Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MGSMamba: Multi-Stage Grouped Structure and Bidirectional State Space Modeling for Robust 3D Point Cloud Classification

Yang Yu

yuyang@swpu.edu.cn

Southwest Petroleum University

Zengyi Zhang

Southwest Petroleum University

Qiang Zeng

Southwest Petroleum University

Qijun Hu

Southwest Petroleum University

Xingpeng Zhang

Southwest Petroleum University

Leping He

Southwest Petroleum University

Qijie Cai

Southwest Petroleum University

Yu Bai

Southwest Petroleum University

Research Article

Keywords: Point cloud classification and segmentation, Grouping structure, Multi-stage modeling, Bidirectional state space

Posted Date: May 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9439284/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MGSMamba: Multi-Stage Grouped Structure and Bidirectional State Space Modeling for Robust 3D Point Cloud Classification

Yang Yu[a], Zengyi Zhang[a], Qiang Zeng[a], Qijun Hu[b,*], Xingpeng Zhang[a], Leping He[d], Qijie Cai[d], Yu Bai[d]

[a]*School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu, China*
[b]*School of Civil Engineering, Southwest Jiaotong University, Chengdu, China*
[c]*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China*
[d]*School of Civil Engineering and Geomatics, Southwest Petroleum University, Chengdu, China*

Abstract

Point cloud classification is a fundamental task in 3D vision and robotics, yet effectively capturing local geometric structures while modeling global contextual dependencies remains challenging. Existing Mamba-based frameworks excel in sequential modeling but often underutilize spatial and structural information intrinsic to unordered 3D point clouds. To address these limitations, we propose MGSMamba, a multi-stage point cloud classification framework integrating grouping structures with bidirectional state space modeling. Our framework implements a hierarchical three-level feature extraction process: intra-group structural modeling via a group-wise embedding module (GEM) incorporating local normals and point density, inter-group semantic interaction via a cross-group feature learning module (CGLM) with grouped affine normalization (GANorm) and rotational feature-enhanced pooling (RFEP), and global contextual understanding through a shifted bidirectional state space module (Shifted Bi-SSM). Experiments on ModelNet40, ScanObjectNN, and ShapeNet-Part datasets demonstrate that MGSMamba outperforms state-of-the-art methods in accuracy, robustness, and generalization, with a 2–4% improvement in classification accuracy across benchmarks while preserving invariance to point order. Our approach offers a scalable, interpretable, and reproducible framework for point cloud analysis with potential applications in robotics, autonomous navigation, and 3D scene understanding. The source code is available at GitHub: `https://github.com/2359063436/MGSMamba`, and detailed documentation is provided in the repository for full reproducibility.

Keywords: Point cloud classification and segmentation, Grouping structure, Multi-stage modeling, Bidirectional state space.

*Corresponding author.
Email address: `yuyang@swpu.edu.cn` (Yang Yu)

1. Introduction

Three-dimensional point cloud data, typically acquired through LiDAR, depth cameras, and stereo vision systems, contain rich spatial and structural information. Classification and segmentation are fundamental tasks in point cloud analysis and play pivotal roles in a wide range of applications, including object detection [1, 2], medical imaging [3, 4], and 3D reconstruction [5, 6]. Compared to traditional 2D image data, point clouds provide greater precision and flexibility in representing the geometric and spatial distribution of objects. However, their inherent irregularity and unstructured nature pose significant challenges for deep feature learning [7]. To address these challenges, researchers have developed various deep learning techniques for point cloud processing, such as PointNet [8], PointNet++ [9], and DGCNN [10]. Among these, Transformer-based methods have achieved considerable success due to their strong capacity for global context modeling and their ability to capture complex 3D structures [11, 12, 13]. The self-attention mechanism, in particular, facilitates the establishment of long-range dependencies throughout the point cloud, enabling significant advances in classification, segmentation, and object detection tasks [14, 15]. Despite their modeling strength, Transformer architectures are constrained by the quadratic complexity of the self-attention mechanism. As the number of input points increases, both computational and memory demands grow rapidly [16, 17], posing serious scalability issues and limiting their efficiency in large-scale point cloud applications.

Subsequently, due to Mamba's strong ability to capture sequential dependencies and its adoption of hardware-efficient scanning mechanisms that achieve nearly linear computational complexity [18], it has attracted significant attention and achieved success in 2D vision tasks [19, 20, 21]. Mamba has now rapidly expanded into various 3D data fields [22, 23]. In particular, it shows great potential in point cloud analysis [24, 25, 26]. However, unlike inherently sequential data such as audio and text, point clouds are unordered, unstructured, and often exhibit non-uniform density [27]. These characteristics present significant challenges for the direct application of Mamba in point cloud scenarios. Consequently, a key research question arises: how can Mamba be adapted to effectively perceive and model the spatial structures of point cloud data to enhance feature representation? To address this challenge, various Mamba-based enhancement strategies have been proposed, primarily focusing on optimizing point cloud serialization methods and positional encoding techniques. On the one hand, serialization aims to transform unordered point clouds into structured sequences that are compatible with Mamba's state space modeling framework. Existing methods, including axial sorting [24], local cluster-based ordering [28], and hierarchical structural sorting [26], partially preserve the spatial distribution of points, thereby enhancing sequence modeling effectiveness. On the other hand, positional encoding explicitly incorporates the spatial location information of points [25], enabling the model to capture both relative and absolute positional relationships among points. This approach helps mitigate the adverse effects of input order on modeling performance.

Although the aforementioned methods have advanced global modeling of point clouds, they remain limited in effectively representing and leveraging local geometric information. Previous studies have demonstrated that incorporating geometric attributes, such as surface normals, as auxiliary features significantly enhances a model's ability to perceive and understand local surface structures [8, 9, 29]. However, current Mamba-based

methods generally overlook the effective integration of geometric features such as normals and point density. Moreover, existing approaches often incorporate normals independently on a per-point basis, failing to model their spatial contextual relationships comprehensively. Additionally, these methods continue to exhibit limitations in structural feature modeling, as they primarily focus on point cloud embedding, ordering, and global modeling processes without systematically improving local feature extraction capabilities. Although some studies have introduced convolutional branches [30] or local feature aggregation mechanisms [7] to strengthen local representations, significant challenges remain in modeling fine-grained geometric structures and capturing complex local relationships.

To address the aforementioned challenges, we propose a multi-stage self-learning framework for point cloud feature extraction called MGSMamba, which integrates grouped structures with bidirectional state space modeling. The framework employs a three-level feature extraction mechanism that sequentially achieves intra-group structural modeling, inter-group semantic interaction, and global contextual understanding. By systematically integrating local geometric details with global structural semantics, MGSMamba significantly enhances the model's perceptual capabilities and the effectiveness of feature representation.

Our main contributions can be summarized as follows:

(1) A group-wise embedding module (GEM) is designed to incorporate local normal information and point density, thereby enhancing intra-group feature extraction and improving the accuracy of point cloud geometric feature representation.

(2) A cross-group feature learning module (CGLM) is developed to optimize the propagation and aggregation of inter-group features through grouped affine normalization (GANorm) and rotation feature-enhanced pooling (RFEP), thereby enhancing the model's expressive power and generalization ability.

(3) A shifted bidirectional state space module (Shifted Bi-SSM) is introduced to break the symmetry between the forward and backward branches by applying a channel-wise rolling perturbation to the reverse branch, while preserving the model's invariance to the order of points.

(4) Extensive experiments conducted on the ModelNet40, ScanObjectNN, and ShapeNet-Part datasets demonstrate the efficiency and effectiveness of MGSMamba.

The remainder of this paper is organized as follows: Section 2 presents a concise review of deep learning methods for point cloud modeling; Section 3 introduces the MGSMamba framework and its constituent modules; Section 4 offers a comprehensive evaluation and analysis of the proposed method using three public datasets; finally, Section 5 concludes the paper.

2. Related Work

In recent years, deep learning–based methods for point cloud analysis have achieved significant performance improvements, driven by a variety of innovative techniques proposed by researchers. These methods can be broadly categorized into three types: multilayer perceptron (MLP)–based methods, Transformer-based methods, and Mamba-based methods.

MLP-based methods have established a strong foundation in point cloud deep learning. Early works such as PointNet [8] and PointNet++ [9] pioneered frameworks that employed shared multilayer perceptrons (MLPs) to efficiently extract features from unordered point sets, providing both a theoretical foundation and technical support for subsequent point cloud representation learning. In recent years, the integration of pretraining strategies and data augmentation techniques has further enhanced the performance and robustness of MLP-based methods. Meanwhile, significant progress in 3D geometric modeling [31] and structure-preserved visual computing [32] has also provided valuable guidance for developing effective point cloud learning frameworks.For instance, PointContrast [33] utilizes contrastive learning to optimize feature representations, enabling the model to learn semantically consistent features from unlabeled point clouds and thereby improve downstream task performance. Building upon this, DepthContrast [34] leverages spatial relationships and temporal consistency to enhance the model's resilience to noise and sparse sampling. These developments highlight the significant potential of self-supervised learning and data augmentation in advancing point cloud representation learning and provide stronger empirical support for MLP-based architectures. Nevertheless, MLP-based methods exhibit limited capacity for global modeling and often struggle to capture the complex spatial dependencies inherent in point clouds.

