LGSSM: Local-to-global State Space Model for Serialized Point Cloud Semantic Segmentation

Hao Wu, Li Yan, Hucheng Li, Qimeng Li, Longze Zhu, Junjie Yuan, Hong Xie*

School of Geodesy and Geomatics, Hubei Luojia Laboratory, Wuhan University

Keywords: Serialization, Local-to-global, State space model, Semantic segmentation

Abstract

Point clouds have become essential data for describing real-world objects. Accurate and efficient 3D semantic segmentation plays a crucial role in environment understanding and scene reconstruction. However, current segmentation methods still face challenges from unordered data, high computational complexity, limited scene perception, and insufficient generalization. To address these issues, we propose a local-to-global semantic segmentation method based on a state-space model (LGSSM). Specifically, the proposed method uses three-dimensional serialization encoding to serialize point clouds along the x, y, and z directions, effectively addressing the inherent disorder of point clouds and enhancing spatial representation. Then, the local state space model extracts fine-grained local geometric structural information and the global state space model captures the overall scene representation, improving the modeling ability for both short and long distances. Finally, the serialized context aggregation module is utilized to fuse contextual features to promote spatial semantic consistency. Extensive experiments conducted on ScanNet, ScanNet200, and S3DIS demonstrate that our model achieves state-of-the-art segmentation accuracy compared with other existing methods.

1. Introduction

In three-dimensional space, LiDAR can efficiently obtain high-precision geometric information of observed objects. Semantic segmentation aims to divide unordered and unstructured point clouds into regions with semantic or geometric consistency, making it a crucial step toward comprehensive scene understanding. It has been widely applied in various fields such as autonomous driving (Chen et al., 2024), robotics (Wu et al., 2025), resource survey (Wilkes et al., 2023), and 3D reconstruction (Yan et al., 2025).

Unlike 2D images, point clouds lack a regular topological structure and fixed pixel adjacency relationships. In addition, point clouds suffer from uneven distribution, density variation, noise, and occlusion. These data characteristics pose significant challenges for semantic segmentation. Early point cloud segmentation methods mainly relied on traditional geometric features and clustering algorithms, such as normal vector analysis (Rusu et al., 2008), region growing (Rabbani et al., 2006), RANSAC-based model fitting (Fischler and Bolles, 1981), and spectral clustering (Golovinskiy and Funkhouser, 2009). Although these handcrafted feature methods can achieve semantic segmentation in specific scenarios, they are sensitive to parameters and suffer from limited generalization capability.

Deep learning has demonstrated outstanding performance in point cloud semantic segmentation. Projection-based (Cortinhal et al., 2020) methods benefit from powerful models in the 2D image domain, allowing projected point clouds to be processed through image segmentation networks for semantic segmentation. However, these methods inevitably lose the inherent geometric properties of point clouds. Voxel-based (Zhou et al., 2020) methods represent point clouds using 3D grids and perform efficient feature extraction through sparse 3D convolutions. But voxelization fails to preserve local structural information. PointNet++ (Qi et al., 2017) is a pioneering work that enables direct feature extraction from point clouds. Point-based methods further enhance global feature extraction and contextual information fusion. However, most of these approaches rely on explicit neighborhood searches and attention mechanisms, resulting in high computational complexity. Meanwhile, they struggle to capture long-range spatial relationships, which limits their generalization across different scenes.

Mamba (Gu and Dao, 2024) achieves global modeling with linear complexity by compressing contextual information through a state space model (SSM), demonstrating outstanding performance in both natural language processing (Zhao et al., 2025) and computer vision (Wang et al., 2025) tasks. However, it lacks local structural modeling abilities, making it difficult to capture fine-grained geometric details. Moreover, state space models are effective primarily for structured data, and the inherent disorder of point clouds poses significant challenges. In addition, a single serialization method reduces the diversity of 3D spatial information and limits broader applicability. Furthermore, depth analysis and comprehensive understanding of state space models remain insufficient.

To address the above problems, we propose a local-to-global state space model for serialized point cloud semantic segmentation (LGSSM). First, to address the challenge of limited spatial perception in single serialization methods, we propose a three-dimensional serialization encoding module based on z-order space-filling curves to enhance spatial diversity. Then, we introduce a local-to-global state space model. This model constructs a bidirectional SSM based on the serialization encoding module, achieving global scene perception through forward and backward propagation. For the local module, the bidirectional serialized encodings are divided into blocks, from which the state space model extracts fine-grained local geometric features. Feature aggregation is used to gain part-to-whole scene understanding. Finally, the serialized context aggregation module achieves multi-level and multi-scale scene feature fusion to enhance semantic consistency.

Our main contributions are as follows:

- We design a 3D space serialization encoding method that utilizes shuffled orders to enable the model to comprehensively perceive the three-dimensional space.

- The combination of the local-to-global and bidirectional state space model simultaneously obtains both part and whole information, achieving fine-grained local geometry perception and long-distance modeling.

- We introduce the serialized context aggregation module at different stages to effectively fuse contextual information and enhance semantic consistency.

- The LGSSM model was constructed and evaluated on multiple point cloud datasets. Experimental results demonstrate that our method achieves state-of-the-art semantic segmentation accuracy.

2. Related Work

2.1 Point cloud semantic segmentation

Learning-based point cloud semantic segmentation achieves higher accuracy and stronger generalization compared to traditional methods. Existing methods can be divided into projection-based methods, voxel-based methods, and point-based methods. Projection-based methods transform irregular point clouds into structured 2D representations, enabling scene segmentation through convolutional neural networks. RangeViT (Ando et al., 2023) projects point clouds into range images and performs global feature modeling through combining 2D convolutions with vision transformers. SPCNet (Zheng et al., 2024) leverages spherical projection and sparse convolution within a spherical frustum to obtain local and global contextual features. However, projecting point clouds onto planes leads to geometric information loss and feature distortion. It is also highly sensitive to sensor viewpoints and parameters, limiting generalization ability across different scenes.

