POINT-FOCUSED ATTENTION MEETS CONTEXT-SCAN STATE SPACE: ROBUST BIOLOGICAL VISUAL PERCEPTION FOR POINT CLOUD REPRESENTATION

Kanglin Qu[1,2], **Pan Gao**[1,2*], **Qun Dai**[1,2*] **and Yuanhao Sun**[3]
[1]College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics
[2]Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education
[3]School of Mathematical Sciences, Beijing University of Posts and Telecommunications
klinqu@163.com,{Pan.Gao,daiqun}@nuaa.edu.cn,sunyh@bupt.edu.cn

ABSTRACT

Synergistically capturing intricate local structures and global contextual dependencies has become a critical challenge in point cloud representation learning. To address this, we introduce PointLearner, a point cloud representation learning network that closely aligns with biological vision which employs an active, foveation-inspired processing strategy, thus enabling local geometric modeling and long-range dependency interactions simultaneously. Specifically, we first design a point-focused attention, which simulates foveal vision at the visual focus through a competitive normalized attention mechanism between local neighbors and spatially downsampled features. The spatially downsampled features are extracted by a pooling method based on learnable inducing points, which can flexibly adapt to the non-uniform distribution of point clouds as the number of inducing points is controlled and they interact directly with point clouds. Second, we propose a context-scan state space that mimics eye's saccade inference, which infers the overall semantic structure and spatial content in the scene through a scan path guided by the Hilbert curve for the bidirectional S6. With this focus-then-context biomimetic design, PointLearner demonstrates remarkable robustness and achieves state-of-the-art performance across multiple point cloud tasks. The code is available at https://github.com/Point-Cloud-Learning/PointLearner.

1 INTRODUCTION

As a fundamental data form in 3D vision, point clouds have been widely applied in numerous tasks such as autonomous driving, robot navigation, and augmented reality due to their ability to precisely represent the geometric structure and spatial details of objects (Yan et al., 2024a; Zhou et al., 2024a; Yan et al., 2024b; An et al., 2025; Resani & Nasihatkon, 2025; Zhou et al., 2025b; He et al., 2024; Xu et al., 2024; Zhou et al., 2025a; Zhang et al., 2023; 2024a; Liang et al., 2025b; 2026). Currently, local attention networks (Zhao et al., 2021; Wu et al., 2024a) represent the mainstream paradigm for point cloud representation learning. By ingeniously computing attention within local neighbors/windows, they successfully reduce computational complexity to a linear relationship with the number of input points. However, such networks inevitably narrow the perceptual field, sacrificing the global perception capability of the attention mechanism, thereby hindering the effective modeling of long-range dependencies between objects in a scene. Recently, inspired by the exceptional long-range modeling capability with linear complexity of the selective state space model (S6) in Mamba (Gu & Dao, 2023), several studies (Zhang et al., 2024b; Han et al., 2024; Köprücü et al., 2024; Schöne et al., 2024; Wang et al., 2024) have attempted to introduce it into point cloud representation learning to overcome the trade-off between long-range interactions and computational resources. However, the bidirectional S6 still relies on compressing all contextual information into the history-hidden state for global connectivity, resulting in insufficient locality learning. Based on the above analysis, how to synergistically capture local fine-grained structures and global contextual dependencies has become a critical challenge in point cloud representation learning.

[*] Corresponding author.

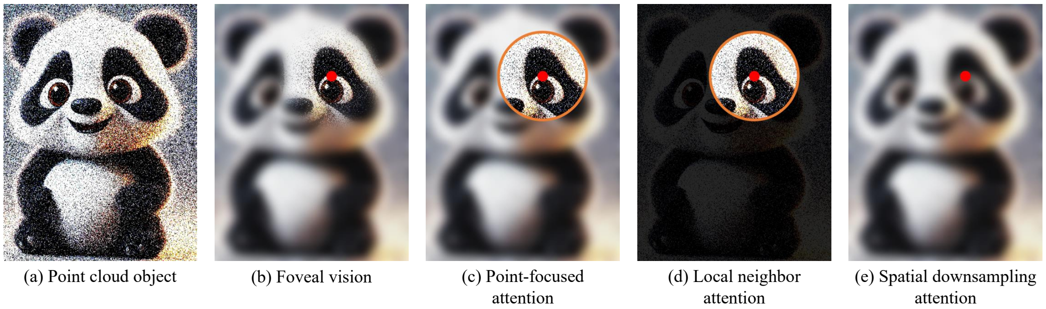

(a) Point cloud object
(b) Foveal vision
(c) Point-focused attention
(d) Local neighbor attention
(e) Spatial downsampling attention

Figure 1: A comparison of different information perception ways, where the red dot indicates the visual focus and the orange circle denotes the local awareness range. Our point-focused attention aligns more closely with natural foveal vision than the perceptual modes in (d) and (e).

In the biological visual system, foveal vision is responsible for perceiving the most significant information within the visual field, exhibiting pronounced spatial non-uniformity: as shown in Fig. 1(b), the region near the visual focus possesses extremely high acuity, enabling fine discrimination of detailed features; while visual acuity decreases with increasing eccentricity, resulting in coarser processing of peripheral features. This pattern not only optimizes the utilization of finite neural resources but also aligns with the intrinsic structure of information distribution in natural scenes (Wandell, 1995). What is more, biological vision is not a static process rather than a dynamic one that acquires information on a series of serialized visual foci via eye's continuous saccade movements, thereby inferring the overall semantic structure and spatial content within a scene (Stewart et al., 2020). The above operational mechanism endows the biological visual system with powerful perceptual capability to synergistically model local geometries and long-range dependencies.

Inspired by this, we explore PointLearner, a bionic-designed network that closely aligns with biological vision, enabling local geometric modeling and global context awareness simultaneously. In our network, the proposed point-focused attention emulate foveal vision perception at the visual focus, and context-scan state space is used to implement inference process during eye saccade:

- The point-focused attention adopts a dual-branch design, where one branch performs fine-grained attention modeling for each point's local neighbors, while the other branch establishes coarse-grained attention relationships between each point and spatially downsampled features. By computing attention weights from both branches within a single softmax calculation, the point-focused attention couples coarse- and fine-grained features in a competitive normalized manner with linear complexity. This makes each point adaptively and efficiently fuse local structures with global semantics, aligning with the intrinsic information distribution in natural scenes — like foveal vision perception. Notably, we develop an induced point pooling method, which is able to flexibly adapt to the non-uniform distribution of point clouds, through trainable vectors (dubbed inducing points) directly performing attention interactions with data points, as well as a controllable number of inducing points, thereby effectively downsampling spatial features.
- The context-scan state space utilizes the Hilbert curve for serialization of the point-focused attention feature, and further employs the the bidirectional S6 for geometric inference. As demonstrated in the eye tracking machine vision experiments (Newport et al., 2023), the Hilbert curve is more aligined with eye's saccade inference, having the property of locality preservation and the peculiarity of its self-similar rotating replication. By employing a biological inspired focus-then-context pipeline, the overall semantic structure and fine-grained spatial content can be efficiently integrated for point cloud representation learning.

To validate the effectiveness of PointLearner, we conduct extensive experiments on multiple standard point cloud datasets, including ModelNet40 (Wu et al., 2015), ScanObjectNN (Uy et al., 2019), ShapeNet (Yi et al., 2016), and S3DIS (Armeni et al., 2016). Experimental results demonstrate that our network, by emulating the operational mechanism of biological vision, effectively captures detail-rich local structures and coherent global context. It exhibits remarkable robustness and point cloud representation capability, yielding state-of-the-art results across multiple point cloud tasks.

In summary, the contributions of this paper stem from the following aspects:

(1) We propose PointLearner for point cloud learning, a bottom-to-up framework that aligns with biological vision, which achieves local refinement modeling and long-range dependency interactions by natural visual perception. This bio-inspired network attains state-of-the-art results on various point cloud tasks and demonstrates significant robustness.

(2) We design a point-focused attention simulating foveal vision perception at the visual focus. By computing attention weights for a point's local neighbors and spatially downsampled features within a single softmax calculation, it enables each point to adaptively and efficiently select the most effective receptive field information from both local structures and global semantics via a competitive normalized attention mechanism with linear complexity.

(3) We introduce a context-scan state space mimicking eye's saccade inference. By the Hilbert curve with excellent locality-preserving property, it guides the bidirectional S6 to accurately infer the entire scene along a scanning path that maintains high-fidelity spatial proximity between points.

2 RELATED WORK

2.1 ATTENTION-BASED NETWORKS

The attention mechanism (Vaswani et al., 2017; Shi, 2024; Su et al., 2024) has been widespread in point cloud representation learning, due to its ability to enable dynamic interactions between elements and global modeling. Several studies have enhanced the performance of the attention mechanism in point cloud tasks by refining attention modules (Guo et al., 2021; Mazur & Lempitsky, 2021; Yan et al., 2020) or designing pre-training strategies (Chen et al., 2023; Yu et al., 2022; Pang et al., 2022; Liu et al., 2022; Qi et al., 2023). Although these global attention networks have achieved impressive results, they perform attention computations directly on the entire point cloud, regarding each point as a token. This incurs prohibitive computational overheads due to the quadratic complexity of the attention mechanism and the large number of points in point clouds. Thus, some works ingeniously design local attention networks, which can be categorized into local neighbor-based and window-based methods. Local neighbor-based methods (Zhao et al., 2021; Wu et al., 2022; Nie et al., 2022; Xiang et al., 2023; Zhang et al., 2022; Liu et al., 2024b) apply the attention mechanism to point neighborhoods constructed for each point using neighbor search techniques such as the K-Means, ball query, and K-nearest neighbors (KNN), while window-based methods (Lai et al., 2022; Park et al., 2022; He et al., 2022; Fan et al., 2022; Sun et al., 2022; Liu et al., 2023b; Wu et al., 2024a) partition the 3D space into non-overlapping windows through voxelization or space-filling curves, transforming attention computations on all points to these spatial windows. Although local attention networks exhibit linear complexity, setting on the number of neighbors or window size constrains the receptive field of the attention mechanism, hindering the modeling of long-range dependencies. Departing from local neighbors, this paper incorporates a spatial downsampling branch and context-scan state space to extract global dependencies with the biological visual mechanism, overcoming the trade-off between global modeling and computational resources.