Transformer-based methods have significantly advanced point cloud analysis by enabling global feature modeling through self-attention and permutation invariance. Originally designed for natural language processing, transformers effectively handle unordered point sets and model long-range dependencies. Inspired by Vision Transformers (ViT), approaches such as Point-BERT [35] utilize masked modeling with self-supervised pretraining to enhance semantic understanding, while Point-MAE [36] employs masked autoencoding to reconstruct missing regions and learn robust representations. Despite these strengths, the quadratic complexity of self-attention limits scalability to large point clouds. To mitigate this, OctFormer [37] introduces an octree-based local attention mechanism that spatially inputs spatially to reduce computational costs without sacrificing performance. PointGPT [38] models point blocks autoregressively, demonstrating strong generalization in few-shot and downstream tasks. PCT [14] leverages permutation invariance combined with farthest point sampling and nearest neighbor search to effectively capture local context, achieving state-of-the-art results in classification and segmentation. Although transformer-based methods show promising performance, their high computational cost remains a critical challenge, driving research toward more efficient yet expressive architectures for large-scale point cloud modeling.

Mamba-based methods have introduced efficient paradigms for the sequential modeling of point cloud data. To adapt to the unordered and irregular structure of 3D point clouds, recent approaches represent them as sequences while preserving spatial geometry. For instance, PointMamba [24] encodes point blocks sorted along the x, y, and z axes, while PCM [25] generates six sequence variants by permuting coordinate orders to enhance viewpoint diversity. NIMBA [28] addresses spatial continuity by sorting along the y-axis and grouping adjacent clusters, and OctreeMamba [26] imposes a global z-order based on the octree structure to capture causal dependencies. While these methods effectively model long-range relationships, they often overlook local geometric details. To mitigate this, [7] introduces Local Norm Pooling and bidirectional SSMs to improve both local feature extraction and global modeling. Nevertheless, fully leveraging fine-grained geometric structures remains a key challenge.

3. Methods

3.1. MGSMamba Overview

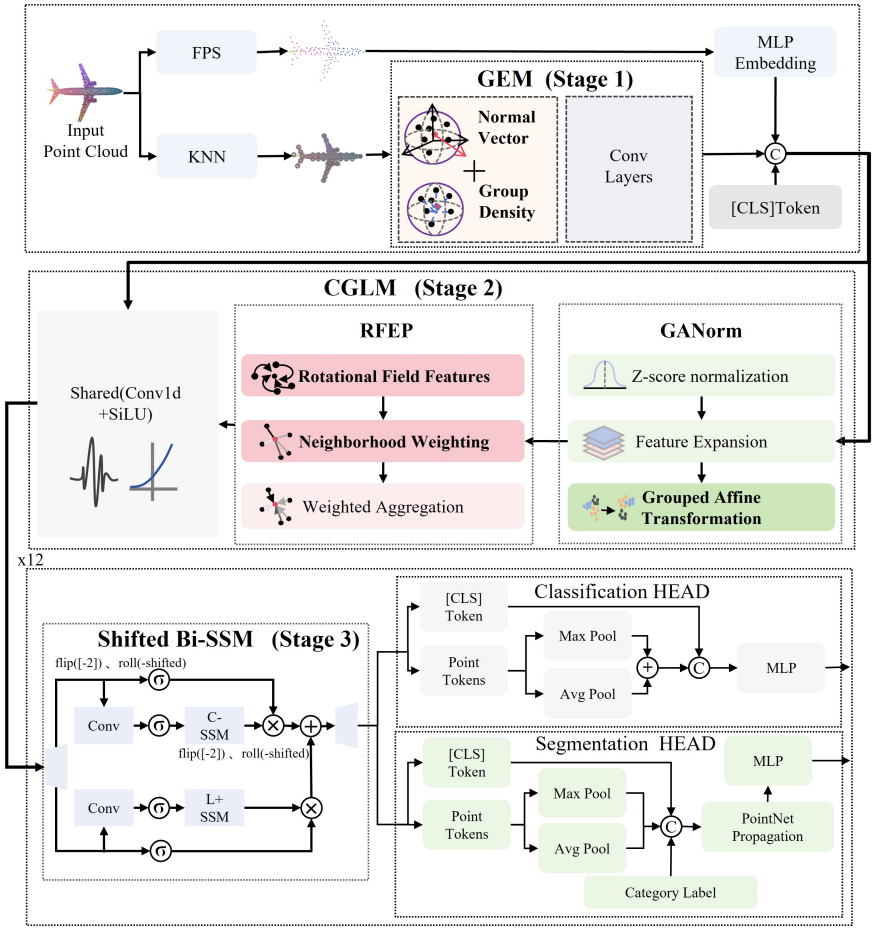

Fig. 1 Overview of the proposed MGSMamba

As illustrated in Figure 1, MGSMamba constructs a multi-stage framework for self-learning of point cloud features, consisting of three main stages: intra-group structural modeling, inter-group semantic interaction, and global contextual understanding.

Initially, the input point cloud is grouped using Farthest Point Sampling (FPS) and k-Nearest Neighbors (KNN). The first stage is implemented by the group-wise embedding module (GEM), which extracts initial features by integrating convolution, normalization, and pooling operations. This module also explicitly incorporates local normal vectors and point density, and constructs learnable positional encodings to enhance local geometric representations.

Subsequently, the extracted intra-group features are processed through stacked Mamba blocks. The second stage—inter-group semantic interaction—is implemented via the

cross-group feature learning module (CGLM). This module facilitates local feature propagation through grouped affine normalization (GANorm), performs effective feature aggregation by combining rotation feature-enhanced pooling (RFEP), and utilizes shared convolution layers to maintain consistent feature dimensions. Together, these operations enhance cross-group feature transmission and semantic fusion efficiency.

Building upon these stages, a shifted bidirectional state space module (Shifted Bi-SSM) is introduced to achieve the third stage: global contextual understanding. The Shifted Bi-SSM captures multi-dimensional feature dependencies using both forward and channel-wise rolled bidirectional structures and fuses information from these two branches, thereby enhancing the model's global representation capability. For downstream tasks, the output from the 12th Mamba block serves as input to the classification head for classification. In segmentation tasks, multi-scale features aggregated from the 4th, 8th, and 12th layers are fed into the segmentation head, improving semantic parsing performance.

3.2. Group-wise Embedding Module

To improve the representation of local geometric structures within point cloud groups, we propose a group-wise embedding module (GEM) that explicitly integrates geometric priors into the feature extraction process. As illustrated in Fig. 2, the module first applies two consecutive stages of convolution, normalization, and pooling to each group, producing preliminary local features.

To construct the local geometric descriptor f_g that captures intrinsic geometric properties, we compute both the normal vectors n and point density ρ for each group. For each point $p_i = (x_i, y_i, z_i)$ in the group, the normal vector n is estimated as:

$$C = \frac{1}{k}\sum_{i=1}^{k}(p_i - \bar{p})(p_i - \bar{p})^\top, \quad n = \arg\min_{v} v^\top C v, \tag{1}$$

where $\bar{p}$ is the center point of the group, and n is also the eigenvector corresponding to the smallest eigenvalue of C. The point density ρ is computed as:

$$\rho = \frac{k}{V}, \tag{2}$$

where k is the number of each group and V is the volume of each group.

The normal vectors and point density together form the local geometric descriptor f_g, explicitly capturing the structural properties of each group. To further incorporate spatial priors, a learnable positional token t_{pos} is added for each group. The final group representation is then constructed as:

$$F_C = \text{Concat}(f_{init}, f_g, t_{pos}), \tag{3}$$

where f_{init} denotes the preliminary features extracted by the convolutional layers. By jointly leveraging learned features, normal vectors, point density, and positional tokens, the GEM module delivers a discriminative, geometry-aware representation that facilitates deeper geometric modeling in subsequent network layers.