Voxel-based methods divide point clouds into regular voxel grids and use 3D convolutions to extract object features. Cylinder3D (Zhu et al., 2021) leverages cylindrical voxel space and asymmetric 3D convolution kernels to obtain local geometric details and global semantic features. SVASeg (Zhao et al., 2022) constructs hash tables based on voxels and captures contextual information through sparse multi-head attention mechanisms. However, voxel-based methods can easily cause the loss of fine-grained local geometric details and boundary information. In addition, high-resolution voxelization can result in significant computational and memory overhead, making it challenging to balance segmentation accuracy and overall efficiency.

Point-based methods perform semantic segmentation directly on original point clouds, avoiding discretization loss and preserving the inherent spatial structure and fine-grained information. LACV-Net (Zeng et al., 2024) introduces a local perception module and a global descriptive vector to simultaneously capture the fine-grained geometric features and overall semantic information. Swin3D (Yang et al., 2025a) proposes a pre-trained 3D transformer network that obtains local and global information through the hierarchical sliding-window attention mechanism. Although point-based methods can effectively capture fine-grained local geometric features, they have high computational complexity and their segmentation accuracy is greatly affected by sampling density and noise.

2.2 3D space serialization

The inherent unordered nature of point clouds makes it challenging to achieve fast and accurate scene semantic segmentation. The serialization of 3D spatial data aims to transform irregular and unordered point clouds into structured representations that can be efficiently processed by neural networks. HSFC-PCAC (Chen et al., 2022) designs a scan order based on the hilbert curve, ensuring that spatially adjacent points remain close to each other. SFC-Net (Xiang et al., 2022) introduces relative angular features and space-filling curves to obtain geometric relationships between neighboring points and overall structural information. FlatFormer (Liu et al., 2023) proposes a point cloud Transformer network that reduces redundant computations in overlapping regions and enhances feature interaction efficiency through a flattened window attention mechanism. OctFormer (Wang, 2023) leverages an octree structure to efficiently capture multi-scale spatial relationships in a hierarchical manner. Point Transformer v3 (Wu et al., 2024) uses shuffle order and linear attention to reduce computational complexity and improve semantic segmentation accuracy. However, a single serialization method struggles to effectively encode point clouds, leading to limited perception of spatial structures and fine-grained details.

2.3 State Space Model

State Space Model (SSM) exhibits strong global modeling abilities and computational efficiency in natural language processing and sequence modeling through combining linear recurrence and convolution. S5 (Smith et al., 2022) uses explicitly parameterized convolution kernels to extract long-range features and improve training stability and inference speed. VMamba (Liu et al., 2024) introduces a bidirectional scanning mechanism in the 2D selective scanning module, achieving linear-complexity context modeling. PointMamba (Liang et al., 2024) serializes irregular point clouds and performs global modeling and efficient feature propagation through the selective state space model. Point Cloud Mamba (Zhang et al., 2025) proposes a consistent traverse serialization method that efficiently serializes 3D point clouds into one-dimensional inputs, ensuring that spatially adjacent points remain close within the sequence. In addition, it combines point prompts with coordinate-mapped positional encoding to enhance spatial awareness. Grid Mamba (Yang et al., 2025b) integrates grid multi-view scanning, grid sparsity pooling, and the grid mamba module to achieve efficient and high-precision point cloud semantic segmentation. Since SSM tends to extract features along the sequence dimension, it lacks strong local interaction capabilities like convolution or attention mechanisms, making it difficult to fully capture fine-grained geometric structures.

3. Method

The overall architecture of LGSSM is illustrated in Fig. 1. Point cloud serializations in three different directions are encoded through 3D serialization encoding, effectively overcoming the inherent unordered characteristic. Shuffle order is used to randomly select the serialization method, enhancing spatial perception. Feature initialization is achieved through the embedding module (Section 3.1). Then, based on the U-Net architecture, LGamba extracts both local and global multi-level features, enhancing the perception of fine-grained details and improving overall scene understanding (Section 3.2). Finally, serialized context aggregation is used for feature fusion to enhance semantic consistency. Fast and accurate scene semantic