2.2 SSM-BASED NETWORKS

With excellent long-range modeling capability with linear complexity, the state space model (SSM) (Gupta et al., 2022; Gu et al., 2022; Smith et al., 2023; Mehta et al., 2023; Gu et al., 2020; 2021; Gu & Dao, 2023) has gained significant prominence in the field of natural language processing (NLP). Among these, Mamba (Gu & Dao, 2023) stands out as the most influential work. Its core component called S6 introduces an input-driven selective mechanism, enabling flexible selection of relevant information and achieving breakthrough performance in long sequence modeling. Furthermore, S6 incorporates a hardware-aware algorithm inspired by FlashAttention (Dao et al., 2022), significantly improving both training and inference efficiencies. These outstanding properties have motivated the extension of S6 from NLP to computer vision, including image recognition (Liu et al., 2025; Shaker et al., 2025; Fu et al., 2025; He et al., 2025), video understanding (Chen et al., 2025; Wang et al., 2023; Li et al., 2024a), and medical image segmentation (Xing et al., 2024; Ma et al., 2024; Ruan & Xiang, 2024). Recently, S6 has also been applied to point cloud representation learning. To adapt to the causal nature and unidirectional modeling of S6, existing methods design serializations strategies based on space-filling curves (Liang et al., 2024; Li et al., 2025; Liu et al., 2024a) or axis ordering (Zhang et al., 2024b; Han et al., 2024; Köprücü et al., 2024; Schöne et al., 2024)

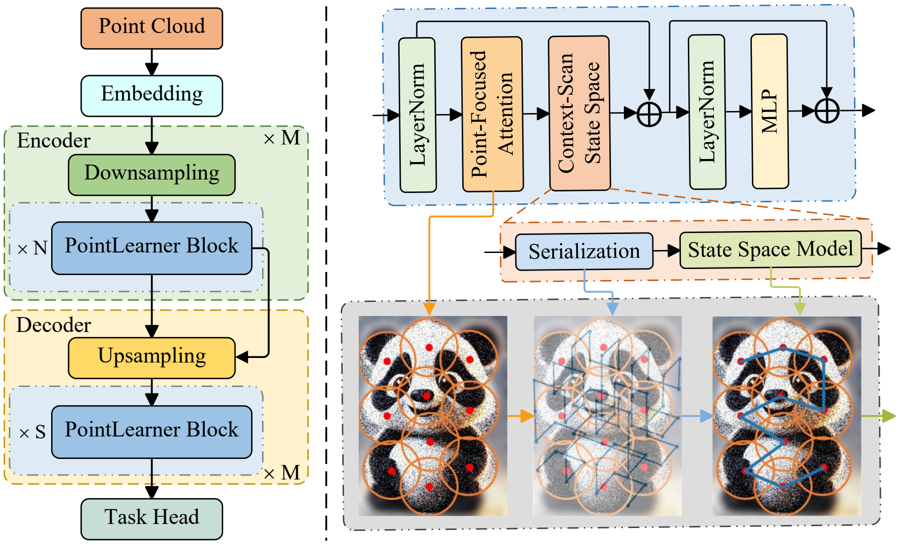


Figure 2: **Left**: Pipeline of PointLearner. **Right**: Architecture of PointLearner block, where the line between the red dots represent the saccade path guided by the serialization, which is used for geometric inference by the state space model.

to establish inter-point structural dependencies for the bidirectional S6's geometric inference. However, the bidirectional S6 still relies on compressing all context information into the history-hidden state to achieve global connectivity, leading to insufficient locality learning. Our work captures local structure information by a point-focused attention that matches foveal vision at the visual focus, addressing the above shortcoming.

3 Methodology

3.1 Overview

As shown in Fig. 2, PointLearner follows the standard Point Transformer-style architecture (Zhao et al., 2021; Wu et al., 2024a). A point cloud is first fed into an embedding layer formed by an MLP to be projected into a high-dimensional space, followed by an encoder-decoder structure that performs residual-based hierarchical feature aggregation, where the downsampling and upsampling layers use the Farthest Point Sampling (FPS) and linear interpolation (Qi et al., 2017b), respectively, and finally an appropriate task head is invoked based on specific requirements. In this study, our network is validated on point cloud recognition and segmentation tasks, where the recognition head first applies average pooling to the encoder's output, then produces global category logits by an MLP; the segmentation head processes the decoder's output with an MLP to predict per-point category logits. As the core component of the encoder-decoder structure, PointLearner block adopts the Transformer architecture for flexible integration into the network. By incorporating the point-focused attention first and then context-scan state space, it endows the network with information perception capability akin to the biological vision system, achieving local geometric modeling and long-range dependency interactions simultaneously. We will elaborate on these key modules next.

3.2 Point-focused attention

The point-focused attention adopts a dual-branch design to emulate foveal vision at the visual focus, where the local neighbor branch provides fine-grained perception near each query point, while the spatial downsampling branch simultaneously maintains each query point's coarse-grained awareness for global semantics, *i.e.*, high acuity at the focus and low acuity in the periphery.

Fine-grained perception. Given a point set $\boldsymbol{P} = \{\boldsymbol{p}_i = (x_i, y_i, z_i)\}_{i=1}^{N} \in \mathbb{R}^{N\times 3}$ with a corresponding point feature set $\boldsymbol{F} = \{\boldsymbol{f}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times D}$, the **L**ocal **N**eighbor **B**ranch on a point $\boldsymbol{p}_i$ is

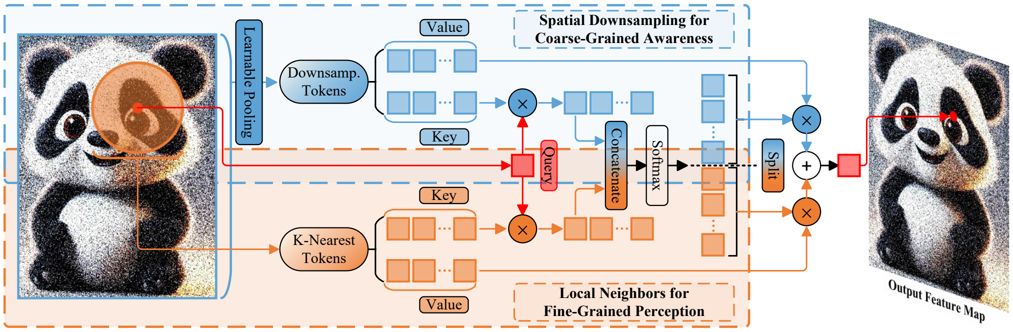


Figure 3: Diagram of point-focused attention with the competitive normalized attention mechanism.

computed as follows:

$$
\begin{aligned}
\left(\boldsymbol{Q}^l, \boldsymbol{K}^l, \boldsymbol{V}^l\right) &= \left(\boldsymbol{W}_q^l, \boldsymbol{W}_k^l, \boldsymbol{W}_v^l\right)\boldsymbol{F} \\
\boldsymbol{\mathcal{A}}_i^l &= softmax\left(\left\langle \boldsymbol{Q}_i^l, \boldsymbol{K}_{\boldsymbol{\mathcal{N}}_i}^l\right\rangle / \sqrt{D}\right) \\
\mathrm{LNB}\left(\boldsymbol{p}_i\right) &= \boldsymbol{\mathcal{A}}_i^l \boldsymbol{V}_{\boldsymbol{\mathcal{N}}_i}^l
\end{aligned}
\tag{1}
$$

where $\boldsymbol{W} \in \mathbb{R}^{D \times D}$ represents a transformation matrix, and $\boldsymbol{\mathcal{N}}_i \in \mathbb{R}^K$ denotes the indices of the local neighbors of a point $\boldsymbol{p}_i$ in $\boldsymbol{P}$, determined by KNN.

Coarse-grained awareness. Given a point feature set $\boldsymbol{F} = \{\boldsymbol{f}_i\}_{i=1}^N \in \mathbb{R}^{N \times D}$ and its corresponding spatially downsampled feature set $\boldsymbol{S} = \{\boldsymbol{s}_i\}_{i=1}^M \in \mathbb{R}^{M \times D}$, the **S**patial **D**ownsampling **B**ranch on a point $\boldsymbol{p}_i$ is computed as follows:

$$
\begin{aligned}
\left(\boldsymbol{Q}^s, \boldsymbol{K}^s, \boldsymbol{V}^s\right) &= \left(\boldsymbol{W}_q^s\boldsymbol{F}, \boldsymbol{W}_k^s\boldsymbol{S}, \boldsymbol{W}_v^s\boldsymbol{S}\right) \\
\boldsymbol{\mathcal{A}}_i^s &= softmax\left(\boldsymbol{Q}_i^s, \boldsymbol{K}^s / \sqrt{D}\right) \\
\mathrm{SDB}\left(\boldsymbol{p}_i\right) &= \boldsymbol{\mathcal{A}}_i^s \boldsymbol{V}^s
\end{aligned}
\tag{2}
$$

Notably, the non-uniformity of point clouds prevents them from achieving spatial downsampling through simple and effective average pooling like 2D images. Moreover, as a commonly used downsampling method for point clouds, FPS often requires setting a small sampling rate to attain sufficient coverage for extracting global information, which significantly increase computational cost in Eq. (2). Thus, we develop an induced point pooling inspired by the inducing point method in the sparse Gaussian (Snelson & Ghahramani, 2005). Specifically, M D-dimensional vectors defined as $\boldsymbol{I} \in \mathbb{R}^{M \times D}$ are termed inducing points, and they are trainable parameters. The **I**nduced **P**oint **P**ooling on a point feature set $\boldsymbol{F} = \{\boldsymbol{f}_i\}_{i=1}^N \in \mathbb{R}^{N \times D}$ is computed as follows:

$$
\begin{aligned}
\left(\boldsymbol{K}^p, \boldsymbol{V}^p\right) &= \left(\boldsymbol{W}_k^p, \boldsymbol{W}_v^p\right)\boldsymbol{F} \\
\boldsymbol{S} = \mathrm{IPP}\left(\boldsymbol{F}\right) &= softmax\left(\boldsymbol{I}, \boldsymbol{K}^p / \sqrt{D}\right)\boldsymbol{V}^p
\end{aligned}
\tag{3}
$$

where $\boldsymbol{S} \in \mathbb{R}^{M \times D}$ represents the spatially downsampled features extracted by the induced point pooling on $\boldsymbol{F}$. By using a controllable number of trainable inducing points to directly learn how to induct point clouds adaptively through attention interactions, the induced point pooling flexibly adapts to the non-uniform distribution of point clouds to integrate global semantics, thereby effectively downsampling spatial features, as shown in Appendix C.2.