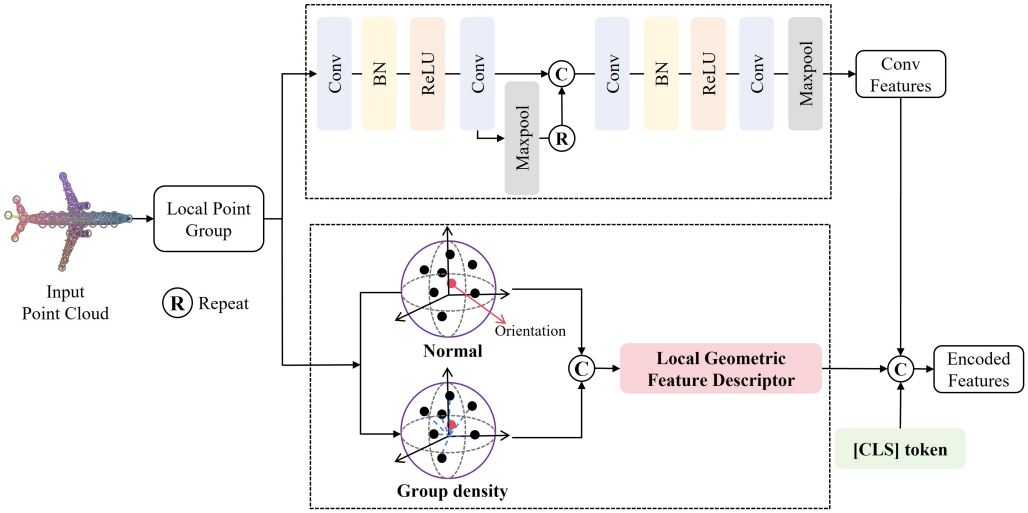

Fig. 2 Architecture of the Group-wise Embedding Module (GEM)

3.3. Cross-group Feature Learning Module

Local geometric features are widely recognized as essential for point cloud feature learning. Typically, these features are extracted by constructing KNN-based local graphs and performing feature fusion on them. Mamba3D [7] simplifies the local feature extraction process into two primary steps: feature propagation and aggregation. Building on this optimized design, we propose a cross-group feature learning module that further enhances the quality of local geometric feature extraction through GANorm (propagation) and RFEP (aggregation). The CGLM is illustrated in Figure 3.

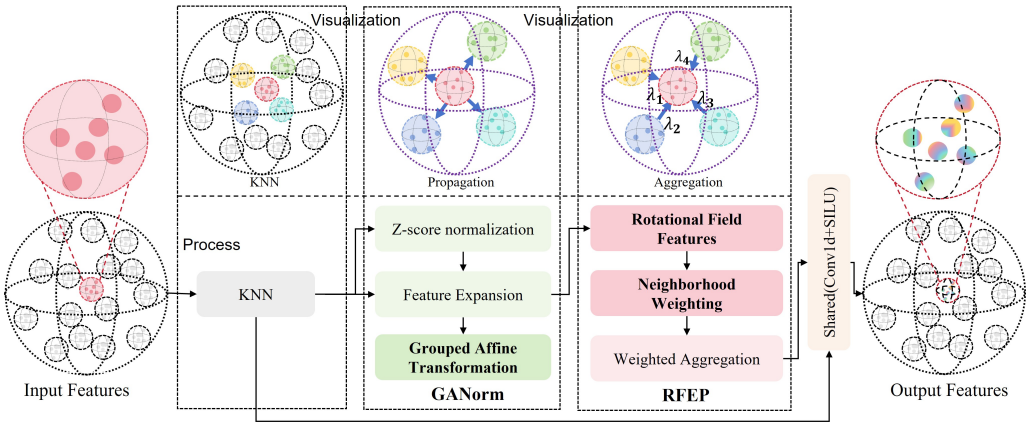

Fig. 3 Structural Diagram of CGLM

3.3.1. Grouped Affine Normalization

The primary objective of feature propagation is to enable neighboring points within a KNN-based local graph to capture relative features with respect to the central point. However, in K-norm [7], feature normalization is applied, followed by a globally shared affine transformation for feature scaling and shifting. Although this approach shows promising performance in local information propagation, it does not fully account for feature differences across distinct groups. Specifically, the same scaling factor α and offset β are shared across all groups, which limits the model's flexibility in handling groups with varying statistical characteristics and consequently reduces the expressiveness of feature propagation.

To address this issue, grouped affine normalization (GANorm) is proposed to enhance the model's adaptability to local features. First, a KNN-based grouping strategy is employed to extract neighborhood information from the input point cloud coordinates and features. Normalization is then applied separately to the neighborhood features and coordinates to ensure stable and consistent feature distributions, thereby alleviating the inherent scale variation problem in point cloud data. Subsequently, the center point features are concatenated with the neighborhood features in GANorm to enrich the feature representation, thereby enhancing the expressiveness of local geometric structures. Moreover, to overcome the limitations of global affine transformations, a group-adaptive affine transformation is introduced, which independently learns the scaling and shifting parameters, α_{group} and β_{group}, for each neighborhood within a feature group. This improvement enables adaptive adjustment of each group according to its own feature distribution, allowing the model to flexibly capture inter-group feature variations and enhance the adaptability of feature propagation. The computation of GANorm is shown in Equation (4):

$$\begin{cases} F_K, K_{xyz} = Grouping(F_C, C_{xyz}), \\ F_K^{\sim} = \frac{F_K - F_C}{\sqrt{\text{Var}(F_K - F_C) + \epsilon}}, \\ F_{ck} = [F_K^{\sim} \oplus F_C] \times \alpha_{group} + \beta_{group} \end{cases} \tag{4}$$

where F_K denotes the neighborhood features for each group, K_{xyz} represents the neighborhood coordinates of each group, F_C represents the feature of the centroid point, and C_{xyz} denotes the coordinate of the centroid point. $F_K^{\sim}$ denotes the normalized neighborhood features, $\oplus$ indicates concatenation, and α_{group} and β_{group} are group-specific learnable affine transformation parameters. The final output F_{ck} corresponds to the feature after propagation.

Through GANorm, independent learnable affine transformation parameters are introduced for the neighborhood of each group, following the construction and normalization of local feature maps. This design significantly enhances the model's adaptability to diverse local structures. Specifically, when processing point cloud data with pronounced local variations, the proposed method effectively captures the distinct characteristics of each group, thereby improving both the representational capacity and generalization ability of the model.

3.3.2. Rotation Feature Enhanced Pooling

The objective of feature aggregation is to fuse propagated features to update point cloud feature representations. Common feature aggregation methods include max pooling, average pooling, and weighted combinations of these techniques. However, these

conventional approaches often suffer from significant information loss during aggregation, especially when processing point clouds with complex geometric structures. Although K-pooling [7] partially mitigates this issue through exponentially weighted averaging, it still fails to fully capture the geometric relationships among neighboring points, thereby neglecting the impact of spatial transformations on feature representations.

To further optimize the feature aggregation process, rotational feature enhanced pooling (RFEP) is proposed. This method explicitly models geometric relationships among neighboring points to more accurately capture local geometric features in point cloud data. Specifically, the coordinate differences between neighboring groups and their corresponding center points are first computed using the KNN strategy. This relative positional information is subsequently used to extract rotational field features, thereby enhancing the model's sensitivity to geometric transformations such as rotation and translation. The computation of these relative coordinate differences is defined in Equation (5):

$$d_{xyz} = K_{xyz} - C_{xyz} + \epsilon \tag{5}$$

where K_{xyz} represents the coordinates of the center point of the neighboring groups, and C_{xyz} represents the coordinates of the center point.

Subsequently, the rotational field feature F_{rot} is computed using the cross product operation to characterize the relative rotational information among neighboring groups, as expressed in Equation (6):

$$F_{\text{rot}} = d_{xyz}(i) \times d_{xyz}(i+1) \tag{6}$$

where $\times$ denotes the cross product operation. The rotational field feature F_{rot} captures the geometric relationships between neighboring points, enhancing the model's robustness to rotational and translational variations, thereby ensuring accurate representation of the local geometric structure.

To further enhance the accuracy of feature aggregation, a group-adaptive weighting mechanism is introduced to enable adaptive adjustment of the contributions from different neighborhood features to the final representation. Specifically, an exponential transformation is applied to the rotational field features to generate weighting coefficients for each feature group. The computation of the weighted neighborhood features is formulated in Equation (7):

$$F_{weighted} = F_{ck} \cdot \exp(\gamma_{group} \cdot F_{rot}) \tag{7}$$

where γ_{group} is a learnable parameter specific to each group, F_{ck} denotes the neighborhood features propagated through GANorm, F_{rot} represents the rotation field features, and F_{weighted} corresponds to the weighted neighborhood feature.

Subsequently, the weighted features are aggregated through a pooling operation. A weighted exponential pooling method is employed to amplify strongly correlated features while attenuating weakly correlated ones, thereby producing the final aggregated feature representation $F_C{}^{\sim}$. The computation is expressed in Equation (8):

$$F_C^{\sim} = \frac{\sum_{i=1}^{K} F_{weighted} \cdot \exp(F_{weighted})}{\sum_{i=1}^{K} \exp(F_{weighted}) + \epsilon} \tag{8}$$

RFEP accurately characterizes local geometric structures by computing rotational field features among neighboring points and incorporates weighted pooling strategies during feature aggregation to enhance the precision of feature representation. This method not only effectively captures local geometric details but also ensures the stability and robustness of the aggregated results when processing point cloud data under geometric transformations.