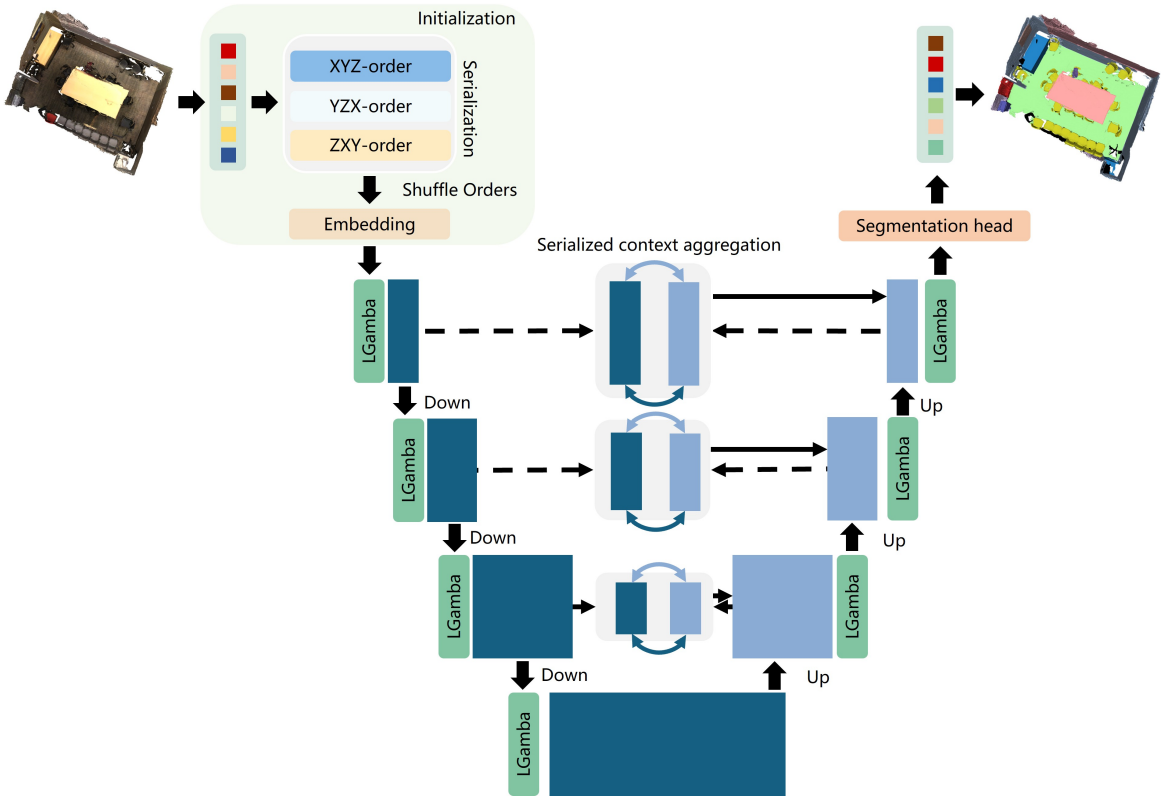

Figure 1. The overall pipeline of LGSSM. First, the three-dimensional z-order space-filling curve serializes the point cloud to address the inherent disorder problem. Shuffle orders are used to enhance model stability. The initial features are extracted through the embedding module. Then, the multi-level LGamba and serialized sampling extract multi-scale features based on the U-Net architecture. Next, serialized context aggregation is applied to fuse contextual scene features. Finally, accurate and efficient scene semantic segmentation is achieved.

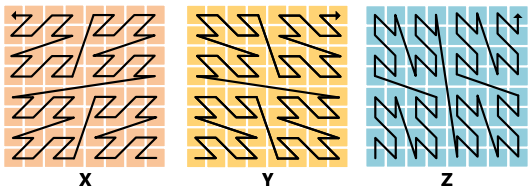

Figure 2. 3D space serialization encoding.

segmentation is achieved through the segmentation head (Section 3.3).

3.1 3D space serialization encoding

The z-order space-filling curve (Morton, 1966) helps overcome point cloud disorder to achieve efficient serialized encoding. However, a single 3D spatial encoding method struggles to adapt to the complex directionality and structure, which results in incomplete spatial structure and serialization bias.

To address this problem, we implement spatial scene serialization along the x, y, and z dimensions based on the Z-order curve, as illustrated in Fig. 2. By changing the traversal order of the coordinate axes in 3D space and combining it with the z-order curve, three different serialization encodings can be generated: XYZ-order, YZX-order and ZXY-order. Through serialization encoding in three different spatial directions, relative position and geometric structure information can be effectively obtained.

$$\mathrm{SSE}(p, b, e) = (b \ll k) | \varphi_e^{-1}(\Phi(p)/g) \tag{1}$$

Where p denotes the point cloud, b represents the batch index, e refers to the z-order serialization encoding, k denotes the number of offset bits, $\ll$ indicates a left shift operation, φ_e^{-1} represents the inverse mapping, Φ is the coordinates with arbitrary permutation along spatial dimensions and g is the grid size. To enhance comprehensive scene perception and generalization ability, we adopt shuffle order (Wu et al., 2024) to randomly select from three different serialization encoding methods. It can reduce the reliance on a single structured method and enhance adaptability and robustness through multi-scale and multi-directional serialization encoding.

$$\mathrm{Embedding} = \mathrm{GELU}(\mathrm{BN}(\mathrm{Subconv}(x))) \tag{2}$$

Finally, sparse convolution is used to initialize the serialized point cloud features.

3.2 Local-to-global state space model

Mamba uses the state space model to achieve long-range scene modeling with linear complexity. However, the local geometric structural details are often neglected, making it difficult to obtain fine-grained information. A single serialization method and unidirectional propagation limit its ability to comprehensively perceive the scene. To this end, we propose a local-to-global

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-169-2026 | © Author(s) 2026. CC BY 4.0 License.