Competitive normalized fusion. A straightforward approach for the point-focused attention is to sum the attention outputs from both branches to achieve the multi-scale fusion PFA $(\boldsymbol{p}_i)$ of the local structures and global semantics at a point $\boldsymbol{p}_i$, thereby matching foveal vision perception, as follows:

$$
\mathrm{PFA}\left(\boldsymbol{p}_i\right) = \mathrm{LNB}\left(\boldsymbol{p}_i\right) + \mathrm{SDB}\left(\boldsymbol{p}_i\right) = \boldsymbol{\mathcal{A}}_i^l \boldsymbol{V}_{\boldsymbol{\mathcal{N}}_i}^l + \boldsymbol{\mathcal{A}}_i^s \boldsymbol{V}^s
\tag{4}
$$

However, such *shallow* multi-scale feature fusion struggles to align with deep dynamic interactions between local fine-grained features and global coarse-grained semantics inherent in foveal vision perception. To address this, as illustrated in Fig. 3, we update the point-focused attention to a competitive normalized attention variant by computing the attention weights of both branches within a single softmax calculation, thereby coupling fine- and coarse-grained features, as follows:

$$\begin{gathered}\boldsymbol{\mathcal{A}}_i = softmax\left(Concat\left(\boldsymbol{Q}_i^l, \boldsymbol{K}_{\boldsymbol{\mathcal{N}}_i}^l, \boldsymbol{Q}_i^s, \boldsymbol{K}^s\right)/\sqrt{D}\right) \\ \boldsymbol{\mathcal{A}}_i^l, \boldsymbol{\mathcal{A}}_i^s = split\left(\boldsymbol{\mathcal{A}}_i, [K, M]\right) \\ \mathrm{PFA}\left(\boldsymbol{p}_i\right) = \mathrm{LNB}\left(\boldsymbol{p}_i\right) + \mathrm{SDB}\left(\boldsymbol{p}_i\right) = \boldsymbol{\mathcal{A}}_i^l \boldsymbol{V}_{\boldsymbol{\mathcal{N}}_i}^l + \boldsymbol{\mathcal{A}}_i^s \boldsymbol{V}^s\end{gathered} \tag{5}$$

where $Concat$ denotes channel-level concatenation, $split$ serves to partition channels according to specified sizes, and K and M represent the number of local neighbors determined by KNN and the number of inducing points used for downsampling spatial features, respectively. Compared to the simple version in Eq. (4), as demonstrated in Tab. 8, Eq. (5) enhances deep dynamic interactions through the competitive mechanism between fine- and coarse-grained features, without introducing additional computational overheads. This effectively simulates the perceptual process of foveal vision adaptively selecting the most effective receptive field information, thereby better aligning with the intrinsic structure of information distribution in natural scenes.

Complexity analysis. Based on the above settings and considering the feature transformation, the computational complexity $\Omega\left(\mathrm{PFA}\right)$ of the point-focused attention is as follows:

$$\begin{aligned}\Omega\left(\mathrm{PFA}\right) &= \Omega\left(\mathrm{LNB}\right) + \Omega\left(\mathrm{SDB}\right) + \Omega\left(\mathrm{IPP}\right) \\ &= 6ND^2 + 2MD^2 + 2NKD + 4NMD\end{aligned} \tag{6}$$

Since K and M are typically small, it can be observed that $\Omega\left(\mathrm{PFA}\right)$ scales linearly with the number of points, indicating that PFA aligns with the optimized utilization of finite neural resources in foveal visual. For the complexity of each module in PFA, please refer to Appendix D.

3.3 CONTEXT-SCAN STATE SPACE

The context-scan state space simulates eye's saccade inference through a dynamic saccade mechanism. It serializes a point cloud to provide a scanning path for the state space model, enabling the inference of the overall semantic structure and spatial content in a scene, *i.e.*, the serialization performs continuous scanning and the state space model is responsible for information integration, as illustrated in Fig. 2 (right).

Serialization. Space-filling curves are trajectories that cover high-dimensional regions by continuous parametric mapping. Their core function is to transform high-dimensional geometric structures into one-dimensional sequences while preserving local neighborhood relationships, meaning that spatially adjacent elements remain adjacent in the sequence. When applied to point clouds, the high-dimensional space refers to the 3D Euclidean space containing point coordinates. Common space-filling curves include the Hilbert curve and Z-Order curve. The former is highly valued for its superior locality-preserving property, while the latter is renowned for its high efficiency.

Inspired by the spatial proximity of space-filling curves, recent S6-based works (Liang et al., 2024; Li et al., 2025; Liu et al., 2024a) utilize them to serialize point clouds, establishing more reliable inter-point structural dependencies for geometric reasoning compared to axis-ordering based serialization (Zhang et al., 2024b; Han et al., 2024; Köprücü et al., 2024; Schöne et al., 2024). However, to provide richer spatial information, these serialization strategies often concatenate the results from multiple space-filling curves. This concatenation introduces a longer sequence, leading to redundancy and negatively impacting efficiency. More importantly, concatenating sequences with different spatial relationships can easily cause confusion. In addition, by serializing randomly located points using the Hilbert and Z-Order curves, respectively, it can be intuitively observed that the Hilbert curve outperforms the Z-Order curve in preserving spatial proximity, as shown in Appendix C.1. This finding is consistent with the prior research (Nordin & Telles, 2023) and also better aligns with the inherent pattern of eye movements during visual search, which typically involve continuous scanning along spatially adjacent regions. Therefore, the Hilbert curve is employed for serialization to establish reliable inter-point structural dependencies while guiding a high-fidelity spatially adjacent scanning path for accurate scene inference.

State space model. S6 is a forward recurrence process based on hidden states, where each position in input sequence can only access prior information and cannot obtain content from subsequent positions, as shown in Appendix A. This unidirectional modeling is unsuitable for visual data requiring

Network	Operator	OA
†IDPT (Zha et al., 2023)	Attention	93.4
†Inter-MAE (Liu et al., 2023a)	Attention	93.6
†CrossNet (Wu et al., 2023b)	Attention	93.4
†ACT (Dong et al., 2023)	Attention	93.7
LFT-Net (Gao et al., 2023)	Attention	93.2
OctFormer (Wang, 2023)	Attention	92.7
†ReCon (Qi et al., 2023)	Attention	92.5
†DAPT (Zhou et al., 2024b)	Attention	93.5
†LCM (Zha et al., 2024)	Attention	93.6
†Point-PEFT (Tang et al., 2024)	Attention	93.4
PointStack (Wijaya et al., 2024)	Attention	93.3
GAD (Li et al., 2024b)	Attention	93.8
PointConT (Liu et al., 2024b)	Attention	93.5
†PointGST (Liang et al., 2025a)	Attention	93.4
†PointMamba (Liang et al., 2024)	SSM	93.6
OctMamba (Liu et al., 2024a)	SSM	92.7
PCM (Zhang et al., 2024b)	SSM	93.4
†Mamba3D (Han et al., 2024)	SSM	93.4
NIMBA (Köprücü et al., 2024)	SSM	92.1
STREAM (Schöne et al., 2024)	SSM	92.7
PoinTramba (Wang et al., 2024)	Hybrid	92.7
PointLearner	Hybrid	**94.2**

Table 1: Experimental results on ModelNet40 dataset. †: Pre-training strategy.

Network	Operator	Ins. mIoU
APES (Wu et al., 2023a)	Attention	85.8
†ACT (Dong et al., 2023)	Attention	86.1
†PointGPT (Chen et al., 2023)	Attention	86.2
†ReCon (Qi et al., 2023)	Attention	86.4
†Point2Vec (Zeid et al., 2023)	Attention	86.3
†IDPT (Zha et al., 2023)	Attention	85.7
GAD (Li et al., 2024b)	Attention	86.3
†MaskFeat3D (Yan et al., 2024b)	Attention	86.3
†MVNet (Yan et al., 2024a)	Attention	86.1
†LCM (Zha et al., 2024)	Attention	86.3
†DAPT (Zhou et al., 2024b)	Attention	85.5
†Point-PEFT (Tang et al., 2024)	Attention	85.1
†PointGST (Liang et al., 2025a)	Attention	85.7
†Mamba3D (Han et al., 2024)	SSM	85.6
†PointMamba (Liang et al., 2024)	SSM	86.2
PCM (Zhang et al., 2024b)	SSM	84.3
NIMBA (Köprücü et al., 2024)	SSM	85.5
PoinTramba (Wang et al., 2024)	Hybrid	85.7
PointLearner	Hybrid	**86.9**

Table 2: Experimental results on ShapeNet dataset. †: Pre-training strategy.

global learning. To address this, most works (Li et al., 2025; Han et al., 2024; Köprücü et al., 2024; Wang et al., 2024) propose the bidirectional S6, by introducing the bidirectionality from Vision Mamba (Zhu et al., 2024) into S6, to achieve global modeling over point sequences. Specifically, two S6 modules are deployed in parallel: a forward S6 and a backward S6. The former performs a forward-scanning recurrent along the input sequence, while the latter processes it in reverse order. In this way, each point in the input sequence possesses a global receptive field, which aligns with how the eye perform back-and-forth scanning to infer information when recognizing an indistinct object. Hence, we adopt the bidirectional S6 for scene inference.

4 Experiments

To validate PointLearner, we conduct experimental comparisons on multiple point cloud tasks. Additionally, we perform robustness checking, and explore the effectiveness and characteristics of these biomimetic designs through extensive ablation studies. For detailed descriptions on datasets and evaluation metrics, please refer to Appendix B.

4.1 Experimental comparisons

Object recognition. Table 1 lists the quantitative results of our network and recent works on ModelNet40 dataset. It can be observed that the performance of the previous state-of-the-art attention networks has been saturated, confined to a narrow range of 93.2% to 93.8%. Our PointLearner breaks through this performance bottleneck, achieving a state-of-the-art 94.2% OA. This result demonstrates that integrating the advantages of both attention and SSM paradigms within a biological visual system framework is an effective avenue to advancing point cloud representation learning.

Part segmentation. Table 2 lists the quantitative results of our network and recent works on ShapeNet dataset. Overall, the performance of emerging SSM-based methods confirms their limitations in handling fine-grained local features. The proposed PointLearner achieves 86.9% Ins. mIoU, significantly outperforming existing state-of-the-art methods based on either attention or SSM. This strongly demonstrates that the hybrid paradigm of the attention and SSM, guided by the biological vision system, can effectively synergize the strengths of both operators, exhibiting powerful capabilities in local geometric modeling and long-range dependency interactions.

Semantic segmentation. Table 3 lists the quantitative results of our network and recent works on S3DIS dataset. In the more challenging task of point cloud semantic segmentation, the performance of the networks with different architectures exhibits significant disparities, highlighting the dual

Network	Operator	mIoU
PointVector (Deng et al., 2023)	Attention	72.3
SPT (Robert et al., 2023)	Attention	68.9
SpoTr (Park et al., 2023)	Attention	70.8
†ACT (Dong et al., 2023)	Attention	61.2
†ReCon (Qi et al., 2023)	Attention	60.8
†IDPT (Zha et al., 2023)	Attention	53.1
†MM-3Dscene (Xu et al., 2023)	Attention	71.9
Retro-FPN (Xiang et al., 2023)	Attention	73.0
PTv3 (Wu et al., 2024a)	Attention	73.4
KPConvX-L (Thomas et al., 2024)	Attention	73.5
†DAPT (Zhou et al., 2024b)	Attention	56.2
†Point-PEFT (Tang et al., 2024)	Attention	56.0
†MVNet (Yan et al., 2024a)	Attention	73.8
GAD (Li et al., 2024b)	Attention	62.9
†Swin3D (Yang et al., 2025)	Attention	72.5
†PointGST (Liang et al., 2025a)	Attention	58.6
PCM (Zhang et al., 2024b)	SSM	63.4
HydraMamba (Qu et al., 2025)	SSM	73.6
Pamba (Li et al., 2025)	SSM	73.5
PointLearner	Hybrid	**74.3**

Table 3: Experimental results on S3DIS dataset. †: Pre-training strategy.