3.4. Shifted Bi-SSM

The Channel Bidirectional State Space Model (Bi-SSM) [7] addresses the pseudo-sequential bias present in Vision Mamba for point cloud modeling by utilizing a channel-wise flipping strategy, as illustrated in Fig. 4. The complete mathematical formulation is presented in Equation (9):

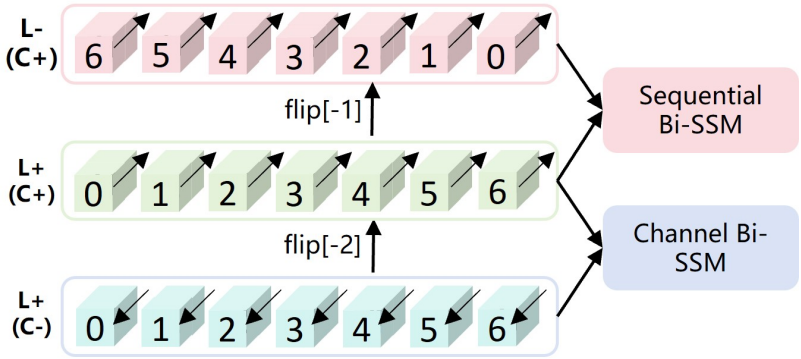

Fig. 4 Illustration of Sequential Bi-SSM and Channel Bi-SSM

$$\begin{cases} F_{fusion} = [L + SSM(F_C^{\sim L+})] + [C - SSM(F_C^{\sim C-})], \\ Bi - SSM(F_C^{\sim}) = F_C^{\sim} + Linear(F_{fusion}) \end{cases} \tag{9}$$

where $F_C^{\sim L+}$ denotes the original input features fed into the forward state space module (L+SSM), and $F_C^{\sim C-}$ represents the channel-wise flipped features processed by the reverse modeling branch (C-SSM).

While the bidirectional state-space model (Bi-SSM) mitigates pseudo-order bias through symmetric forward and backward branches, its high structural symmetry often induces representational redundancy, limiting the diversity of captured features. To address this limitation, we propose the Shifted Bi-SSM, which incorporates a lightweight channel-wise perturbation mechanism between the two branches. Specifically, a fixed rolling shift along the channel dimension is applied to both the input and output of the backward branch. This operation disrupts the exact symmetry and promotes complementary feature learning, resulting in more discriminative bidirectional representations. Formally, given the reversed feature sequence $F_C^{\sim C-}$ (obtained by flipping the forward features), a channel-wise cyclic shift with a fixed offset is applied before it enters the reverse SSM module, and a corresponding inverse shift is applied after processing. This simple yet

effective modification preserves the bidirectional flow while enhancing representational divergence across directions.

$$F_C^{\sim C-'} = \text{Roll}(F_C^{\sim C-}, shift),$$

where $shift$ denotes a predefined shift value. This operation introduces a relative positional perturbation, allowing the reverse branch to more effectively capture cross-regional geometric variations and reduce redundancy with the forward path.

To ensure consistent feature alignment, an inverse shift followed by a flip is applied to the output of the C-SSM:

$$\hat{F}_{C-} = \text{Flip}(\text{Roll}^{-1}(\text{C-SSM}(F_C^{\sim C-'}), shift)).$$

The final fused representation is obtained as follows:

$$\begin{cases} F_{\text{fusion}} = \text{L+SSM}(F_C^{\sim L+}) + \hat{F}_{C-}, \\ \text{Shifted-Bi-SSM}(F_C^{\sim}) = F_C^{\sim} + \text{Linear}(F_{\text{fusion}}) \end{cases}$$

Compared to the original Bi-SSM, the proposed Shifted Bi-SSM retains order-agnostic modeling capabilities while breaking the forward-backward symmetry through channel-wise shifting. This enhancement increases feature diversity and improves sensitivity to local geometric variations and directional patterns. Importantly, this is achieved without explicitly encoding point order, thereby maintaining robustness to input permutations.

4. Experiment

4.1. Datasets

ModelNet40 [39] is a widely used, high-quality synthetic 3D CAD dataset comprising 40 categories and 12,311 clean samples. Due to its well-defined structure and balanced classes, it serves as a standard benchmark for 3D shape classification. We follow the official split, using 9,843 samples for training and 2,468 for testing. Data preprocessing and augmentation procedures follow those of PointNet [8] and PCT [14], including anisotropic scaling, random rotations, and translations to enhance generalization and robustness.

ScanObjectNN [40] is a challenging 3D point cloud dataset derived from real-world indoor scans, comprising approximately 15,000 objects across 15 categories. It is notable for including realistic imperfections such as inherent sensor noise, partial occlusions, and cluttered backgrounds, thereby providing a more authentic benchmark for evaluating model robustness in practical applications. The dataset offers three progressively complex variants: **OBJ_ONLY** (containing cleaned and isolated object points); **OBJ_BG** (incorporating real background points alongside the target objects); and **PB_T50_RS** (the most challenging variant, featuring 50% background points, random perturbations, and simulated occlusions). In our experiments, we follow the official dataset splits to ensure a fair and reproducible assessment of model performance under realistic conditions.

ShapeNetPart [41] is a large-scale semantic segmentation dataset derived from ShapeNet [42], comprising 16,880 models across 16 common categories. Each model includes multiple semantic parts spanning 50 labels, with at least two distinct parts per object. We follow the official split, using 13,998 samples for training and 2,874 for validation, and report category mean IoU (mIoU_C) and instance mean IoU (mIoU_I) to evaluate segmentation performance.

4.2. Implementation Details

All experiments were conducted using the PyTorch framework on a computing platform equipped with NVIDIA RTX 4090 GPUs.

Classification. For the classification task, the Mamba block depth was set to 12. The input point cloud was divided into 128 groups, each containing 32 points. The model was optimized using the AdamW optimizer with an initial learning rate of 0.0005 and a weight decay of 0.05. A cosine annealing schedule was applied to dynamically adjust the learning rate. The batch size was fixed at 32, and the total training duration was 300 epochs.

Semantic Segmentation. For the segmentation task, the initial learning rate was set to 0.0002, and the batch size was set to 16. All other training configurations remained consistent with those used in the classification task.

Pre-training strategy. The pre-training phase utilizes the PointMAE framework [36], which is based on a masked point modeling strategy. The overall model employs an encoder–decoder architecture, where the encoder processes only the visible tokens to extract feature representations. Following the pre-training configuration of Mamba3D [7], the decoder is simplified to consist solely of Mamba blocks without the cross-group feature learning module, thereby accelerating model convergence while maintaining performance.

4.3. Comparisons with the state-of-the-art

We compared the proposed MGSMamba with current state-of-the-art (SOTA) methods on three benchmark datasets: ModelNet40 [39], ScanObjectNN [40], and ShapeNet-Part [41]. The comparison results are summarized in Tables 1, 2, and 3, respectively. These methods are grouped into four categories based on their architectural design: MLP-based, Transformer-based, Mamba-based, and Transformer-Mamba hybrid methods.

Comparative experiments on the ModelNet40 dataset are summarized in Table 1. Compared to various self-supervised Transformer-based methods, MGSMamba consistently outperforms existing Transformer models in overall accuracy. Without employing a voting strategy, MGSMamba achieves a 0.5% improvement over the best-performing Transformer model, PointConT [50], and a 1.6% increase compared to the baseline model PointMamba [24]. Moreover, MGSMamba attains a 0.6% gain over the current leading Mamba model, Mamba3D [50]. When applying the voting strategy, MGSMamba further boosts classification accuracy to 94.4% on ModelNet40, substantiating the effectiveness and superiority of the proposed framework on synthetic datasets. After adopting the pretraining strategy, the classification accuracy of LGABMamba on the ModelNet40 dataset improves further to 95.4%, surpassing various comparison models and achieving state-of-the-art performance under current pretraining conditions.

To intuitively demonstrate the feature extraction capabilities and classification performance of the proposed model, we conducted a visualization analysis of the training processes of LGABMamba, PointTramba, and Mamba3D on the ModelNet40 dataset, as illustrated in Figure 5. The results show that LGABMamba achieves higher accuracy while maintaining superior training stability and faster convergence, highlighting its efficient feature modeling capabilities and well-designed optimization mechanism.