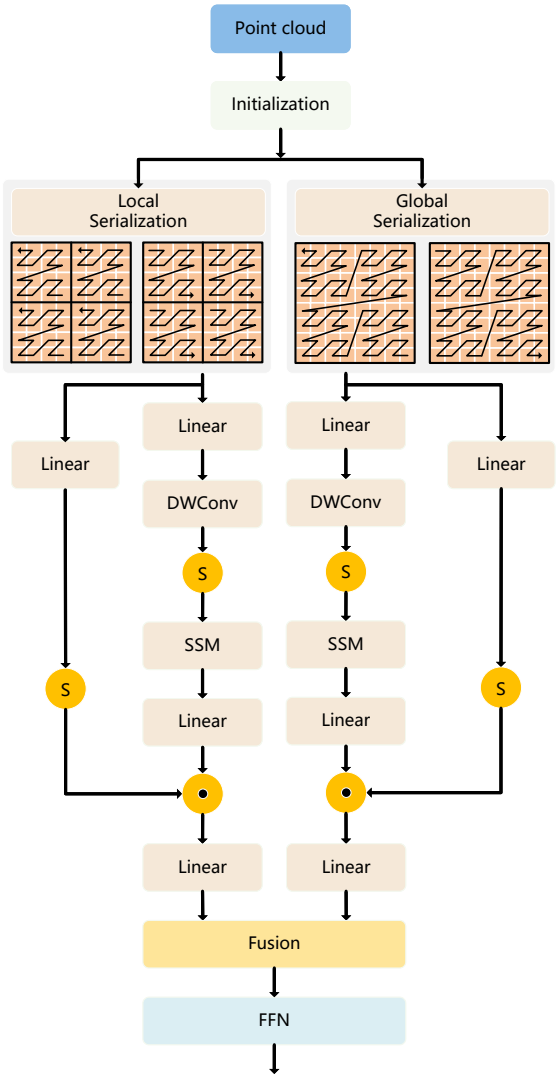

Figure 3. Local-to-global state space model.

bidirectional state space model (LGamba) that strengthens the information propagation and filtering mechanisms, enabling the acquisition of both local fine-grained details and comprehensive scene understanding, as illustrated in Figure 3.

$$\begin{aligned} F'_m &= \text{Silu}(\text{DWConv}(\text{Linear}(F'))) \\ F''_m &= \text{LN}(\text{SSM}(F'_m)) \\ \text{Mamba}(F) &= \text{Linear}(F''_m \cdot \text{Silu}(\text{Linear}(F'))) \end{aligned} \tag{3}$$

Here, F' represents the initialized serialized features, DWConv denotes depthwise convolution, and SiLU refers to the activation function. In the global state space model, a bidirectional state space structure is constructed based on 3D spatial serialization encoding through backward propagation to refine global information and filter irrelevant information. Bidirectional Mamba performs long-range modeling in different directions to obtain relevant dependencies, enhancing the entire scene understanding.

$$F_g = F' + \text{BiMamba}(F') \tag{4}$$

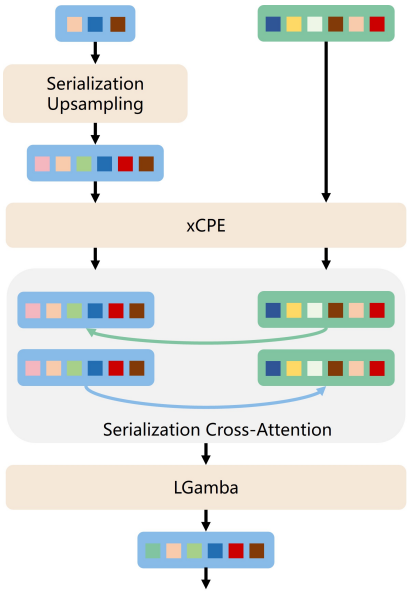

Figure 4. Serialized context aggregation module.

Where BiMamba represents bidirectional Mamba. In the local state space model, the point cloud is divided into $F'_n = \{F'_i, i = 1, 2, 3, \ldots, \frac{N}{S}\}$ non-overlapping small blocks based on serialized encoding. Then, bidirectional Mamba extracts the features of each small block to obtain local geometric structural information and restores the features to the original shape.

$$\begin{aligned} F'_l &= \text{Block}(F', S) \\ F''_l &= F'_l + \text{BiMamba}(F'_l) \\ F_l &= \text{Restore}(F''_l, L) \end{aligned} \tag{5}$$

Here, Block denotes the dividing operation, and Restore represents the restoration to the original shape. A feed-forward network composed of SubConv, batch normalization, and GELU is used to extract high-level semantic information.

In LGSSM, local and global serialization encoding enhances spatial diversity and stability. The local state space model is used to capture local relative geometric relationships. The global state space model achieves efficient long-range modeling with linear complexity. Local and global features are fused to enhance their complementarity. Finally, a feed-forward network enables comprehensive spatial perception of the entire scene.

3.3 Serialized context aggregation module

Multi-level feature fusion can effectively enhance comprehensive scene perception and understanding. To achieve this, we design a serialized context aggregation module to integrate contextual information, as shown in Figure 4.

Given the downsampled and upsampled point clouds (P_i, P_{i-1}) and corresponding features (F_i, F_{i-1}), the downsampled point cloud aligns spatial and feature dimensions through serialized uppooling. We use conditional positional embedding (Wu et al., 2024) to improve its awareness of spatial positions and geometric structures. Then, a serialized cross-attention mechanism is applied to fuse contextual features, achieving cross-layer semantic alignment and relational perception.