Network	Operator	OA
ADS (Hong et al., 2023)	Attention	87.5
†IDPT (Zha et al., 2023)	Attention	84.9
†ACT (Dong et al., 2023)	Attention	88.2
†Joint-MAE (Guo et al., 2023)	Attention	86.1
SpoTr (Park et al., 2023)	Attention	88.6
†LCM (Zha et al., 2024)	Attention	87.8
†Inter-MAE (Liu et al., 2023a)	Attention	85.4
†PointGPT (Chen et al., 2023)	Attention	86.9
†PointDif (Zheng et al., 2024)	Attention	87.6
†DAPT (Zhou et al., 2024b)	Attention	85.1
GAD (Li et al., 2024b)	Attention	82.6
†Point-PEFT (Tang et al., 2024)	Attention	85.0
†MaskFeat3D (Yan et al., 2024b)	Attention	87.7
†MVNet (Yan et al., 2024a)	Attention	86.7
†PointGST (Liang et al., 2025a)	Attention	85.6
†PointMamba (Liang et al., 2024)	SSM	89.3
PCM (Zhang et al., 2024b)	SSM	88.1
†Mamba3D (Han et al., 2024)	SSM	88.2
NIMBA (Köprücü et al., 2024)	SSM	84.2
STREAM (Schöne et al., 2024)	SSM	85.3
PoinTramba (Wang et al., 2024)	Hybrid	88.9
PointLearner	Hybrid	**89.8**

Table 4: Experimental results on ScanObjectNN dataset. †: Pre-training strategy.

challenges of local geometric modeling and global contextual inference in this task. Our network attains state-of-the-art performance with 74.3% mIoU, which benefits from its bio-inspired visual perception design, maintaining the sensitivity of the attention mechanism to both local geometries and global semantics while incorporating the scene inference capability of SSM.

4.2 Robustness checking

Biological vision exhibits remarkable robustness in to low-quality scenes. To verify that our biomimetic network inherits this property, we conduct experiments from two perspectives: strong noise corruption and varying sampling densities.

Robustness to strong noise. ScanObjectNN is a challenging dataset collected from real-world scenes. We conduct experiments on its most difficult variant, PB_T50_RS, to validate PointLearner's robustness to strong noise. Tab. 4 lists the quantitative results of our network and recent works on ScanObjectNN dataset. The performance of the hybrid architecture PoinTramba (88.9% OA) has demonstrated the potential of hybrid methods in terms of robustness. Our PointLearner further elevates the performance to 89.8% OA, surpassing all the existing models. This result indicates that the synergistic mechanism of the attention and SSM adopted by PointLearner effectively inherits the robust inference capability of the biological vision system under low-quality perception conditions.

Robustness to varying sampling densities. Sensor data captured directly from the real world often suffers from severe irregular sampling issues. Consequently, during the testing phase, we randomly discard data points, as shown in Fig. 4 (left), to validate our network's robustness to non-uniformly sparse data. Fig. 4 (right) presents the quantitative results of our network alongside the top-performing attention method (GAD (Li et al., 2024b)) and SSM method (PCM (Zhang et al., 2024b)) from Tab. 1 on ModelNet40 dataset with varying numbers of sampling points. Intuitively, our network exhibits superior robustness to variations in sampling density compared to the other methods, with only a 2.2% performance drop when the number of test points decreases from 1024 to 256. This result benefits from the competitive normalized attention in the point-focused attention, which balances descriptiveness and robustness by appropriately weighting local structures and global perception, as well as the powerful global inference capability of the context-scan state space.

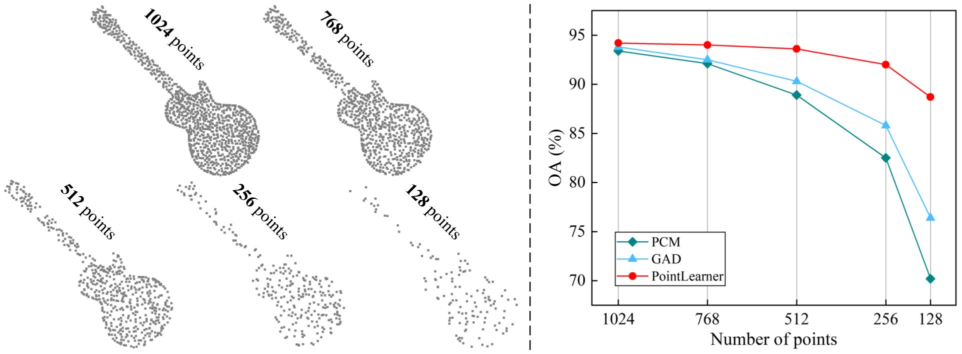


Figure 4: **Left**: Point clouds with different point densities. **Right**: Quantitative results of our network and other works on ModelNet40 dataset with different numbers of sampling points.

Networks	Operator	Params	Latency	Memory	mIoU
HydraMamba (Qu et al., 2025)	SSM	63.14M	54ms	**5.9G**	73.6
PTv3 (Wu et al., 2024a)	Attention	46.17M	**49ms**	6.3G	73.4
Swin3D (Yang et al., 2025)	Attention	71.15M	365ms	10.7G	72.5
PointLearner	Hybrid	52.78M	63ms	6.5G	**74.3**

Table 5: Params, latency, and memory footprint of our network and previous state-of-the-art methods in a single inference on S3DIS dataset.

4.3 Efficiency analysis

To further analyze the computational overheads of PointLearner, we compared it with several previous state-of-the-art methods on S3DIS dataset, and the params, latency, and memory footprint in a single inference are selected as evaluation metrics. Specifically, to ensure a fair comparison, the latency and memory footprint in a single inference are taken as the average values obtained over the entire S3DIS test set with each scene on the same RTX 4090 GPU. Tab. 5 presents the params, latency, and memory footprint of our network and multiple previous state-of-the-art methods in a single inference. Swin3D is a ***heavy*** attention network that, unlike PTv3, does not incorporate efficiency improvements for attention computations. Furthermore, it can be observed that, with a similar number of parameters, PointLearner, as a hybrid network, achieves a superior trade-off between computational overheads and performance compared to the fully optimized PTv3 and the HydraMamba that benefits from S6's excellent properties. We attribute this to the following two aspects: (1) PointLearner is able to achieve powerful local geometric modeling and long-range dependency interactions from the perspective of biological vision through hybrid operators, which simultaneously requires less layer stacking; (2) Although our basic block contains multiple components, they are essentially ***lightweight*** with linear complexity, while also benefiting from the hardware-optimized algorithms of FlashAttention (Dao et al., 2022).

4.4 Ablation study

We conduct ablation study on ModelNet40 dataset, to investigate the effectiveness of the designed bio-inspired visual processing modules. To ensure a fair comparison, all experiments are conducted on an RTX 4090 GPUs with identical configurations, and results are averaged over three runs.

Local neighbor branch. The point-focused attention employs the local neighbor branch to achieve the fine-grained perception, akin to foveal vision. To validate its necessity, we compare ablation results with and without the local neighbor branch in Tab. 6. The trade-off between accuracy and efficiency further indicates that the local neighbor branch enhances the network's ability to learn local geometric structures while maintaining low computational complexity. This closely aligns with the biological vision mechanism where refined local perception optimizes global inference.

Spatial downsampling branch. Through the spatial downsampling branch, the point-focused attention simultaneously maintains coarse-grained perception of global semantics for each query point. To validate its importance, we compare ablation results with and without this branch in Tab. 7.

LNB	Params	FLOPs	Throughput	OA
✓	7.36M	0.610G	163FPS	**94.17**
×	6.57M	0.504G	221FPS	92.11

Table 6: Ablation results w/ or w/o the local neighbor branch.

SDB	Params	FLOPs	Throughput	OA
✓	7.36M	0.610G	163FPS	**94.17**
×	6.35M	0.578G	183FPS	93.06

Table 7: Ablation results w/ or w/o the spatial downsampling branch.

Fusion	Params	FLOPs	Throughput	OA
Competitive	7.36M	0.610G	163FPS	**94.17**
Additive	7.36M	0.610G	166FPS	93.43

Table 8: Ablation results with different multi-scale fusion.

SSM	Params	FLOPs	Throughput	OA
Bid. S6	7.36M	0.610G	163FPS	**94.17**
Uni. S6	6.06M	0.553G	181FPS	93.08

Table 9: Ablation results with both state space models.

PFA	CSSS	Params	FLOPs	Throughput	OA
✓	×	4.57M	0.487G	198FPS	92.93
×	✓	5.37M	0.463G	231FPS	91.94
✓	✓	7.36M	0.610G	163FPS	**94.17**

Table 10: Ablation results with PFA and CSCC.

Quantitative analysis demonstrates that this branch effectively establishes coarse-grained associations between query points and global semantics at a low computational cost, simulating the integration mechanism of peripheral information in biological vision.

Normalized fusion. In the point-focused attention, by coupling attention weights within a single softmax calculation, the proposed competitive normalized fusion replaces simple additive multi-scale fusion to enhance deep dynamic interactions. To validate its advantage, Tab. 8 compares the ablation results of both multi-scale fusion strategies. Quantitative analysis demonstrates that the competitive fusion mechanism effectively simulates the dynamic interaction and adaptive selection between fine- and coarse-grained features in foveal vision. This achieves superior semantic fusion effect while maintaining nearly identical computational overheads.

State space model. The context-scan state space leverages the state space model to infer the entire scene along the scanning path provided by the serialization, based on inter-point structural dependencies. To validate that the adopted bidirectional S6 achieves stronger inference performance compared to the unidirectional S6, we compared the ablation results of both state space models. As shown in Tab. 9, although the bidirectional S6 introduces a slight increase in parameters and computational complexity, it constructs a global receptive field for each point through the forward and backward scanning. This effectively simulates the eye's back-and-forth saccade, thereby enhancing inference ability in complex scenes.

