H. Zhao et al., "M2CS: A multimodal and campus-scapes dataset for dynamic SLAM and moving object perception," *J. Field Robot.*, vol. 42, no. 3, pp. 787–805, May 2025, doi: 10.1002/rob.22468.

[5] X. Y. Tian, T. Jiang, L. Yun, Y. Wang, Y. Wang, and H. Zhao, "Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 64318–64330.

[6] X. Sun, B. Guo, C. Li, N. Sun, Y. Wang, and Y. Yao, "Semantic segmentation and roof reconstruction of urban buildings based on LiDAR point clouds," *ISPRS Int. J. Geo-Information*, vol. 13, no. 1, p. 19, Jan. 2024, doi: 10.3390/ijgi13010019.

[7] P. Tian et al., "3-D semantic terrain reconstruction of monocular close-up images of Martian terrains," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4600716, doi: 10.1109/TGRS.2024.3378192.

[8] W. Gao, L. Nan, B. Boom, and H. Ledoux, "SUM: A benchmark dataset of semantic urban meshes," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 108–120, Sep. 2021, doi: 10.1016/j.isprsjprs.2021.07.008.

[9] C. Huang, S. Fang, H. Wu, Y. Wang, and Y. Yang, "Low-altitude intelligent transportation: System architecture, infrastructure, and key technologies," *J. Ind. Inf. Integr.*, vol. 42, Nov. 2024, Art. no. 100694.

[10] H. Li et al., "MVPNet: A multi-scale voxel-point adaptive fusion network for point cloud semantic segmentation in urban scenes," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, Aug. 2023, Art. no. 103391, doi: 10.1016/j.jag.2023.103391.

[11] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "RangeNet++: Fast and accurate LiDAR semantic segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4213–4220.

[12] J. Yang, C. Lee, P. Ahn, H. Lee, E. Yi, and J. Kim, "PBP-Net: Point projection and back-projection network for 3D point cloud segmentation," presented at the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct. 2020, pp. 8469–8475.

[13] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal ConvNets: Minkowski convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3070–3079.

[14] H. Tang et al., "Searching efficient 3D architectures with sparse point-voxel convolution," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, 2020, pp. 685–702.

[15] H. Zhou et al., "Cylinder3D: An effective 3D framework for driving-scene LiDAR semantic segmentation," 2020, *arXiv:2008.01550*.

[16] Spconv Contributors, "SpConv: Spatially sparse convolution library," 2022. [Online]. Available: https://github.com/traveller59/spconv

[17] 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.

[18] C. R. Qi, Y. Li, H. Su, and L. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108.

[19] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6410–6419.

[20] Q. Hu et al., "RandLA-Net: Efficient semantic segmentation of large-scale point clouds," Proc. presented at the IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2020.

[21] C. Liu et al., "Context-aware network for semantic segmentation toward large-scale point clouds in urban environments," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5703915, doi: 10.1109/TGRS.2022.3182776.

[22] Z. Luo, Z. Zeng, W. Tang, J. Wan, Z. Xie, and Y. Xu, "Dense dual-branch cross attention network for semantic segmentation of large-scale point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5700216, doi: 10.1109/TGRS.2023.3341894.

[23] 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.

[24] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16239–16248.

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

[26] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.

[27] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz, "Analysis of the clustering properties of the Hilbert space-filling curve," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 1, pp. 124–141, Jun. 2001.

[28] H. Sagan, *Space-Filling Curves*. Cham, Switzerland: Springer, 2012.

[29] P.-S. Wang, "OctFormer: Octree-based transformers for 3D point clouds," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–11, Aug. 2023, doi: 10.1145/3592131.

[30] X. Wu et al., "Point transformer v3: Simpler, faster, stronger," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 4840–4851.

[31] D. Li, J. Zhao, C. Chang, Z. Chen, and J. Du, "LGMamba: Large-scale ALS point cloud semantic segmentation with local and global state-space model," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, pp. 1–5, 2025, doi: 10.1109/LGRS.2024.3521395.

[32] J. A. Orenstein, "Spatial query processing in an object-oriented database system," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 1986, pp. 326–336.

[33] D. Hilbert, *Dritter Band: Analysis? Grundlagen Der Mathematik? Physik Verschiedenes: Nebst Einer Lebensgeschichte*. Springer-Verlag, 2013.

[34] J. Xu, R. Zhang, J. Dou, Y. Zhu, J. Sun, and S. Pu, "RPVNet: A deep and efficient range-point-voxel fusion network for LiDAR point cloud segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16004–16013.

[35] S. Fan, Q. Dong, F. Zhu, Y. Lv, P. Ye, and F.-Y. Wang, "SCF-Net: Learning spatial contextual features for large-scale point cloud segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14499–14508.

[36] H. Shuai, X. Xu, and Q. Liu, "Backward attentive fusing network with local aggregation classifier for 3D point cloud semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 4973–4984, 2021, doi: 10.1109/TIP.2021.3073660.