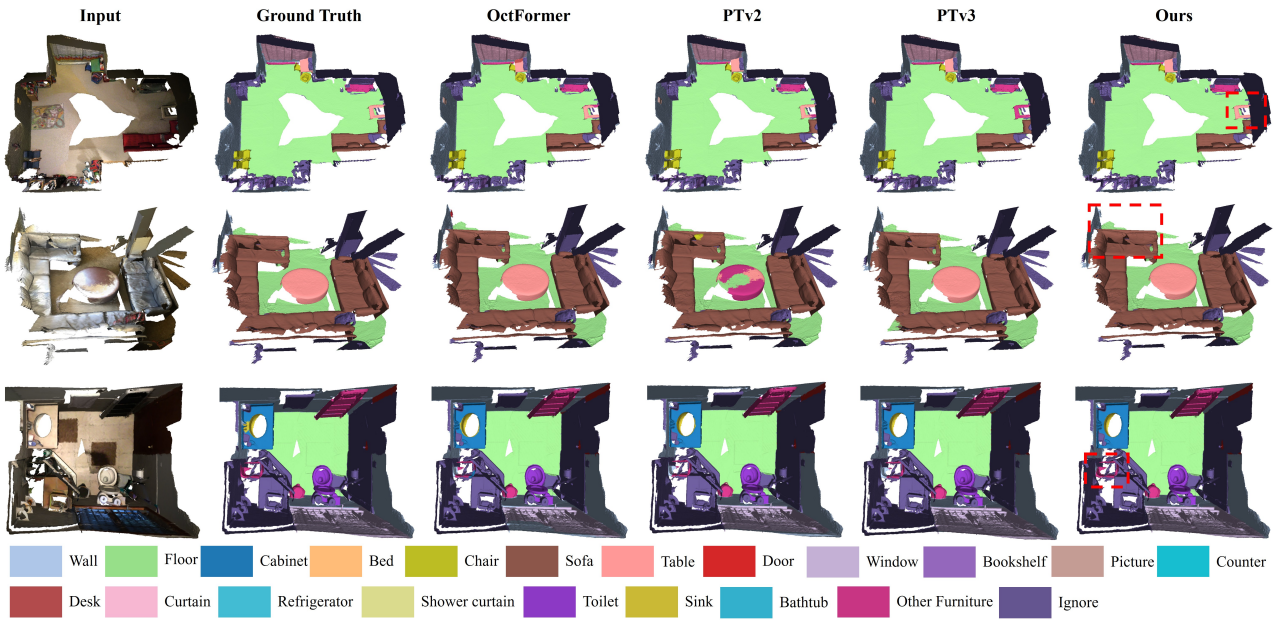

Figure 5. Visualization of Indoor Scannet Dataset.

Table 1. Comparison of semantic segmentation performance of different methods on the ScanNet dataset. Bold indicates the best performance.

Method	Reference	Backbone	mIoU
PointNet++	NeurIPS 17	CNN	53.5
PointConv	CVPR 19	CNN	61.0
MinkowskiNet	CVPR 19	CNN	72.2
KPConv	ICCV 19	CNN	69.2
PointNeXt	NeurIPS 22	CNN	71.5
PointMetaBase	CVPR 23	CNN	72.8
OA-CNNs	CVPR 24	CNN	76.1
Point Transformer	ICCV 21	Transformer	70.6
StratifiedFormer	CVPR 22	Transformer	74.3
Point Transformer v2	NeurIPS 22	Transformer	75.4
OctFormer	TOG 23	Transformer	75.7
Point Transformer v3	CVPR 24	Transformer	77.5
Grid Mamba	NC 25	Mamba	76.8
Ours		Mamba	**78.0**

$$\hat{F}_{i-1} = \text{FlashAttention}\left(\frac{(\bar{F}_{i-1}W_Q)(F_iW_K)^\top}{\sqrt{d}}, F_iW_V\right)W_O \tag{6}$$

Here, $\hat{F}_{i-1}$ represents the fused features. $\bar{F}_{i-1}$ denotes the serialized upsampling features. W_Q, W_K, W_V and W_O are projection matrices. The contextual features are aggregated to enhance semantic consistency. Finally, LGamba is used to further extract high-level semantic features, achieving fast and accurate point cloud semantic segmentation.

4. Experiment

4.1 Datasets

We evaluate the semantic segmentation performance of different methods on three datasets: ScanNet (Dai et al., 2017), ScanNet200 (Dai et al., 2017), and S3DIS (Armeni et al., 2016). ScanNet uses a structured-light sensor to capture 1513 indoor scenes covering 707 unique spaces. It reconstructs scenes using RGB-D data, with each scene containing approximately 150 thousand points on average. The dataset annotates 20 semantic categories and is divided into a training set of 1201 scenes and a validation set of 312 scenes. ScanNet200 uses the same data as ScanNet but expands the semantic categories to 200, enabling finer-grained classification and increasing the difficulty of semantic segmentation. S3DIS is collected from 6 buildings at Stanford University, comprising 271 rooms with a total area exceeding 6000 square meters and approximately 273 million points. Each point contains XYZ coordinates, RGB color, and semantic labels from 13 categories. The dataset covers various indoor scenes, including offices, conference rooms, and corridors.

4.2 Experimental Setting

LGSSM training and inference are conducted on an NVIDIA RTX 4090 GPU. The batch size for our model is set to 6. The AdamW optimizer is used with an initial learning rate of 0.0001 and a weight decay of 0.001. The model is trained for 100 epochs using the data augmentation strategy (Yang et al., 2025b). In the encoder stage, the number of LGamba blocks is set to [1, 2, 4, 1], while in the decoder stage, it is set to [1, 1, 1, 1]. The local patch size is configured as [1024, 1024, 1024, 1024]. The evaluation metrics include overall accuracy (OA), mean intersection-over-union (mIoU), and per-class IoU. We compare with other existing methods as follows: PointNet++ (Qi et al., 2017), PointConv (Wu et al., 2019), MinkowskiNet (Choy et al., 2019), KPConv (Thomas et al., 2019), PointNeXt (Qian et al., 2022), PointMetaBase (Lin et al., 2023), OA-CNNs (Peng et al., 2024), Point Transformer (Zhao et al., 2021), StratifiedFormer (Lai et al., 2022), Point Transformer v2 (Wu et al., 2022), OctFormer (Wang, 2023), Point Transformer v3 (Wu et al., 2024) Grid Mamba (Yang et al., 2025b), SupCon (Khosla et al., 2020), CSC-Pretrain (Hou et al., 2021), LGround (Rozenberszki et al., 2022) and Swin3D (Yang et al., 2025a).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-169-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 2. Comparison of detailed semantic segmentation results of different methods on ScanNet. Bold indicates the best performance.