PFA & CSSS. The proposed network incorporates two key modules: Point-Focused Attention (PFA) and Contextual Scan State Space (CSSS). PFA fuses local neighbors and spatially downsampled features based on a competitive normalized attention mechanism, simulating foveal vision at the visual focus. On this basis, CSSS further infers the overall semantic structure and spatial content within a scene through point cloud serialization and state space model, mimicking eye's saccade inference. The combination of PFA and CSSS closely aligns with biological vision, and their complementarity is fully validated by the ablation results presented in Tab. 10.

5 Conclusion

In this paper, we introduce PointLearner, a point cloud representation learning network inspired by biological vision mechanisms. First, we propose a point-focused attention architecture that mimics the foveal vision of the human eye, enabling fine-grained perception around each query point while maintaining coarse-grained awareness of more distant regions. Building on this locally attentive representation, we further introduce a context-scan state space model to globally scan point clouds, drawing inspiration from the saccadic movements of the human eye during visual inference. This allows the model to integrate local details and global context, leading to a comprehensive understanding of the underlying geometric structure. Extensive experiments across multiple datasets and tasks demonstrate that PointLearner outperforms existing methods and exhibits strong robustness to noise and varying sampling densities.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China under Grant (62476126, 62272227).

REFERENCES

Z. C. An, G. L. Sun, Y. Liu, R. J. Li, M. Wu, M. M. Cheng, E. Konukoglu, and S. J. Belongie. Multimodality helps few-shot 3d point cloud semantic segmentation. In *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–15, Singapore, Apr. 2025.

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

G. Y. Chen, M. L. Wang, Y. Yang, K. Yu, L. Yuan, and Y. F. Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–13, New Orleans, LA, USA, Dec. 2023.

S. R. Chen, Y. X. Luo, Y. Ma, Y. Qiao, and Y. L. Wang. H-mba: Hierarchical mamba adaptation for multi-modal video understanding in autonomous driving. arXiv:2501.04302, 2025.

T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–16, New Orleans, LA, USA, Dec. 2022.

X. Deng, W. Y. Zhang, Q. Ding, and X. M. Zhang. Pointvector: A vector representation in point cloud analysis. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9455–9465, Vancouver, BC, Canada, Jun. 2023.

R. P. Dong, Z. K. Qi, L. F. Zhang, J. B. Zhang, J. J. Sun, Z. Ge, L. Yi, and K. S. Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–16, Kigali, Rwanda, May 2023.

L. Fan, Z. Q. Pang, T. Y. Zhang, Y. X. Wang, H. Zhao, F. Wang, N. Y. Wang, and Z. X. Zhang. Embracing single stride 3d object detector with sparse transformer. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8448–8458, New Orleans, LA, USA, Jun. 2022.

Y. X. Fu, M. Lou, and Y. Z. Yu. Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19077–19087, Nashville, TN, USA, Jun. 2025.

Y. B. Gao, X. B. Liu, J. Li, Z. J. Fang, X. Y. Jiang, and K. M. S. Huq. Lft-net: Local feature transformer network for point clouds analysis. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 24(2):2158–2168, Feb. 2023.

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

A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré. Hippo: Recurrent memory with optimal polynomial projections. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1474–1487, Virtual Event, Dec. 2020.

A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 572–585, Virtual Event, Dec. 2021.

A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. In *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–15, Virtual Event, Apr. 2022.

M. H. Guo, J. X. Cai, Z. N. Liu, T. J. Mu, R. R. Martin, and S. M. Hu. Pct: Point cloud transformer. *Computational Visual Media (CVM)*, 7(2):187–199, Apr. 2021.

Z. Y. Guo, R. R. Zhang, L. T. Qiu, X. Z. Li, and P. A. Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 791–799, Macao, SAR, China, Aug. 2023.

A. Gupta, A. Gu, and J. Berant. Diagonal state spaces are as effective as structured state spaces. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–12, New Orleans, LA, USA, Dec. 2022.

X. Han, Y. Tang, Z. X. Wang, and X. Z. Li. Mamba3d: Enhancing local features for 3d point cloud analysis via state space model. In *Proc. ACM International Conference on Multimedia (MM)*, pp. 4995–5004, Melbourne, Australia, Oct. 2024.

C. H. He, R. H. Li, S. Li, and L. Zhang. Voxel set transformer: A set-to-set approach to 3d object detection from point clouds. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8407–8417, New Orleans, LA, USA, Jun. 2022.

H. Y. He, J. N. Zhang, Y. X. Cai, H. X. Chen, X. B. Hu, Z. Y. Gan, Y. B. Wang, C. J. Wang, Y. S. Wu, and L. Xie. Mobilemamba: Lightweight multi-receptive visual mamba network. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4497–4507, Nashville, TN, USA, Jun. 2025.

X. W. He, S. L. Cheng, D. K. Liang, S. Bai, X. Wang, and Y. Y. Zhu. Latformer: Locality-aware point-view fusion transformer for 3d shape recognition. *Pattern Recognit (PR)*, 151:110413, Jul. 2024.

C. Y. Hong, Y. Y. Chou, and T. L. Liu. Attention discriminant sampling for point clouds. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14383–14394, Paris, France, Oct. 2023.

N. Köprücü, D. Okpekpe, and A. Orvieto. Nimba: Towards robust and principled processing of point clouds with ssms. arXiv:2411.00151, 2024.

X. Lai, J. H. Liu, L. Jiang, L. W. Wang, and H. S. Zhao. Stratified transformer for 3d point cloud segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8490–8499, New Orleans, LA, USA, Jun. 2022.

K. C. Li, X. H. Li, Y. Wang, Y. N. He, Y. L. Wang, L. M. Wang, and Y. Qiao. Videomamba: State space model for efficient video understanding. arXiv:2403.06977, 2024a.

Z. H. Li, P. Gao, K. You, C. Yan, and M. Paul. Global attention-guided dual-domain point cloud feature learning for classification and segmentation. *IEEE Transactions on Artificial Intelligence (TAI)*, 5(10):5167–5167, Oct. 2024b.

Z. Y. Li, Y. B. Ai, J. H. Lu, C. X. Wang, J. C., Deng, H. Z. Chang, Y. Z. Liang, W. F. Yang, S. F. Zhang, and T. Z. Zhang. Pamba: Enhancing global interaction in point clouds via state space model. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 5092–5100, Philadelphia, PA, USA, Feb. 2025.

D. K. Liang, X. Zhou, W. Xu, X. K. Zhu, Z. K. Zou, X. Q. Ye, X. Tan, and X. Bai. Pointmamba: A simple state space model for point cloud analysis. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–14, Vancouver, BC, Canada, Dec. 2024.

D. K. Liang, T. R. Feng, X. Zhou, Y. M. Zhang, Z. K. Zou, and X. Bai. Parameter-efficient fine-tuning in spectral domain for point cloud learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 47(12):10949–10966, Aug. 2025a.

D. K. Liang, W. Hua, C. S. Shi, Z. K. Zou, X. Q. Ye, and X. Bai. Sood++: Leveraging unlabeled data to boost oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 48(1):840–858, Sep. 2025b.

D. K. Liang, D. Y. Zhang, X. Zhou, S. F. Tu, T. R. Feng, X. F. Li, Y. M. Zhang, M. Y. Du, X. Tan, and X. Bai. Seeing the future, perceiving the future: A unified driving world model for future generation and perception. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–11, Vienna, Austria, Jun. 2026.

H. T. Liu, M. Cai, and Y. J. Lee. Masked discrimination for self-supervised learning on point clouds. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 657–675, Tel Aviv, Israel, Oct. 2022.

J. M. Liu, Y. Wu, M. G. Gong, Z. X. Liu, Q. G. Miao, and W. P. Ma. Inter-modal masked autoencoder for self-supervised learning on point clouds. *IEEE Transactions on Multimedia (TMM)*, 26:3897–3908, Sep. 2023a.

J. M. Liu, R. J. Yu, Y. Wang, Y. Zheng, T. C. Deng, W. C. Ye, and H. S. Wang. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. arXiv:2403.06467, 2024a.

L. Y Liu, M. Zhang, J. H. Yin, T. W. Liu, W. Ji, Y. R. Piao, and H. C. Lu. Defmamba: Deformable visual state space model. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8838–8847, Nashville, TN, USA, Jun. 2025.

Y. H. Liu, B. Tian, Y. S. Lv, L. X. Li, and F. Y. Wang. Point cloud classification using content-based transformer via clustering in feature space. *IEEE/CAA Journal of Automatica Sinica (JAS)*, 11(1): 231–239, Jan. 2024b.

Z. J. Liu, X. Y. Yang, H. T. Tang, S. Yang, and S. Han. Flatformer: Flattened window attention for efficient point cloud transformer. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1200–1211, Vancouver, BC, Canada, Jun. 2023b.

J. Ma, F. F. Li, and B. Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv:2401.04722, 2024.

K. Mazur and V. Lempitsky. Cloud transformers: A universal approach to point cloud processing tasks. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10695–10704, Montreal, QC, Canada, Oct. 2021.

H. Mehta, A. Gupta, A. Cutkosky, and B. Neyshabur. Long range language modeling via gated state spaces. In *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–16, Kigali, Rwanda, May 2023.

Robert Ahadizad Newport, Sidong Liu, and Antonio Di Ieva. Integrating eye gaze into machine learning using fractal curves. In *Gaze Meets Machine Learning Workshop*, pp. 113–126. PMLR, 2023.

D. Nie, R. Lan, L. Wang, and X. F. Ren. Pyramid architecture for multi-scale processing in point cloud segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17263–17273, New Orleans, LA, USA, Jun. 2022.

A. Nordin and A. Telles. Comparing the locality preservation of z-order curves and hilbert curves. 2023.

Y. T. Pang, W. X. Wang, F. E. H. Tay, W. Liu, Y. H. Tian, and L. Yuan. Masked autoencoders for point cloud self-supervised learning. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 604–621, Tel Aviv, Israel, Oct. 2022.

C. Park, Y. Jeong, M. S. Cho, and J. Park. Fast point transformer. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16928–16937, New Orleans, LA, USA, Jun. 2022.

J. Park, S. Lee, S. Kim, Y. Y. Xiong, and H. J. Kim. Self-positioning point-based transformer for point cloud understanding. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21814–21823, Vancouver, BC, Canada, Jun. 2023.

C. R. Qi, H. Su, K. C. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77–85, Honolulu, HI, USA, Jul. 2017a.

C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5099–5108, Long Beach, CA, USA, Dec. 2017b.

Z. K. Qi, R. P. Dong, G. F. Fan, Z. Ge, X. Y. Zhang, K. S. Ma, and L. Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *Proc. International Conference on Machine Learning (ICML)*, pp. 28223–28243, Honolulu, HI, USA, Jul. 2023.

K. L. Qu, P. Gao, Q. Dai, and Y. H. Sun. Hydramamba: Multi-head state space model for global point cloud learning. In *Proc. ACM International Conference on Multimedia (MM)*, pp. 333–342, Dublin Ireland, Oct. 2025.