To further assess the robustness of MGSMamba in more challenging real-world scenarios, we conducted comparative experiments on the ScanObjectNN dataset. The results, summarized in Table 2, demonstrate that when trained from scratch, MGSMamba

Table 1 Comparison of classification accuracies (%) on the ModelNet40 dataset, with the best performance highlighted in bold

Method	Backbone	Param.(M)	FLOPs(G)	ModelNet40 ↑
PointNet [8]	MLP	3.5	0.5	89.2
PointNet++ [9]	MLP	1.5	1.7	90.7
DGCNN [10]	MLP	1.8	2.4	92.2
PointCNN [43]	MLP	0.6	-	92.9
PointNeXt [44]	MLP	1.4	3.6	92.9
DRNet [45]	MLP	-	-	93.1
SimpleView [46]	MLP	-	-	93.9
GBNet [47]	MLP	8.8	-	93.8
PRA-Net [48]	MLP	2.3	-	93.7
MVTN [49]	MLP	11.2	43.7	93.8
PCT [14]	Transformer	2.9	2.3	93.2
PointConT [50]	Transformer	-	-	93.5
PointMamba [24]	Mamba	12.3	3.6	92.4
SI-Mamba [51]	Mamba	12.3	3.6	92.7
PCM [25]	Mamba	34.2	45.0	92.6
Mamba3D [7]	Mamba	16.9	3.9	93.4
Pointramba [52]	Hybrid	19.5	-	92.8
MGSMamba w/o vot.	Mamba	17.0	3.9	**94.0**
MGSMamba w/ vot.	Mamba	17.0	3.9	**94.4**
with Self-supervised Pre-training				
Point-BERT [35]	Transformer	22.1	4.8	93.4
PointMAE [36]	Transformer	22.1	4.8	94.4
MaskPoint [53]	Transformer	22.1	4.8	93.8
PointMamba [24]	Mamba	12.3	3.6	93.6
SI-Mamba [51]	Mamba	12.3	3.6	93.4
ZigzagPointMamba [54]	Mamba	12.3	3.6	93.15
Mamba3D [7]	Mamba	16.9	3.9	95.1
MGSMamba w/ vot.	Mamba	17.0	3.9	**95.4**

achieves an accuracy of 93.96% on the OBJ_BG subset, significantly outperforming existing methods and improving by 5.66% over the baseline model PointMamba [24]. With the application of a voting strategy, the accuracy further increases to 94.49%, highlighting its strong modeling capability and robustness. On the most challenging PB_T50_RS subset, MGSMamba attains an overall accuracy of 91.92%, which improves to 92.88% after voting is applied. These results surpass the current state-of-the-art Mamba model, Mamba3D [50], further confirming the superior performance of MGSMamba in complex real-world scenarios. After employing the pretraining strategy, MGSMamba demonstrated further improvements in classification performance on the ScanObjectNN dataset, achieving 95.42% accuracy on the OBJ_BG subset, 93.12% on the OBJ_ONLY subset, and 93.63% on the PB_T50_RS subset. This pretraining strategy effectively enhanced LGABMamba's adaptability to noise, occlusion, and background clutter, thereby significantly improving its ability to model and discriminate point cloud features in complex real-world environments.

To evaluate the modeling capability of MGSMamba in additional point cloud analysis tasks, we further tested it on the point cloud semantic segmentation task using the ShapeNetPart dataset. The results are summarized in Table 3. The experimental results show that, without employing the voting strategy, MGSMamba achieved category mean IoU (mIoU_C) and instance mean IoU (mIoU_I) of 83.9% and 85.8%, respectively. Compared to the MLP-based DGCNN [10], MGSMamba improved mIoU_C and mIoU_I

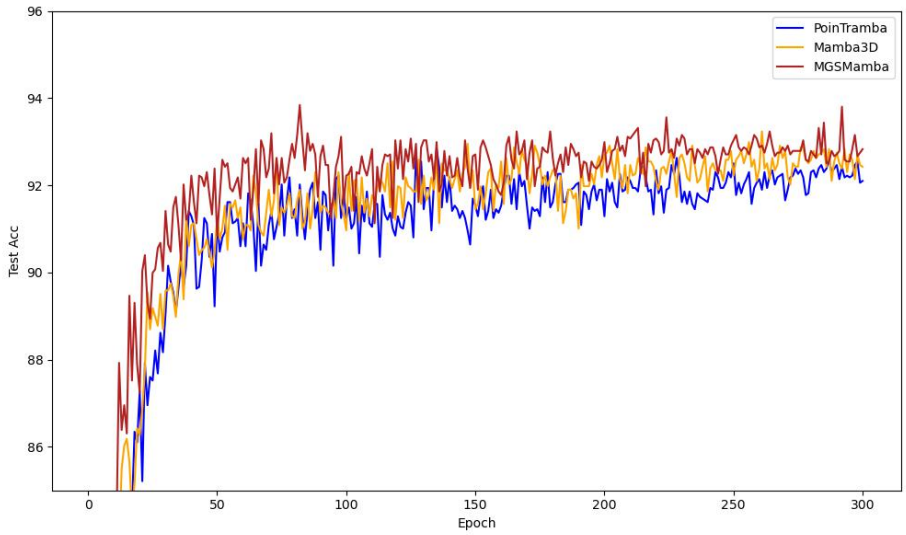

Fig. 5 Comparison of training accuracy among MGSMamba, PointMamba, and Mamba3D on the ModelNet40 dataset

by 1.6% and 0.6%, respectively. Relative to the current best-performing Mamba model, Mamba3D [7], it achieved gains of 0.2% and 0.1%, respectively. These results further confirm the superior performance of MGSMamba in semantic segmentation tasks. After applying the pretraining strategy, MGSMamba achieved further improvements on the ShapeNetPart dataset, with mIoU_C and mIoU_I reaching 84.3% and 86.2%, respectively, where mIoU_I attained the current state-of-the-art performance.

4.4. Ablation Study

This section conducts a systematic ablation study on the ScanObjectNN (OBJ_BG) dataset to evaluate the influence of different components within the MGSMamba architecture. By selectively replacing or removing key modules, we quantitatively assess the contribution of each component to feature representation quality and model robustness, thereby confirming the effectiveness of the overall design.

4.4.1. Component Ablation

In the component ablation study of MGSMamba, we replaced individual modules with their counterparts from Mamba3D [7] within a unified architecture for comparative analysis. The results, presented in Table 4, show that removing the GEM module causes a 0.16% drop in overall accuracy, demonstrating its important role in integrating normal information and enhancing spatial feature representation. In contrast, omitting the GANorm module leads to a larger accuracy decrease of 0.4%, underscoring its effectiveness in adaptively modeling feature variations across different groups. Moreover, removing the RFEP module results in a 0.33% reduction in accuracy, highlighting its critical function in improving the rotation invariance of point cloud features. Finally, excluding the Shifted Bi-SSM causes a 0.16% accuracy decline, indicating its contribution

Table 2 Comparison of classification accuracies (%) on three variants of the ScanObjectNN dataset, with the best performance highlighted in bold

Method	Backbone	Param.(M)	FLOPs(G)	OBJ_BG ↑	OBJ_ONLY ↑	PB_T50_RS ↑
PointNet [8]	MLP	3.5	0.5	73.3	79.2	68.0
PointNet++ [9]	MLP	1.5	1.7	82.3	84.3	77.9
DGCNN [10]	MLP	1.8	2.4	82.8	86.2	78.1
PointCNN [43]	MLP	0.6	-	86.1	85.5	78.5
DRNet [45]	MLP	-	-	-	-	80.3
SimpleView [46]	MLP	-	-	-	-	80.5±0.3
GBNet [47]	MLP	8.8	-	-	81.0	81.0
PRA-Net [48]	MLP	2.3	-	-	81.0	81.0
MVTN [49]	MLP	11.2	43.7	92.6	**92.3**	82.8
PointMLP [55]	MLP	12.6	31.4	-	-	85.4±0.3
PointNeXt [44]	MLP	1.4	3.6	-	-	87.7±0.4
P2P-HorNet [56]	MLP	-	34.6	-	-	89.3
DeLA [57]	MLP	5.3	1.5	-	-	90.4
PCT [14]	Transformer	2.9	2.3	-	-	-
Point-BERT [35]	Transformer	22.1	4.8	79.9	80.6	77.2
PointMAE [36]	Transformer	22.1	4.8	86.8	86.9	86.9
SPoTr [58]	Transformer	1.7	10.8	-	-	88.60
PointConT [50]	Transformer	-	-	-	-	90.30
PointMamba [24]	Mamba	12.3	3.6	88.30	87.78	82.48
SI-Mamba [51]	Mamba	12.3	3.6	92.25	91.39	87.30
PCM [25]	Mamba	34.2	45.0	-	-	88.10±0.3
Mamba3D [7]	Mamba	16.9	3.9	92.94	92.08	91.81
Pointramba [52]	Hybrid	19.5	-	92.3	90.9	84.5
MGSMamba w/o vot.	Mamba	17.0	3.9	**93.96**	91.96	**91.92**
MGSMamba w/ vot.	Mamba	17.0	3.9	**94.49**	**92.36**	**92.88**
with Self-supervised Pre-training						
Point-BERT [35]	Transformer	22.1	4.8	87.43	88.12	83.07
PointMAE [36]	Transformer	22.1	4.8	90.02	88.29	85.18
MaskPoint [53]	Transformer	22.1	4.8	89.30	88.10	84.30
PointMamba [24]	Mamba	12.3	3.6	94.32	92.60	89.31
SI-Mamba [51]	Mamba	12.3	3.6	94.32	91.91	89.10
ZigzagPointMamba [54]	Mamba	12.3	3.6	94.15	92.10	88.65
Mamba3D [7]	Mamba	16.9	3.9	95.18	92.60	93.34
MGSMamba w/ vot.	Mamba	17.0	3.9	**95.42**	**93.12**	**93.63**

to the effective fusion of forward and backward features and the overall enhancement of feature representation capability.