Model	wall	floor	cabinet	bed	chair	sofa	table	door	window	book	picture	counter	desk	curtain	refrigerator	shower	toilet	sink	bathtub	other
PointNet++	85.97	95.79	71.12	83.38	92.06	83.97	75.69	**79.19**	70.41	82.90	37.59	69.18	67.02	78.21	68.35	65.48	92.33	68.44	88.26	60.37
OctFormer	86.30	95.89	67.61	84.69	**92.59**	**85.70**	75.80	69.16	67.52	82.28	33.55	69.38	66.48	78.35	68.15	71.08	91.12	68.72	87.74	58.63
Grid Mamba	87.28	95.62	70.36	**85.49**	91.46	81.00	77.03	72.03	70.83	83.57	33.02	67.98	70.27	79.75	69.99	**79.27**	**96.15**	**72.28**	88.38	65.59
Ours	**87.72**	**95.91**	**73.73**	82.64	92.25	85.58	**79.11**	70.70	**72.20**	**84.04**	**38.83**	**71.15**	**75.25**	**80.91**	**76.00**	73.08	94.12	71.69	**91.08**	**65.75**

Table 3. Comparison of semantic segmentation performance of different methods on the ScanNet200 dataset. Bold indicates the best performance.

Method	Reference	Backbone	mIoU
MinkowskiNet	CVPR 19	CNN	25.3
SupCon	NeurIPS 20	CNN	26.0
CSC-Pretrain	CVPR 21	CNN	24.9
LGround	ECCV 22	CNN	27.2
OA-CNNs	CVPR 24	CNN	33.3
Point Transformer v2	NeurIPS 22	Transformer	30.2
OctFormer	TOG 23	Transformer	32.6
Point Transformer v3	CVPR 24	Transformer	35.2
Grid Mamba	NC 25	Mamba	35.6
Ours		Mamba	**35.7**

Table 4. Comparison of semantic segmentation performance of different methods on the S3DIS dataset. Bold indicates the best performance.

Method	Reference	Backbone	6-Fold
MinkowskiNet	CVPR 19	CNN	65.4
PointNeXt	NeurIPS 22	CNN	74.9
Swin3D	CVM 25	Transformer	76.9
Point Transformer v1	ICCV 21	Transformer	65.4
Point Transformer v2	NeurIPS 22	Transformer	73.5
Point Transformer v3	CVPR 24	Transformer	77.7
Grid Mamba	NC 25	Mamba	77.9
Ours		Mamba	**78.1**

Table 5. Ablation study of designed modules.

3D space serialization encoding	✓	✓	✓	✓
Local-to-global state space model		✓		✓
Serialized context aggregation			✓	✓
mIoU	76.6	77.6	76.9	**78.0**

4.3 Experimental Results

We conduct a quantitative evaluation of our Mamba-based model on the ScanNet dataset, with detailed results reported in Table 1. Compared to CNN-based methods such as PointNeXt and PointMetaBase, our method improves mIoU by 6.5% and 5.2%, respectively, and further surpasses OA-CNNs by 1.9% mIoU. Relative to Transformer-based approaches, including StratifiedFormer, PTv2, and OctFormer, our model achieves gains of 3.7%, 2.6%, and 2.3% mIoU, respectively, and still brings a 0.5% improvement over the latest PTv3. In addition, our approach outperforms the Mamba-based Grid Mamba by 1.2% mIoU. These results demonstrate the effectiveness and highlight the superior capability in segmenting complex multi-type objects on the ScanNet dataset. We compare our method with other models, as shown in Fig. 5. It can be seen that our method achieves finer segmentation and clearer boundaries, such as tables, sofas, and other furniture. This is mainly attributed to the global-to-local state space model and the serialized context aggregation method, which fully exploit scene information to enhance the feature differences among different objects, thereby improving semantic segmentation accuracy.

The detailed comparison of semantic segmentation results on the ScanNet dataset is presented in Table 2. Overall, our method achieves the best mIoU among all compared models. Relative to PointNet++, our model shows noticeable improvements on some classes. For structural objects such as wall, window, cabinet, table, desk, and curtain, our method achieves higher IoU scores, while maintaining competitive performance on large planar regions like floor and bed. Compared with OctFormer, our method also produces clear gains on many categories, such as cabinet, table, window, picture, counter, desk, curtain, refrigerator, sink, bathtub, and other, enhancing structural and object-level semantic modeling ability. Compared to the Mamba-based Grid Mamba, our method further improves several important categories, for example, cabinet, table, window, book, picture, counter, desk, curtain, refrigerator, and bathtub, while also maintaining strong performance on most other classes. These improvements indicate that our model can better exploit spatial context and obtain fine-grained geometric details, especially for mid-sized furniture and cluttered regions. Overall, the detailed per-class results show that our method not only improves global mIoU, but also offers more balanced and robust performance across categories—from large structural regions to small or structurally complex objects. This indicates that our model provides consistently stronger segmentation ability in indoor scenes.