H. Resani and B. Nasihatkon. Miracle3d: Memory-efficient integrated robust approach for continual learning on 3d point clouds via shape model construction. In *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–14, Singapore, Apr. 2025.

D. Robert, H. Raguet, and L. Landrieu. Efficient 3d semantic segmentation with superpoint transformer. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17149–17158, Paris, France, Oct. 2023.

J. C. Ruan and S. C. Xiang. Vm-unet: Vision mamba unet for medical image segmentation. arXiv: 2402.02491, 2024.

M. Schöne, Y. Bhisikar, K. Bania, K. K. Nazeer, C. Mayr, A. Subramoney, and D. Kappel. Stream: A universal state-space model for sparse geometric data. arXiv:2411.12603, 2024.

A. M. Shaker, S. T. Wasim, S. H. Khan, J. Gall, and F. S. Khan. Groupmamba: Efficient group-based visual state space model. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14912–14922, Nashville, TN, USA, Jun. 2025.

D. Shi. Transnext: Robust foveal visual perception for vision transformers. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17773–17783, Seattle, WA, USA, Jun. 2024.

J. T. H. Smith, A. Warrington, and S. W. Linderman. Simplified state space layers for sequence modeling. In *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–13, Kigali, Rwanda, May 2023.

E. L. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1257–1264, Vancouver, BC, Canada, Dec. 2005.

E. E. M. Stewart, M. Valsecchi, and A. C. Schütz. A review of interactions between peripheral and foveal vision. *Journal of Vision (JVISION)*, 20(12):2, Nov. 2020.

J. Y. Su, J. B. Ma, S. Y. Tong, E. Z. Xu, and M. H. Chen. Multiscale attention wavelet neural operator for capturing steep trajectories in biochemical systems. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 15100–15107, Vancouver, BC, Canada, Feb. 2024.

P. Sun, M. X. Tan, W. Y. Wang, C. X. Liu, F. Xia, Z. Q. Leng, and D. Anguelov. Swformer: Sparse window transformer for 3d object detection in point clouds. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 426–442, Tel Aviv, Israel, Oct. 2022.

Y. W. Tang, R. Zhang, Z. Guo, X. Z. Ma, B. Zhao, Z. G. Wang, D. Wang, and X. L. Li. Point-peft: Parameter-efficient fine-tuning for 3d pre-trained models. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 5171–5179, Vancouver, BC, Canada, Feb. 2024.

H. Thomas, Y. H. Tsai, T. D. Barfoot, and J. Zhang. Kpconvx: Modernizing kernel point convolution with kernel attention. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5525–5535, Seattle, WA, USA, Jun. 2024.

M. A. Uy, Q. H. Pham, B. S. Hua, D. T. Nguyen, and S. K. Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1588–1597, Seoul, South Korea, Oct. 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, Long Beach, CA, USA, Dec. 2017.

B. A. Wandell. *Foundations of Vision*. Sinauer Associates, 1995.

J. Wang, W. T. Zhu, P. C. Wang, X. Yu, L. D. Liu, M. Omar, and R. Hamid. Selective structured state-spaces for long-form video understanding. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6387–6397, Vancouver, BC, Canada, Jun. 2023.

P. S. Wang. Octformer: Octree-based transformers for 3d point clouds. *ACM Transactions on Graphics (TOG)*, 42(4):155, Jul. 2023.

Z. C. Wang, Z. H. Chen, Y. M. Wu, Z. Zhao, L. P. Zhou, and D. Xu. Pointramba: A hybrid transformer-mamba framework for point cloud analysis. arXiv:2405.15463, 2024.

K. T. Wijaya, D. H. Paek, and S. H. Kong. Advanced feature learning on point clouds using multi-resolution features and learnable pooling. *Remote Sensing (RS)*, 16(11):1835, May 2024.

C. Z. Wu, J. W. Zheng, J. Pfrommer, and J. Beyerer. Attention-based point cloud edge sampling. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5333–5343, Vancouver, BC, Canada, Jun. 2023a.

X. Y. Wu, Y. X. Lao, L. Jiang, X. H. Liu, and H. S. Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–13, New Orleans, LA, USA, Dec. 2022.

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

X. Y. Wu, Z. T. Tian, X. Wen, B. H. Peng, X. H. Liu, K. C. Yu, and H. S. Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19551–19562, Seattle, WA, USA, Jun. 2024b.

X. Y. Wu, D. DeTone, D. Frost, T. W. Shen, C. Xie, N. Yang, J. Engel, R. Newcombe, H. S. Zhao, and J. Straub. Sonata: Self-supervised learning of reliable point representations. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22193–22204, Nashville, TN, USA, Jun. 2025.

Y. Wu, J. M. Liu, M. G. Gong, P. R. Gong, X. L. Fan, A. K. Qin, Q. G. Miao, and W. P. Ma. Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding. *IEEE Transactions on Multimedia (TMM)*, 26:1626–1638, Jun. 2023b.

Z. R. Wu, S. R. Song, A. Khosla, F. Yu, L. G. Zhang, X. O. Tang, and J. X. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1912–1920, Boston, MA, USA, Jun. 2015.

P. Xiang, X. Wen, Y. S. Liu, H. Zhang, Y. Fang, and Z. Z. Han. Retro-fpn: Retrospective feature pyramid network for point cloud semantic segmentation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17826–17838, Paris, France, Oct. 2023.

Z. H. Xing, T. Ye, Y. J. Yang, G. Liu, and L. Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv:2401.13560, 2024.

M. Y. Xu, M. T. Xu, T. He, W. Ouyang, Y. L. Wang, X. G. Han, and Y. Qiao. Mm-3dscene: 3d scene understanding by customizing masked modeling with informative-preserved reconstruction and self-distilled consistency. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4380–4390, Vancouver, BC, Canada, Jun. 2023.

W. Xu, C. S. Shi, S. F. Tu, X. Zhou, D. K. Liang, and X. Bai. A unified framework for 3d scene understanding. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–14, Vancouver, BC, Canada, Dec. 2024.

S. M. Yan, C. Song, Y. K. Kong, and Q. X. Huang. Multi-view representation is what you need for point-cloud pre-training. In *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–15, Vienna, Austria, May 2024a.

S. M. Yan, Y. Q. Yang, Y. X. Guo, H. Pan, P. S. Wang, X. Tong, Y. Liu, and Q. X. Huang. 3d feature prediction for masked-autoencoder-based point cloud pretraining. In *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–12, Vienna, Austria, May 2024b.

X. Yan, C. D. Zheng, Z. Li, S. Wang, and S. G. Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5588–5597, Virtual Event / Seattle, WA, USA, Jun. 2020.

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

L. Yi, V. G. Kim, D. Ceylan, I. C. Shen, M. Y. Yan, H. Su, C. Lu, Q. X. Huang, A. Sheffer, and L. Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (TOG)*, 35(6):210, Dec. 2016.

X. M. Yu, L. L. Tang, Y. M. Rao, T. J. Huang, J. Zhou, and J. W. Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19291–19300, New Orleans, LA, USA, Jun. 2022.

K. A. Zeid, J. Schult, A. Hermans, and B. Leibe. Point2vec for self-supervised representation learning on point clouds. In *Proc. DAGM German Conference (DAGM)*, pp. 131–146, Heidelberg, Germany, Sep. 2023.

Y. H. Zha, J. P. Wang, T. Dai, B. Chen, Z. Wang, and S. T. Xia. Instance-aware dynamic prompt tuning for pre-trained point cloud models. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14161–14170, Paris, France, Oct. 2023.

Y. H. Zha, N. Q. Li, Y. Z. Wang, T. Dai, H. Guo, B. Chen, Z. Wang, Z. H. Ouyang, and S. T. Xia. Lcm: Locally constrained compact point cloud model for masked point modeling. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–14, Vancouver, BC, Canada, Dec. 2024.

Y. H. Zha, Y. Z. Wang, H. Guo, J. P. Wang, T. Dai, B. Chen, Z. H. Ouyang, Y. R. Xue, K. Chen, and S. T. Xia. Pma: Towards parameter-efficient point cloud understanding via point mamba adapter. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16976–16986, Nashville, TN, USA, Jun. 2025.

C. Zhang, H. C. Wan, X. Y. Shen, and Z. Z. Wu. Patchformer: An efficient point transformer with patch attention. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11789–11798, New Orleans, LA, USA, Jun. 2022.

D. Y. Zhang, D. K. Liang, Z. K. Zou, J. Y. Li, X. Q. Ye, Z. Liu, X. Tan, and X. Bai. A simple vision transformer for weakly semi-supervised 3d object detection. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8339–8349, Paris, France, Oct. 2023.

D. Y. Zhang, D. K. Liang, Z. C. Tan, X. Q. Ye, C. Zhang, J. D. Wang, and X. Bai. Make your vit-based multi-view 3d detectors faster via token compression. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 56–72, Milan, Italy, Sep. 2024a.

T. Zhang, X. T. Li, H. B. Yuan, S. P. Ji, and S. C. Yan. Point cloud mamba: Point cloud learning via state space model. arXiv:2403.00762, 2024b.

H. S. Zhao, L. Jiang, J. Y. Jia, P. Torr, and V. Koltun. Point transformer. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16239–16248, Montreal, QC, Canada, Oct. 2021.

X. Zheng, X. S. Huang, G. F. Mei, Y. N. Hou, Z. Y. Lyu, B. Dai, W. L. Ouyang, and Y. S. Gong. Point cloud pre-training with diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22935–22945, Seattle, WA, USA, Jun. 2024.

J. S. Zhou, J. S. Wang, B. R. Ma, Y. S. Liu, T. J. Huang, and X. L. Wang. Uni3d: Exploring unified 3d representation at scale. In *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–14, Vienna, Austria, May 2024a.

X. Zhou, D. K. Liang, W. Xu, X. K. Zhu, Y. H. Xu, Z. K. Zou, and X. Bai. Dynamic adapter meets prompt tuning: Parameter-efficient transfer learning for point cloud analysis. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14707–14717, Seattle, WA, USA, Jun. 2024b.

X. Zhou, D. K. Liang, S. F. Tu, X. W. Chen, Y. K. Ding, D. Y. Zhang, F. Y. Tan, H. S. Zhao, and X. Bai. Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27817–27827, Honolulu, HI, USA, Oct. 2025a.

Y. C. Zhou, J. Y. Gu, T. Y. Chiang, F. B. Xiang, and H Su. Point-sam: Promptable 3d segmentation model for point clouds. In *Proc. International Conference on Learning Representations (ICLR)*, pp. 1–14, Singapore, Apr. 2025b.

L. H. Zhu, B. C. Liao, Q. Zhang, X. L. Wang, W. Y. Liu, and X. G. Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1–10, Vienna, Austria, Jul. 2024.