4.4.2. Ablation study of the RFEP module

Building on the architectural ablation study, we conducted a more detailed evaluation of the RFEP module, with comprehensive results presented in Table 5. To rigorously assess its effectiveness, we replaced RFEP with several widely used local aggregation strategies: AvgPooling, MaxPooling, the average of Max and AvgPooling ((Max+Avg)/2), and K-pool [7]. The results show respective decreases in overall accuracy of 0.67%, 1.88%, 1.71%, and 0.33%. These findings underscore the superior capability of RFEP in modeling local features, as it consistently outperforms alternative methods in capturing discriminative and rotation-invariant representations.

4.4.3. Ablation of RFEP Parameter

To further investigate the influence of the parameter γ within the RFEP module, we conducted targeted ablation experiments, the results of which are summarized in Table 6. The findings reveal that fixing γ at different constant values generally leads to a decrease in model accuracy on the OBJ_BG dataset. This suggests that static parameter settings struggle to accommodate the diverse feature distributions present across samples,

Table 3 Comparison of classification accuracies (%) on the ShapeNetPart dataset, with the best performance highlighted in bold

Method	Backbone$_C$ (%)	mIoU$_C$ (%)	mIoU$_I$ (%)	Param.(M)
PointNet [8]	MLP	80.4	83.7	3.6
PointNet++ [9]	MLP	81.9	85.1	**1.0**
DGCNN [10]	MLP	82.3	85.2	1.3
PointMAE [36]	Transformer	83.4	85.2	1.3
Pointramba [52]	Hybrid	-	85.7	25.4
Mamba3D [7]	Mamba	83.7	85.7	23.0
MGSMamba	Mamba	**83.9**	**85.8**	29.2
with Self-supervised Pre-training				
Point-BERT [35]	Transformer	84.1	85.6	27.1
PointMAE [36]	Transformer	84.2	86.0	27.1
MaskPoint [53]	Transformer	**84.6**	86.0	27.1
PointMamba [24]	Mamba	84.4	86.2	27.1
SI-Mamba [51]	Mamba	-	86.1	27.1
ZigzagPointMamba [54]	Mamba	84.2	85.8	27.1
Mamba3D [7]	Mamba	84.1	85.7	23.0
MGSMamba w/ vot.	Mamba	84.3	**86.2**	29.2

Table 4 Ablation on model component

GEM	GANorm	RFEP	Shifted Bi-SSM	OBJ_BG(%)
				92.94
	✓	✓	✓	93.80
✓		✓	✓	93.46
✓	✓		✓	93.63
✓	✓	✓		93.80
✓	✓	✓	✓	**93.96**

thereby limiting the model's generalization capability. In contrast, treating γ as a learnable parameter enables the model to adaptively adjust during training, resulting in a significant accuracy improvement of 93.96%. These results validate the effectiveness and flexibility of the learnable γ design in optimizing the model's performance.

4.4.4. Time Efficiency Analysis

To systematically evaluate the computational performance of LGABMamba, this section presents a quantitative comparative analysis of the training times of PointMamba, PoinTramba, Mamba3D, and LGABMamba on the ModelNet40 dataset under identical experimental conditions. Each model was trained for 300 batches, and the average training time was recorded as the metric for time efficiency. The experimental results are summarized in Table 7.

The experimental results indicate that the average training time of different models varies under identical experimental conditions. PointMamba exhibits the shortest training time due to its relatively simple network architecture and the absence of explicit local geometric feature modeling. In contrast, PoinTramba incorporates multi-head attention mechanisms and feature importance ranking strategies into its architecture, enhancing feature representation and selection capabilities but increasing model parameters and computational complexity, which significantly prolongs training time. Although MGSMamba employs a more sophisticated architecture than Mamba3D, its average training time remains comparable, demonstrating that MGSMamba achieves a favorable balance between enhanced feature representation and computational efficiency, thereby exhibiting

Table 5 Ablation Results of Different Pooling Methods

Method	**OBJ_BG(%)**
Maxpool	93.29
Avgpool	92.08
(Max+Avgpool)/2	92.25
K-pool	93.63
RFEP	**93.96**

Table 6 Ablation study results of RFEP with different γ parameters

γ	**OBJ_BG(%)**
0.2	92.77
0.3	92.94
0.4	93.13
0.5	92.43
0.6	92.08
$\gamma_{\text{learnable}}$	**93.96**

higher optimization efficiency and better engineering feasibility.

4.5. Visualization

Visualization on the ShapeNetPart dataset shows that our MGSMamba method outperforms PointMamba [24], Pointramba [52], and Mamba3D [7] is more accurate. Additionally, category-wise accuracy comparison plots between MGSMamba and PointMamba were generated, as shown in Figures 6 and 7.

Table 7 Comparison of the average training time of different models on the ModelNet40 dataset

Models	PointMamba	PointMamba	PointMamba	MGSMamba
Time	**48.2**	131.2	67.1	68.5

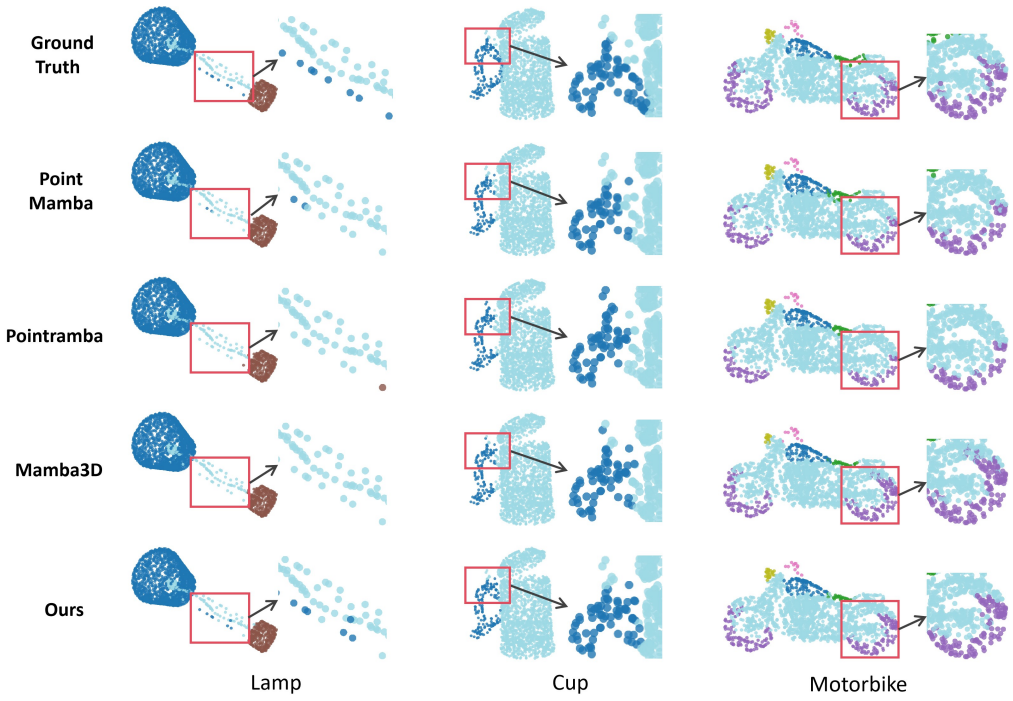

Fig. 6 The qualitative results of part segmentation of our MGSMamba and PointMamba on the ShapeNetPart dataset

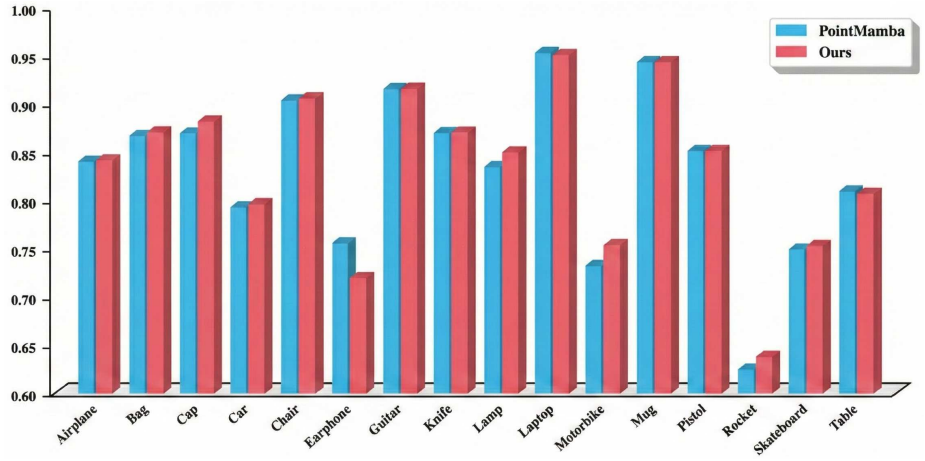

Fig. 7 Comparison of per-category accuracy between MGSMamba and PointMamba on the ShapeNet-Part dataset

Through comparative analysis, it is evident that MGSMamba outperforms PointMamba across most categories, with particularly notable improvements in classes such as "Cap," "Lamp," and "Rocket," where segmentation boundaries tend to be ambiguous or geometric structures are complex. This demonstrates that MGSMamba has an enhanced representational capacity for capturing fine-grained semantic structures and modeling intricate geometric relationships. Additionally, MGSMamba maintains consistent performance advantages in common categories like "Airplane," "Chair," and "Motorbike," further validating the model's generalization ability and robustness. However, it is worth noting that despite its strong overall performance, segmentation accuracy for the "Rocket" category remains considerably lower than for other classes. This challenge primarily arises from the high symmetry and geometric similarity within this class, which complicates the accurate differentiation of semantic parts. The significant improvement observed for MGSMamba in the "Rocket" category can be attributed to its superior local geometric feature extraction capabilities. By effectively incorporating normal information, enhancing the modeling of local feature differences, and increasing robustness to rotational variations, MGSMamba better distinguishes morphologically similar structures with different orientations, enabling more precise semantic segmentation within this challenging category.