Table 3 presents the quantitative comparison on the ScanNet200 benchmark. Among CNN-based approaches, OA-CNNs achieves the strongest result with 33.3% mIoU. Transformer-based methods such as Point Transformer v2 and OctFormer achieve 30.2% and 32.6% mIoU, respectively, while the latest Point Transformer v3 attains 35.2% mIoU. The Mamba-based Grid Mamba slightly increases the score to 35.6%, indicating that state space modeling is competitive with Transformer baselines. In contrast, our method improves the performance to 35.7% mIoU. Compared with Point Transformer v3, our model achieves a +0.5% mIoU improvement. It also surpasses Grid Mamba by +0.1% mIoU. These results demonstrate that our proposed method can more effectively exploit spatial information, achieving higher segmentation accuracy on the challenging ScanNet200 dataset.

A comprehensive comparison between our model and existing methods on the S3DIS dataset is presented in Table 4. Overall, our method achieves the best performance among all compared models. Compared with CNN backbones such as MinkowskiNet and PointNeXt, our method obtains higher mIoU, demonstrating stronger ability in capturing complex indoor geometry and diverse semantic categories. For Transformer-based methods, our model outperforms

Table 6. Efficiency evaluation of different models on the ScanNet dataset.

Methods	Params	Training		Inference	
		Latency	Memory	Latency	Memory
MinkowskiNet	37.9 M	267 ms	4.9 G	90 ms	4.7 G
OctFormer	44.0 M	264 ms	12.9 G	86 ms	12.5 G
Point Transformer v1	7.8 M	904 ms	13.4 G	1450 ms	7.2 G
Point Transformer v2	12.8 M	312 ms	13.4 G	191 ms	18.2 G
Point Transformer v3	46.2 M	151 ms	6.8 G	61 ms	5.2 G
Grid Mamba	49.9 M	282 ms	11.1 G	161 ms	4.1 G
Ours	44.3 M	174 ms	10.4 G	67 ms	5.5 G

Swin3D and Point Transformer, indicating that the proposed method can better exploit both long-range dependencies and fine-grained local structures than traditional self-attention. Furthermore, compared to the Mamba-based Grid Mamba, our method achieves a higher 6-fold mIoU, showing that the improvements within the state-space modeling paradigm effectively enhance feature representation and optimization. These results suggest that our method adapts well to the diversity and scene variations in the S3DIS dataset.

4.4 Ablation Studies

We conduct ablation studies on the proposed modules using the ScanNet dataset to verify their effectiveness. As reported in Table 5, each individual module consistently improves semantic segmentation accuracy. When all modules are combined, the model attains its best overall performance, demonstrating the complementary benefits of each component.

Table 6 presents the number of parameters, training and inference time, and memory consumption of different methods on the ScanNet dataset. Compared with Point Transformer v3, although our method increases the training and inference time and memory consumption, it reduces the number of parameters and improves semantic segmentation accuracy. In contrast to Grid Mamba, which also adopts the Mamba architecture, our method reduces the time required for training and inference, improving scene perception efficiency.

5. Conclusion

In this paper, we propose a local-to-global state space model (LGSSM) for serialized point cloud semantic segmentation. The three-dimensional serialization encoding module based on z-order space-filling curves enhances spatial diversity and effectively handles the inherent unordered characteristics of point clouds. The local-to-global state space modeling captures fine-grained local geometric structures and coherent global scene information. The serialized context aggregation module fuses multi-level and multi-scale features, enhancing spatial semantic consistency. Extensive experiments on ScanNet, ScanNet200, and S3DIS demonstrate the superiority of the proposed architecture. In the future, we will integrate cross-modal information such as RGB images and text into the state space model and use domain generalization and domain adaptation methods to accommodate distributional discrepancies across diverse environments, ultimately achieving robust open-world 3D scene perception.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2025YFB3910302); Zhejiang Province “Vanguard” and “Geese Leading” Research and Development Plan (2025C01073); the National Natural Science Foundation of China (Grants·42371451); the National Natural Science Foundation of China (Grants·42394061); the Natural Science Foundation of Wuhan (No.2024040701010028).

References

Ando, A., Gidaris, S., Bursuc, A., Puy, G., Boulch, A., Marlet, R., 2023. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5240–5250.

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S., 2016. 3d semantic parsing of large-scale indoor spaces. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1534–1543.

Chen, J., Yu, L., Wang, W., 2022. Hilbert space filling curve based scan-order for point cloud attribute compression. *IEEE Transactions on Image Processing*, 31, 4609–4621.

Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., Li, H., 2024. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Choy, C., Gwak, J., Savarese, S., 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3075–3084.

Cortinhal, T., Tzelepis, G., Erdal Aksoy, E., 2020. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. *International Symposium on Visual Computing*, 207–222.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2432–2443.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

Golovinskiy, A., Funkhouser, T., 2009. Min-cut based segmentation of point clouds. *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 39–46.

Gu, A., Dao, T., 2024. Mamba: Linear-time sequence modeling with selective state spaces. *The First Conference on Language Modeling (COLM)*.

Hou, J., Graham, B., Nießner, M., Xie, S., 2021. Exploring data-efficient 3d scene understanding with contrastive scene contexts. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15587–15597.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. *The Advances in Neural Information Processing Systems (NeurIPS)*, 18661–18673.

Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., Qi, X., Jia, J., 2022. Stratified transformer for 3d point cloud segmentation. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8500–8509.

Liang, D., Zhou, X., Xu, W., Zhu, X., Zou, Z., Ye, X., Tan, X., Bai, X., 2024. Pointmamba: A simple state space model for point cloud analysis. *The Advances in Neural Information Processing Systems (NeurIPS)*, 32653–32677.

Lin, H., Zheng, X., Li, L., Chao, F., Wang, S., Wang, Y., Tian, Y., Ji, R., 2023. Meta architecture for point cloud analysis. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17682–17691.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y., 2024. Vmamba: Visual state space model. *The Advances in Neural Information Processing Systems (NeurIPS)*, 103031–103063.