A State space models

SSMs are cyclic processes with latent states, which map a 1-D equation or sequence $x\left(t\right) \in \mathbb{R}^N$ to $y\left(t\right) \in \mathbb{R}^N$ by a latent state $h\left(t\right) \in \mathbb{R}^N$. The process is mathematically denoted as a linear ordinary differential equation as follows

$$y\left(t\right) = \boldsymbol{C}h\left(t\right), h'\left(t\right) = \boldsymbol{A}h\left(t\right) + \boldsymbol{B}x\left(t\right), \tag{7}$$

where the three parameters $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, $\boldsymbol{B} \in \mathbb{R}^N$, and $\boldsymbol{C} \in \mathbb{R}^N$ represent the state matrix, input matrix, and output matrix, respectively. Since the above SSMs run on continuous inputs and are not applicable to discrete inputs such as images and text, they cannot be introduced into deep models. Thus, it is necessary to discretize them, and the zero-order hold is commonly used as a discretization method. The discretized formulas are as follows

$$h_t = \bar{\boldsymbol{A}}h_{t-1} + \bar{\boldsymbol{B}}x_t, \quad y_t = \boldsymbol{C}h_t, \tag{8}$$

where $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$ are the results of discretizing the continuous parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ by a time scale $\boldsymbol{\Delta}$, denoted as

$$\bar{\boldsymbol{A}} = e^{\boldsymbol{\Delta A}}, \ \ \bar{\boldsymbol{C}} = \boldsymbol{C}, \quad \bar{\boldsymbol{B}} = (\boldsymbol{\Delta A})^{-1}\big(e^{\boldsymbol{\Delta A}} - \boldsymbol{I}\big)(\boldsymbol{\Delta B}). \tag{9}$$

Since processing the input and latent state equally, previous approaches focusing on linear time-invariant SSMs (where $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$ are invariant) may fail to capture critical information from context. Hence, Mamba proposes a novel SSM termed S6 by integrating an input-dependent selective mechanism into SSMs, where $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$ are the functions of inputs, indicating Mamba is linear time-variant.

B Datasets and implementation

B.1 Datasets

ModelNet40 dataset contains 12,311 CAD models across 40 categories, with 9,843 samples in the training set and 2,468 samples in the test set. Data preprocessing follows the method of Qi et al. (2017a), where 1,024 points and their normal vectors are uniformly sampled from each sample as input. As per most relevant studies in the literature, the overall accuracy (OA) is adopted as an evaluation metric.

ShapeNet dataset comprises 16,878 samples from 50 parts across 16 categories, with 14,005 samples in the training set and 2,873 in the test set. Data preprocessing is consistent with that applied to ModelNet40 dataset. As per most relevant studies in the literature, the instance mIoU (Ins. mIoU) is adopted as an evaluation metric.

S3DIS dataset comprises 3D point cloud data from 271 indoor scenes across 6 areas, with each point annotated with one of 13 semantic labels. Data preprocessing follows the method of Qi et al. (2017a), where input features include point coordinates, RGB color, and normalized positions. As per most relevant studies in the literature, area 5 is used as the test set, while the remaining areas are used for training, and the mean IoU (mIoU) is adopted as an evaluation metric.

ScanObjectNN (PB_T50_RS variant) contains 14,450 valid samples across 15 categories, with 11,636 samples for training and 2,814 for testing. Except for using only coordinates as input, data preprocessing and evaluation metric are consistent with those used for ModelNet40 dataset.

B.2 Implementation

To better understand the model's structure and implementation, we list detailed network architectures and training settings across different datasets in Tab. 11.

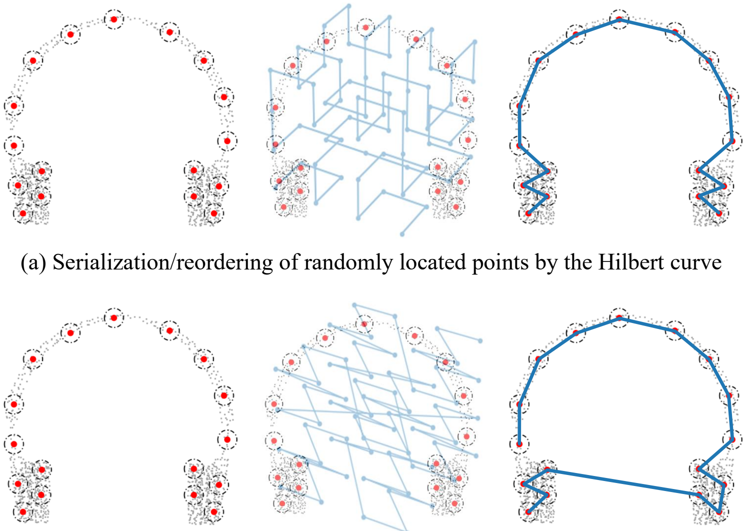
C More ablation studies

C.1 Ablation comparison on the Serialization

Serialization. Based on point cloud serialization, the context-scan state space generates a continuous scanning path with inter-point structural dependencies. To validate the rationale behind selecting

Configurations	ModelNet40	ScanObjectNN	ShapeNet	S3DIS
Training epochs	500	500	600	500
Optimizer & Scheduler	Adamw & CosLR	AdamW & CosLR	Adamw & CosLR	Adamw & CosLR
Weight decay	0.01	0.01	0.01	0.01
Learning rate	8e-4	4e-4	1e-3	1e-3
Warmup epochs	10	20	10	10
Batch size	24	24	24	12
Embedding channels	48	48	96	48
KNN	8	8	8	8
IPP ratio	8	8	8	16
Encoder depth	[1, 1, 1, 1]	[1, 1, 1, 1]	[2, 2, 6, 2]	[1, 2, 3, 1]
Encoder channels	[48, 96, 192, 384]	[48, 96, 192, 384]	[96, 192, 384, 768]	[96, 192, 384, 768]
Decoder depth	-	-	[1, 1, 1, 1]	[1, 1, 1, 1]
Decoder channels	-	-	[768, 384, 192, 96]	[768, 384, 192, 96]
Downsampling stride	[4, 4, 4]	[4, 4, 4]	[4, 4, 4]	[4, 4, 4]
MLP ratio	4	4	4	4
QKV bias	True	True	True	True
Dropout	0.3	0.3	0.3	0.3
Augmentation	RandomScale RandomShift ShufflePoint	ShufflePoint RandomScale RandomRotate	RandomScale RandomShift ShufflePoint	RandomScale RandomFlip RandomJitter ChromaticAutoContrast ChromaticTranslation ChromaticJitter

Table 11: Detailed implementation configurations.



(a) Serialization/reordering of randomly located points by the Hilbert curve

(b) Serialization/reordering of randomly located points by the Z-order curve

Figure 5: Comparison of the serialization of randomly located points by the Hilbert curve (top) and Z-Order curve (bottom).

the Hilbert curve, Tab. 12 compares the ablation results of various serialization strategies. When two serialization methods are employed, they are applied separately to the two directions of the bidirectional S6. Quantitative analysis indicates that while combining multiple serialization strategies can provide richer spatial information, differences in spatial relationships between these strategies may introduce confusion and interfere with the learning of spatial consistency. By leveraging its exceptional spatial locality-preserving property, as shown in Fig. 5, the Hilbert curve establishes a high-fidelity spatially adjacent scanning path with reliable inter-point structural dependencies for the state space model. This aligns with the visual search pattern of eye movements scanning continuously along spatially adjacent regions, thereby achieving an optimal balance between accuracy and efficiency. To further discuss the specific advantages of the Hilbert curve over learnable serial-

Serialization	Params	FLOPs	Throughput	OA
None	7.36M	0.610G	219FPS	91.34
Hilbert	7.36M	0.610G	163FPS	**94.17**
Z-Order	7.36M	0.610G	209FPS	93.06
Hilbert & Trans-Hilbert	7.36M	0.610G	133FPS	93.78
Hilbert & Z-Order	7.36M	0.610G	155FPS	93.52
Learnable Serialization	8.04M	0.723G	168FPS	92.78

Table 12: Ablation results with multiple serialization strategies.

ization strategies in terms of spatial locality preservation, continuity, and computational efficiency, we compare with the learnable serialization strategy from the latest research (Zha et al., 2025). It is intuitively observed that, compared to the learnable serialization, the Hilbert curve exhibits higher computational efficiency and superior spatial locality preservation and continuity. We attribute the poorer performance of the learnable serialization to the fact that it is an adaptive method for determining geometric correlation between points, but this approach possesses much less geometry-specific inductive biases compared to space-filling curves. In summary, the Hilbert curve introduces more precise inductive bias regarding geometric correlation compared to the Z-Order curve and learnable serialization strategies.

C.2 Ablation comparison of IPP and FPS

In our work, we employ the proposed Induced Point Pooling (IPP) for spatial downsampling. To investigate its ability to flexibly adapt to the non-uniform distribution of point clouds for global semantic integration, we present ablation results comparing IPP with the Farthest Point Sampling (FPS) at different sampling rates in Fig. 6, where /N denotes the sampling rate relative to the input number of points, and None indicates the absence of the spatial downsampling branch. Intuitively, at low sampling rates, both methods exhibit comparable performance, indicating that FPS can obtain better global semantics with its excellent spatial coverage when sufficient sampling points are available. However, as the downsampling rate increases, the performance of FPS declines sharply. At a sampling rate of /256, its accuracy approaches the baseline without the downsampling branch, suggesting its inability to capture critical semantics from non-uniform point clouds at high sparsity. In contrast, IPP exhibits a more gradual performance decrease, maintaining an excellent accuracy of 93.39% even at the /256 sampling rate. This confirms that IPP, through trainable induced points that adaptively learn the point cloud distribution, can more flexibly and robustly integrate global semantics, providing effective coarse-grained information for the point-focused attention.

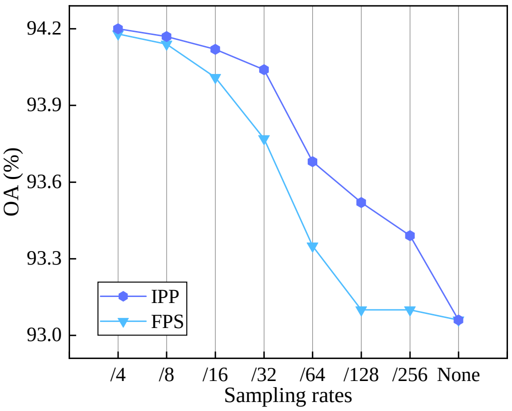


Figure 6: Quantitative results of IPP and FPS.