5. Conclusion

Compared with recent advanced studies, our method shows distinct differences and advantages. Thilakarathna et al. [59] focused on high-precision hardware sensing for VR finger tracking, which belongs to wearable device design rather than 3D point cloud understanding. Chen et al. [60] proposed a multimodal deep learning framework for cardiac prognosis prediction, emphasizing medical image fusion without involving 3D geometric analysis. Wang et al. [61] addressed non-rigid point cloud registration via anisotropic hybrid field harmonization, which targets shape alignment under deformation instead of 3D object classification. Different from the above tasks and methodologies, MGSMamba

concentrates on 3D point cloud classification and segmentation, and achieves superior performance by hierarchical grouped structure learning and shifted bidirectional state space modeling.

This paper addresses the limitations of existing Mamba-based point cloud classification methods in effectively leveraging local geometric information and structural feature representation by proposing a multi-stage point cloud classification framework integrating grouping structures and bidirectional state space modeling (MGSMamba). The proposed framework significantly enhances the capacity to represent spatial structures of point cloud data and facilitates deep fusion of local geometric details with global structural information through a three-stage design encompassing intra-group structural modeling, inter-group semantic interaction, and global context understanding. In the intra-group stage, the GEM is employed to achieve local feature embedding and enrich geometric information, thereby effectively capturing fine-grained structural characteristics. During the inter-group stage, the CGLM utilizes GANorm and RFEP mechanisms to optimize feature propagation and semantic aggregation across groups. Finally, at the global modeling stage, the Shifted Bi-SSM is introduced to substantially enhance the representation capacity and dynamic modeling of global features. Extensive experimental results demonstrate that MGSMamba attains superior accuracy gains across multiple public datasets, fully validating its exceptional generalization ability and practical effectiveness in complex point cloud analysis tasks, while further advancing the development and widespread application of Mamba-based point cloud analysis techniques.

Although MGSMamba demonstrates outstanding performance and strong generalization ability across various point cloud classification and segmentation tasks, several limitations still remain. First, the collaborative optimization between local geometric modeling and global semantic modeling within the multi-stage framework can be further improved. The current GEM and Shifted Bi-SSM modules rely on a relatively static structural design for inter-level information exchange, which limits their adaptability in capturing multi-scale geometric relationships. As a result, the flexibility and robustness of feature fusion may degrade when handling complex scenes or point clouds with highly varying densities. Second, the model may still suffer from feature degradation under noisy or heavily occluded data conditions. Although the CGLM module enhances inter-group information propagation through the incorporation of GANorm and RFEP mechanisms, local feature aggregation remains susceptible to interference in high-noise or incomplete point clouds, thereby affecting the stability and consistency of global modeling.

Future research will focus on several directions. First, we plan to further enhance the multi-scale adaptive modeling capability of MGSMamba by exploring dynamic weighting and cross-layer feature interaction mechanisms to achieve more flexible geometric–semantic fusion. Second, to improve robustness against noise and missing data, we will incorporate uncertainty modeling and self-supervised geometric completion strategies to enhance model stability in complex real-world scenarios. Finally, MGSMamba will be extended to more challenging tasks such as point cloud scene understanding, cross-modal 3D data fusion, and spatiotemporal dynamic modeling, thereby advancing Mamba-based point cloud representation learning toward higher efficiency, robustness, and generalization.

CRediT authorship contribution statement

Yang Yu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Visualization, Writing – review and editing. **Zengyi Zhang**: Methodology, Investigation, Data curation, Software, Validation. **Qiang Zeng**: Data curation, Formal analysis, Visualization, Writing – original draft. **Qijun Hu**: Funding acquisition, Project administration, Supervision, Resources, Conceptualization, Writing – review and editing. **Xingpeng Zhang**: Methodology, Visualization, Validation. **Leping He**: Resources, Data curation, Visualization. **Qijie Cai**: Data curation, Formal analysis. **Yu Bai**: Resources, Visualization.

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.

Data availability

The datasets used are open-source and available in the repository indicated in the related papers.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. U23A2046); the Sichuan Science and Technology Program (Provincial Academy-University Cooperation Project) (2025YFHZ0015).

References

[1] W. Cai, D. Liu, X. Ning, C. Wang, G. Xie, Voxel-based three-view hybrid parallel network for 3d object classification, Displays 69 (2021) 102076.

[2] C. Yan, G. Pang, X. Bai, C. Liu, X. Ning, L. Gu, J. Zhou, Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss, IEEE Transactions on Multimedia 24 (2021) 1665–1677.

[3] X. Ning, K. Gong, W. Li, L. Zhang, Jwsaa: joint weak saliency and attention aware for person re-identification, Neurocomputing 453 (2021) 801–811.

[4] W. Zhangyu, Y. Guizhen, W. Xinkai, L. Haoran, L. Da, A camera and lidar data fusion method for railway object detection, IEEE Sensors Journal 21 (12) (2021) 13442–13454.

[5] Y. Zhou, O. Tuzel, Voxelnet: End-to-end learning for point cloud based 3d object detection, in: IEEE conference on computer vision and pattern recognition, 2018, pp. 4490–4499.

[6] J. Guo, X. Yao, M. Shen, J. Wang, W. Liao, A deep learning network for point cloud of medicine structure, in: International Conference on Information Technology in Medicine and Education (ITME), IEEE, 2018, pp. 683–687.

[7] X. Han, Y. Tang, Z. Wang, X. Li, Mamba3d: Enhancing local features for 3d point cloud analysis via state space model, in: ACM International Conference on Multimedia, 2024, pp. 4995–5004.

[8] C. R. Qi, H. Su, K. Mo, L. J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660.

[9] C. R. Qi, L. Yi, H. Su, L. J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, in: Advances in Neural Information Processing Systems, 2017, pp. 5099–5108.

[10] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, J. M. Solomon, Dynamic graph cnn for learning on point clouds, ACM Transactions on Graphics (tog) 38 (5) (2019) 1–12.

[11] H. Zhao, L. Jiang, J. Jia, P. H. Torr, V. Koltun, Point transformer, in: IEEE/CVF international conference on computer vision, 2021, pp. 16259–16268.

[12] D. Lu, Q. Xie, K. Gao, L. Xu, J. Li, 3dctn: 3d convolution-transformer network for point cloud classification, IEEE Transactions on Intelligent Transportation Systems 23 (12) (2022) 24854–24865.

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

[14] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, S.-M. Hu, Pct: Point cloud transformer, Computational Visual Media 7 (2021) 187–199.

[15] X. Yang, M. Jin, W. He, Q. Chen, Pointcat: Cross-attention transformer for point cloud, arXiv preprint arXiv:2304.03012 (2023).

[16] C. Zhang, H. Wan, X. Shen, Z. Wu, Patchformer: An efficient point transformer with patch attention, in: IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11799–11808.

[17] Z. Liu, X. Yang, H. Tang, S. Yang, S. Han, Flatformer: Flattened window attention for efficient point cloud transformer, in: IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 1200–1211.

[18] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, arXiv preprint arXiv:2312.00752 (2023).

[19] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision mamba: Efficient visual representation learning with bidirectional state space model, in: International Conference on Machine Learning, OpenReview.net, 2024, pp. 1–14.

[20] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, Y. Qiao, Videomamba: State space model for efficient video understanding, in: European Conference on Computer Vision, Springer, 2025, pp. 237–255.

[21] X. Zhang, S. Wu, Q. Wang, K. Wang, Effretmamba: A light-weight mamba for low-light image enhancement, Circuits Syst. Signal Process. 44 (11) (2025) 8691–8711.

[22] Z. Xing, T. Ye, Y. Yang, G. Liu, L. Zhu, Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 578–588.

[23] Z. Zhang, A. Liu, I. Reid, R. Hartley, B. Zhuang, H. Tang, Motion mamba: Efficient and long sequence motion generation, in: European Conference on Computer Vision, Springer, 2025, pp. 265–282.

[24] D. Liang, X. Zhou, W. Xu, X. Zhu, Z. Zou, X. Ye, X. Tan, X. Bai, Pointmamba: A simple state space model for point cloud analysis, in: Advances in Neural Information Processing Systems, 2024, pp. 32653–32677.