[37] Z. Zeng, Y. Xu, Z. Xie, W. Tang, J. Wan, and W. Wu, "Large-scale point cloud semantic segmentation via local perception and global descriptor vector," *Expert Syst. Appl.*, vol. 246, Jul. 2024, Art. no. 123269, doi: 10.1016/j.eswa.2024.123269.

[38] T. Han et al., "Point cloud semantic segmentation with adaptive spatial structure graph transformer," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 133, Sep. 2024, Art. no. 104105, doi: 10.1016/j.jag.2024.104105.

[39] S. Zhang, B. Wang, Y. Chen, S. Zhang, and W. Zhang, "Point and voxel cross perception with lightweight cosformer for large-scale point cloud semantic segmentation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 131, Jul. 2024, Art. no. 103951, doi: 10.1016/j.jag.2024.103951.

[40] D. Liang et al., "PointMamba: A simple state space model for point cloud analysis," 2024, *arXiv:2402.10739*.

[41] T. Zhang et al., "Point cloud mamba: Point cloud learning via state space model," 2024, *arXiv:2403.00762*.

[42] X. Han, Y. Tang, Z. Wang, and X. Li, "Mamba3D: Enhancing local features for 3D point cloud analysis via state space model," presented at the Proc. 32nd ACM Int. Conf. Multimedia, Oct. 2024.

[43] G. Zhang, L. Fan, C. He, Z. Lei, Z. Zhang, and L. Zhang, "Voxel mamba: Group-free state space models for point cloud based 3D object detection," 2024, *arXiv:2406.10700*.

[44] C. Faloutsos, "Multiattribute hashing using gray codes," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 1986, pp. 227–238.

[45] X. Xiang, L. Wang, W. Zong, and G. Li, "Extraction of local structure information of point clouds through space-filling curve for semantic segmentation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 114, Nov. 2022, Art. no. 103027, doi: 10.1016/j.jag.2022.103027.

[46] J. Chen, L. Yu, and W. Wang, "Hilbert space filling curve based scan-order for point cloud attribute compression," *IEEE Trans. Image Process.*, vol. 31, pp. 4609–4621, 2022, doi: 10.1109/TIP.2022.3186532.

[47] W. Chen, X. Zhu, G. Chen, and B. Yu, "Efficient point cloud analysis using Hilbert curve," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 730–747.

[48] Z. Li et al., "Pamba: Enhancing global interaction in point clouds via state space model," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 5, pp. 5092–5100.

[49] Y. Yang, T. Xun, K. Hao, B. Wei, and X.-S. Tang, "Grid mamba: Grid state space model for large-scale point cloud analysis," *Neurocomputing*, vol. 636, Jul. 2025, Art. no. 129985, doi: 10.1016/j.neucom.2025.129985.

[50] X. Jin et al., "UniMamba: Unified spatial-channel representation learning with group-efficient mamba for LiDAR-based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 1407–1417.

[51] H. Liu, M. Yao, X. Xiao, B. Zheng, and H. Cui, "MarsScapes and UDAFormer: A panorama dataset and a transformer-based unsupervised domain adaptation framework for Martian terrain segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2023, Art. no. 4600117, doi: 10.1109/TGRS.2023.3343109.

[52] M. Li, Y. Wu, A. G. O. Yeh, and F. Xue, "HRHD-HK: A benchmark dataset of high-rise and high-density urban scenes for 3D semantic segmentation of photogrammetric point clouds," in *Proc. IEEE Int. Conf. Image Process. Challenges Workshops (ICIPCW)*, Oct. 2023, pp. 3714–3718.

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

[54] 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.

Huchen Li received the B.S. degree from East China University of Technology, Nanchang, China, in 2021, and the M.Sc. degree from Nanjing University of Information Science and Technology, Nanjing, China, in 2024. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan, China.

His research interests include LiDAR point cloud semantic segmentation, multimodal fusion, and occupancy projections.

Wubiao Huang received the B.S. degree from Xiamen University of Technology, Xiamen, China, in 2020, and the M.Sc. degree in photogrammetry and remote sensing from Chang'an University, Xi'an, China, in 2023. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan, China.

His research interests include deep learning, geological hazard identification and analysis, remote sensing semantic segmentation, knowledge graphs, and multimodal large models.

Jiacheng Liu received the M.Sc. degree from Nanjing University of Information Science and Technology, Nanjing, China, in 2024. He is currently pursuing the Ph.D. degree with the Civil and Environmental Engineering, UNSW, Sydney, NSW, Australia.

His research interests include point cloud analysis and processing.

Ke Chen received the M.Sc. degree from Nanjing University of Information Science and Technology, Nanjing, China, in 2023, where he is currently pursuing the Ph.D. degree with the School of Remote Sensing and Geomatics Engineering.

His research interests include point cloud processing and intelligent interpretation.

Fei Deng received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Hubei, China, in 1999, 2002, and 2006, respectively.

He is currently a Professor with Wuhan University. His research interests include 3-D reconstruction, parametric modeling, and machine learning for remote sensing image and point cloud.