Liu, Z., Yang, X., Tang, H., Yang, S., Han, S., 2023. Flatformer: Flattened window attention for efficient point cloud transformer. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1200–1211.

Morton, G. M., 1966. *A computer oriented geodetic data base and a new technique in file sequencing*. International Business Machines Company.

Peng, B., Wu, X., Jiang, L., Chen, Y., Zhao, H., Tian, Z., Jia, J., 2024. Oa-cnns: Omni-adaptive sparse cnns for 3d semantic segmentation. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21305–21315.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *The Advances in Neural Information Processing Systems (NeurIPS)*.

Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., Ghanem, B., 2022. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *The Advances in Neural Information Processing Systems (NeurIPS)*, 23192–23204.

Rabbani, T., Van Den Heuvel, F., Vosselmann, G., 2006. Segmentation of point clouds using smoothness constraint. *International archives of photogrammetry, remote sensing and spatial information sciences*, 36(5), 248–253.

Rozenberszki, D., Litany, O., Dai, A., 2022. Language-grounded indoor 3d semantic segmentation in the wild. *The European Conference on Computer Vision (ECCV)*, 125–141.

Rusu, R. B., Marton, Z. C., Blodow, N., Beetz, M., 2008. Persistent point feature histograms for 3d point clouds. *Intelligent Autonomous Systems*, 119–128.

Smith, J. T., Warrington, A., Linderman, S. W., 2022. Simplified state space layers for sequence modeling. *arXiv preprint: 2208.04933*.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L. J., 2019. Kpconv: Flexible and deformable convolution for point clouds. *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 6411–6420.

Wang, F., Wang, J., Ren, S., Wei, G., Mei, J., Shao, W., Zhou, Y., Yuille, A., Xie, C., 2025. Mamba-reg: Vision mamba also needs registers. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14944–14953.

Wang, P.-S., 2023. Octformer: Octree-based transformers for 3d point clouds. *ACM Transactions on Graphics*, 42(4), 1–11.

Wilkes, P., Disney, M., Armston, J., Bartholomeus, H., Bentley, L., Brede, B., Burt, A., Calders, K., Chavana-Bryant, C., Clewley, D. et al., 2023. TLS2trees: A scalable tree segmentation pipeline for TLS data. *Methods in Ecology and Evolution*, 14(12), 3083–3099.

Wu, W., Chen, C., Yang, B., Zou, X., Liang, F., Xu, Y., He, X., 2025. DALI-SLAM: Degeneracy-aware LiDAR-inertial SLAM with novel distortion correction and accurate multi-constraint pose graph optimization. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221, 92–108.

Wu, W., Qi, Z., Fuxin, L., 2019. Pointconv: Deep convolutional networks on 3d point clouds. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9621–9630.

Wu, X., Jiang, L., Wang, P.-S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H., 2024. Point transformer v3: Simpler faster stronger. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4840–4851.

Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H., 2022. Point transformer v2: Grouped vector attention and partition-based pooling. *The Advances in Neural Information Processing Systems (NeurIPS)*, 33330–33342.

Xiang, X., Wang, L., Zong, W., Li, G., 2022. Extraction of local structure information of point clouds through space-filling curve for semantic segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 114, 103027.

Yan, L., Song, J., Xie, H., Wei, P., Li, G., Zhu, L., Fan, Z., Gong, S., 2025. LiDGS: An efficient 3D reconstruction framework integrating lidar point clouds and multi-view images for enhanced geometric fidelity. *International Journal of Applied Earth Observation and Geoinformation*, 142, 104730.

Yang, Y.-Q., Guo, Y.-X., Xiong, J.-Y., Liu, Y., Pan, H., Wang, P.-S., Tong, X., Guo, B., 2025a. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. *Computational Visual Media*, 11(1), 83–101.

Yang, Y., Xun, T., Hao, K., Wei, B., Tang, X.-s., 2025b. Grid Mamba: Grid State Space Model for large-scale point cloud analysis. *Neurocomputing*, 636, 129985.

Zeng, Z., Xu, Y., Xie, Z., Tang, W., Wan, J., Wu, W., 2024. Large-scale point cloud semantic segmentation via local perception and global descriptor vector. *Expert Systems with Applications*, 246, 123269.

Zhang, T., Yuan, H., Qi, L., Zhang, J., Zhou, Q., Ji, S., Yan, S., Li, X., 2025. Point cloud mamba: Point cloud learning via state space model. *The AAAI Conference on Artificial Intelligence (AAAI)*, 10121–10130.

Zhao, H., Jiang, L., Jia, J., Torr, P. H., Koltun, V., 2021. Point transformer. *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 16259–16268.

Zhao, H., Zhang, M., Zhao, W., Ding, P., Huang, S., Wang, D., 2025. Cobra: Extending mamba to multi-modal large language model for efficient inference. *The AAAI Conference on Artificial Intelligence (AAAI)*, 10421–10429.

Zhao, L., Xu, S., Liu, L., Ming, D., Tao, W., 2022. SVASeg: Sparse Voxel-Based Attention for 3D LiDAR Point Cloud Semantic Segmentation. *Remote Sensing*, 14.

Zheng, Y., Wang, G., Liu, J., Pollefeys, M., Wang, H., 2024. Spherical frustum sparse convolution network for lidar point cloud semantic segmentation. *The Advances in Neural Information Processing Systems (NeurIPS)*, 121827–121858.

Zhou, H., Zhu, X., Song, X., Ma, Y., Wang, Z., Li, H., Lin, D., 2020. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. *arXiv preprint: 2008.01550*.

Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D., 2021. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9939–9948.