[25] T. Zhang, H. Yuan, L. Qi, J. Zhang, Q. Zhou, S. Ji, S. Yan, X. Li, Point cloud mamba: Point cloud learning via state space model, in: AAAI Conference on Artificial Intelligence, 2025, pp. 10121–10130.

[26] J. Liu, R. Yu, Y. Wang, Y. Zheng, T. Deng, W. Ye, H. Wang, Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy, arXiv preprint arXiv:2403.06467 (2024).

[27] K. A. Tychola, E. Vrochidou, G. A. Papakostas, Deep learning based computer vision under the prism of 3d point clouds: a systematic review, The Visual Computer 40 (11) (2024) 8287–8329.

[28] N. Köprücü, D. Okpekpe, A. Orvieto, Nimba: Towards robust and principled processing of point clouds with ssms, arXiv preprint arXiv:2411.00151 (2024).

[29] W. Wu, Z. Qi, L. Fuxin, Pointconv: Deep convolutional networks on 3d point clouds, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9621–9630.

[30] P. Zhou, L. An, Y. Wang, G. Geng, Mlgtm: Multi-scale local geometric transformer-mamba application in terracotta warriors point cloud classification, Remote Sensing 16 (16) (2024) 2920.

[31] W. Yang, L. Xie, W. Qian, C. Wu, H. Yang, Coarse-to-fine cascaded 3d hand reconstruction based on ssgc and mhsa, The Visual Computer 41 (1) (2025) 11–24.

[32] J. Wu, H. Wu, G. Yuan, Detail-aware image denoising via structure preserved network and residual diffusion model, The Visual Computer 41 (1) (2025) 639–658.

[33] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, O. Litany, Pointcontrast: Unsupervised pre-training for 3d point cloud understanding, in: European Conference on Computer Vision, Springer, 2020, pp. 574–591.

[34] Z. Zhang, R. Girdhar, A. Joulin, I. Misra, Self-supervised pretraining of 3d features on any point-cloud, in: IEEE/CVF international conference on computer vision, 2021, pp. 10252–10263.

[35] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, J. Lu, Point-bert: Pre-training 3d point cloud transformers with masked point modeling, in: IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 19313–19322.

[36] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, L. Yuan, Masked autoencoders for point cloud self-supervised learning, in: European conference on computer vision, Springer, 2022, pp. 604–621.

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

[38] G. Chen, M. Wang, Y. Yang, K. Yu, L. Yuan, Y. Yue, Pointgpt: Auto-regressively generative pre-training from point clouds, in: Advances in Neural Information Processing Systems, 2023, pp. 29667 – 29679.

[39] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3d shapenets: A deep representation for volumetric shapes, in: IEEE conference on computer vision and pattern recognition, 2015, pp. 1912–1920.

[40] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, S.-K. Yeung, Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data, in: IEEE/CVF international conference on computer vision, 2019, pp. 1588–1597.

[41] L. Yi, V. G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, L. Guibas, A scalable active framework for region annotation in 3d shape collections, ACM Transactions on Graphics (ToG) 35 (6) (2016) 1–12.

[42] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al., Shapenet: An information-rich 3d model repository, arXiv preprint arXiv:1512.03012 (2015).

[43] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, B. Chen, Pointcnn: Convolution on x-transformed points, in: Advances in Neural Information Processing Systems, 2018, pp. 828–838.

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

[45] S. Qiu, S. Anwar, N. Barnes, Dense-resolution network for point cloud classification and segmentation, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3813–3822.

[46] A. Goyal, H. Law, B. Liu, A. Newell, J. Deng, Revisiting point cloud shape classification with a simple and effective baseline, in: International conference on machine learning, PMLR, 2021, pp. 3809–3820.

[47] S. Qiu, S. Anwar, N. Barnes, Geometric back-projection network for point cloud classification, IEEE Transactions on Multimedia 24 (2021) 1943–1955.

[48] S. Cheng, X. Chen, X. He, Z. Liu, X. Bai, Pra-net: Point relation-aware network for 3d point cloud analysis, IEEE Transactions on Image Processing 30 (2021) 4436–4448.

[49] A. Hamdi, S. Giancola, B. Ghanem, Mvtn: Multi-view transformation network for 3d shape recognition, in: IEEE/CVF international conference on computer vision, 2021, pp. 1–11.

[50] Y. Liu, B. Tian, Y. Lv, L. Li, F.-Y. Wang, Point cloud classification using content-based transformer via clustering in feature space, IEEE/CAA Journal of Automatica Sinica 11 (1) (2023) 231–239.

[51] A. Bahri, M. Yazdanpanah, M. Noori, S. Dastani, M. Cheraghalikhani, G. A. V. Hakim, D. Osowiechi, F. Beizaee, I. Ben Ayed, C. Desrosiers, Spectral informed mamba for robust point cloud processing, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 11799–11809.

[52] Z. Wang, Z. Chen, Y. Wu, Z. Zhao, L. Zhou, D. Xu, Pointramba: A hybrid transformer-mamba framework for point cloud analysis, arXiv preprint arXiv:2405.15463 (2024).

[53] H. Liu, M. Cai, Y. J. Lee, Masked discrimination for self-supervised learning on point clouds, in: European Conference on Computer Vision, Springer, 2022, pp. 657–675.

[54] L. Diao, S. Song, Y. Qian, D. Ren, Zigzagpointmamba: Spatial-semantic mamba for point cloud understanding, arXiv preprint arXiv:2505.21381 (2025).

[55] X. Ma, C. Qin, H. You, H. Ran, Y. Fu, Rethinking network design and local geometry in point cloud: A simple residual MLP framework, in: International Conference on Learning Representations, OpenReview.net, 2022, pp. 1–14.

[56] Z. Wang, X. Yu, Y. Rao, J. Zhou, J. Lu, P2P: tuning pre-trained image models for point cloud analysis with point-to-pixel prompting, in: Advances in Neural Information Processing Systems, 2022, pp. 14388–14402.

[57] B. Chen, Y. Xia, Y. Zang, C. Wang, J. Li, Decoupled local aggregation for point cloud learning, arXiv preprint arXiv:2308.16532 (2023).

[58] J. Park, S. Lee, S. Kim, Y. Xiong, H. J. Kim, Self-positioning point-based transformer for point cloud understanding, in: IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 21814–21823.

[59] R. Thilakarathna, M. Phlernjai, Finger tracking for wearable vr glove using flexible rack mechanism, Virtual Reality & Intelligent Hardware 7 (1) (2025) 1–25.

[60] Y. Chen, M. Jiang, C. Xia, et al., A novel deep learning system for stemi prognostic prediction from multi-sequence cardiac magnetic resonance, Science Bulletin 70 (2025) 4241–4252.

[61] J. Wang, X. Lu, M. Bennamoun, B. Sheng, Non-rigid point cloud registration via anisotropic hybrid field harmonization, IEEE Transactions on Pattern Analysis and Machine Intelligence 47 (9) (2025) 7898–7914.



Yang Yu received the Ph.D. degree from the School of Computer and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China, in 2022. He is currently a Lecturer with the School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu, China.His research interests include remote sensing image processing, point cloud modeling, computer vision, and machine learning.



Zengyi Zhang received the B.Eng. degree in Software Engineering from Southwest Petroleum University, Chengdu, China, in 2025. He is currently pursuing the M.S. degree in Software Engineering at Southwest Petroleum University, Chengdu, China. His research interests include computer vision, machine learning, and point cloud processing.



Qiang Zeng received the B.Eng. degree in space science and technology from Chengdu University of Technology, Chengdu, China, in 2022. He is currently pursuing the M.S. degree at the School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu, China. His research interests include computer vision, machine learning, and point cloud processing.



Qijun Hu received the Ph.D. degree in Road and Railway Engineering from Southwest Jiaotong University, Chengdu, China, in 2008. He is currently a Professor with the School of Civil Engineering, Southwest Jiaotong University, Chengdu, China. His research interests include remote sensing image processing, point cloud processing, computer vision, and machine learning.



Xingpeng Zhang received the Ph.D. degree in Software Engineering from Chongqing University, Chongqing, China, in 2020. He is currently a Lecturer with the School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu, China. His research interests include computer vision, medical and remote sensing image analysis, natural language processing



Leping He received the Ph.D. degree in Underground Engineering from Southwest Petroleum University, Chengdu, China, in 2022. She is currently an Associate Professor with the School of Civil Engineering and Geomatics, Southwest Petroleum University, Chengdu, China. Her research interests include point cloud processing and LiDAR remote sensing.



Yang Yu received the Ph.D. degree in Transportation Planning and Management from Southwest Jiaotong University, Chengdu, China, in 2023. He is currently a Lecturer with the School of Civil Engineering and Geomatics, Southwest Petroleum University, Chengdu, China. His research interests include remote sensing image processing and point cloud processing.



Yu Bai received the Ph.D. degree in Mechanical Engineering from Southwest Petroleum University, Chengdu, China, in 2024. She is currently a Lecturer with the School of Civil Engineering and Geomatics, Southwest Petroleum University, Chengdu, China. Her research interests include point cloud processing and LiDAR remote sensing.