D Complexity analysis

Following the same settings as Eq.(6) and considering the feature transformation, the computational complexities of each module in the point-focused attention are as follows:

$$\begin{aligned} \Omega\,(\mathrm{LNB}) &= 3ND^2 + 2NKD \\ \Omega\,(\mathrm{SSD}) &= ND^2 + 2MD^2 + 2NMD \\ \Omega\,(\mathrm{IPP}) &= 2ND^2 + 2NMD \end{aligned} \tag{10}$$

Finally, we have a complexity of $\Omega\,(\mathrm{PFA}) = 6ND^2 + 2MD^2 + 2NKD + 4NMD$ for the point-focused attention. Since the context-scan state space inherits the linear complexity inherent in the state space model, the overall network exhibits linear complexity.

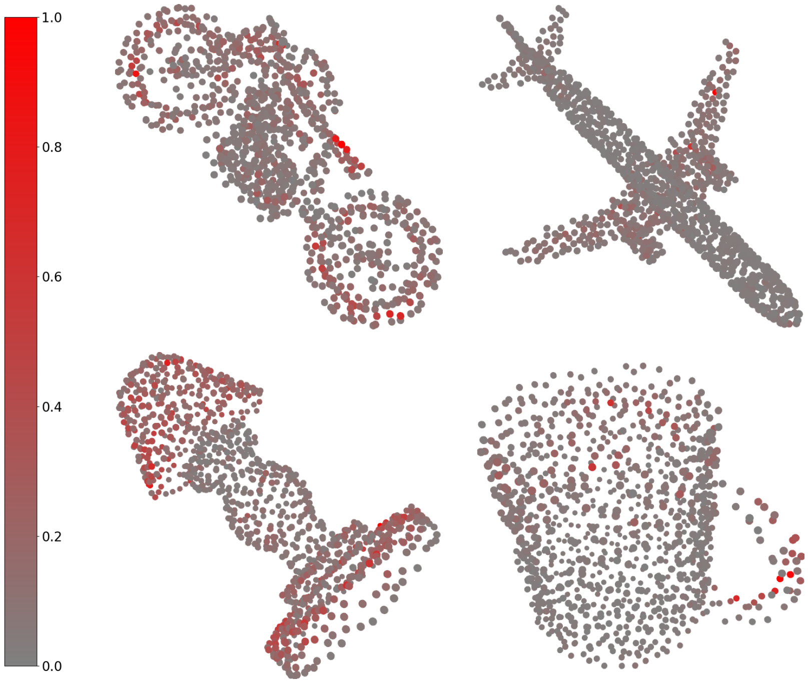


Figure 7: Visualization results of attention heatmaps.

E Visualizations

E.1 Attention heatmaps

To better understand the attention responses and the advantages of the proposed method, we present in Fig. 7 the attention heatmaps of PointLearner for different objects, generated from attention weights in the local neighbor branch of the last point-focused attention layer within the decoder. These attention heatmaps illustrate that, through a fully understanding of the bio-inspired visual perception, our method effectively focuses on critical information for semantic inference to achieve outstanding performance, such as the tires and seat on the motorcycle, as well as the base and cover of the lamp.

E.2 Qualitative comparison

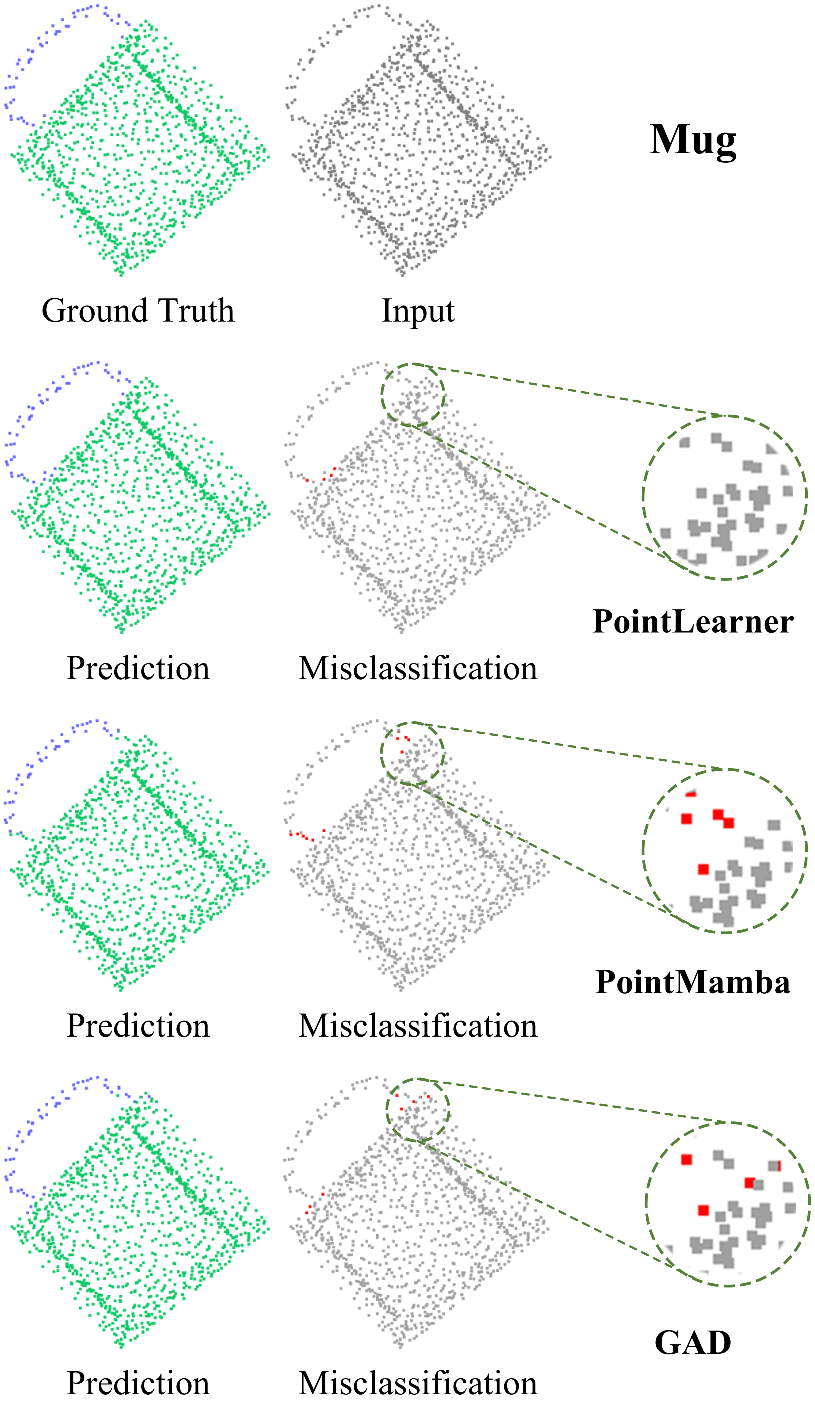
To intuitively demonstrate the performance of our network, we present in Fig. 8 the visualization results of our network alongside the top-performing SSM method (PointMamba (Liang et al., 2024)) and attention method (GAD (Li et al., 2024b)) from Tab. 2 on ShapeNet dataset, where the red points denote that these points are misclassified. The comparison of the visualization results reveals that our network is able to achieve better part segmentation results at the boundaries of objects.

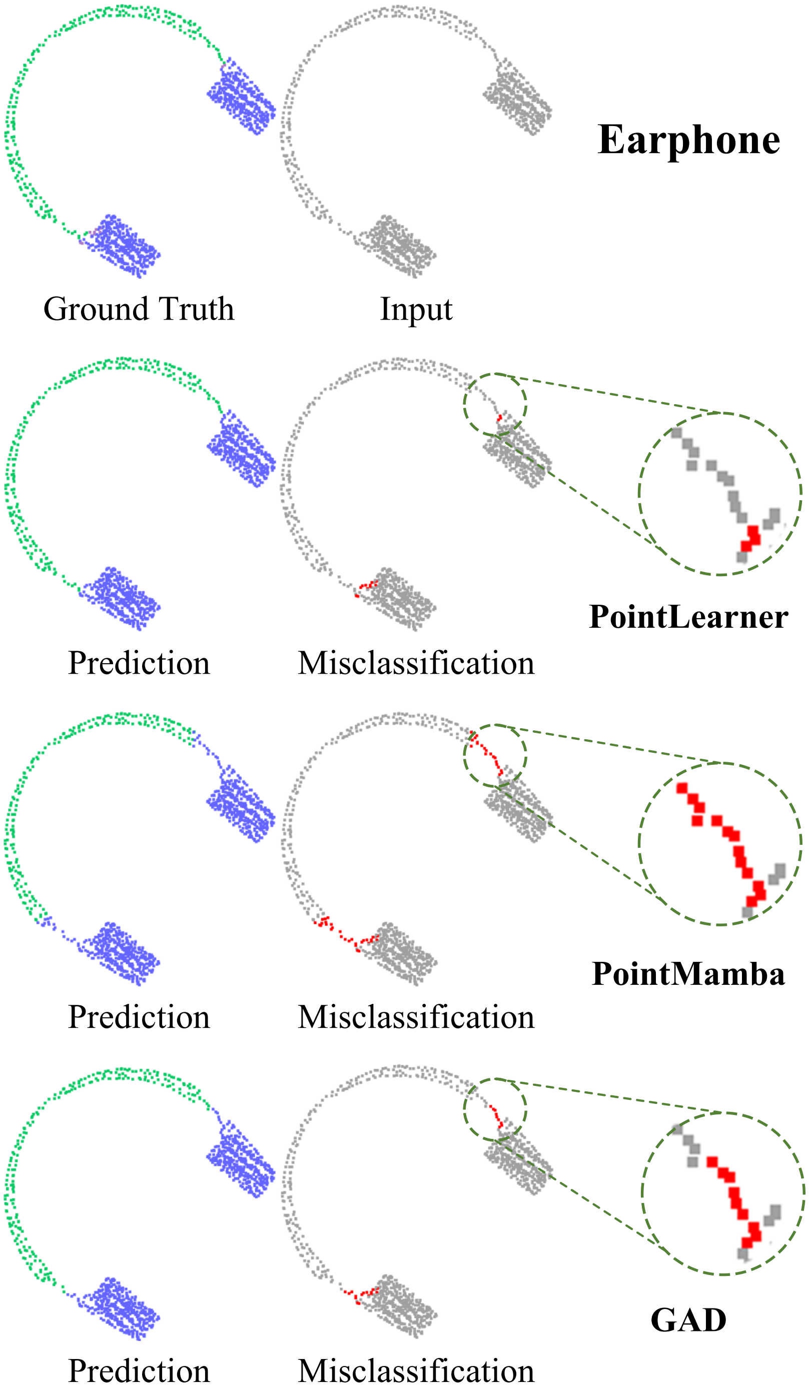
F Future work

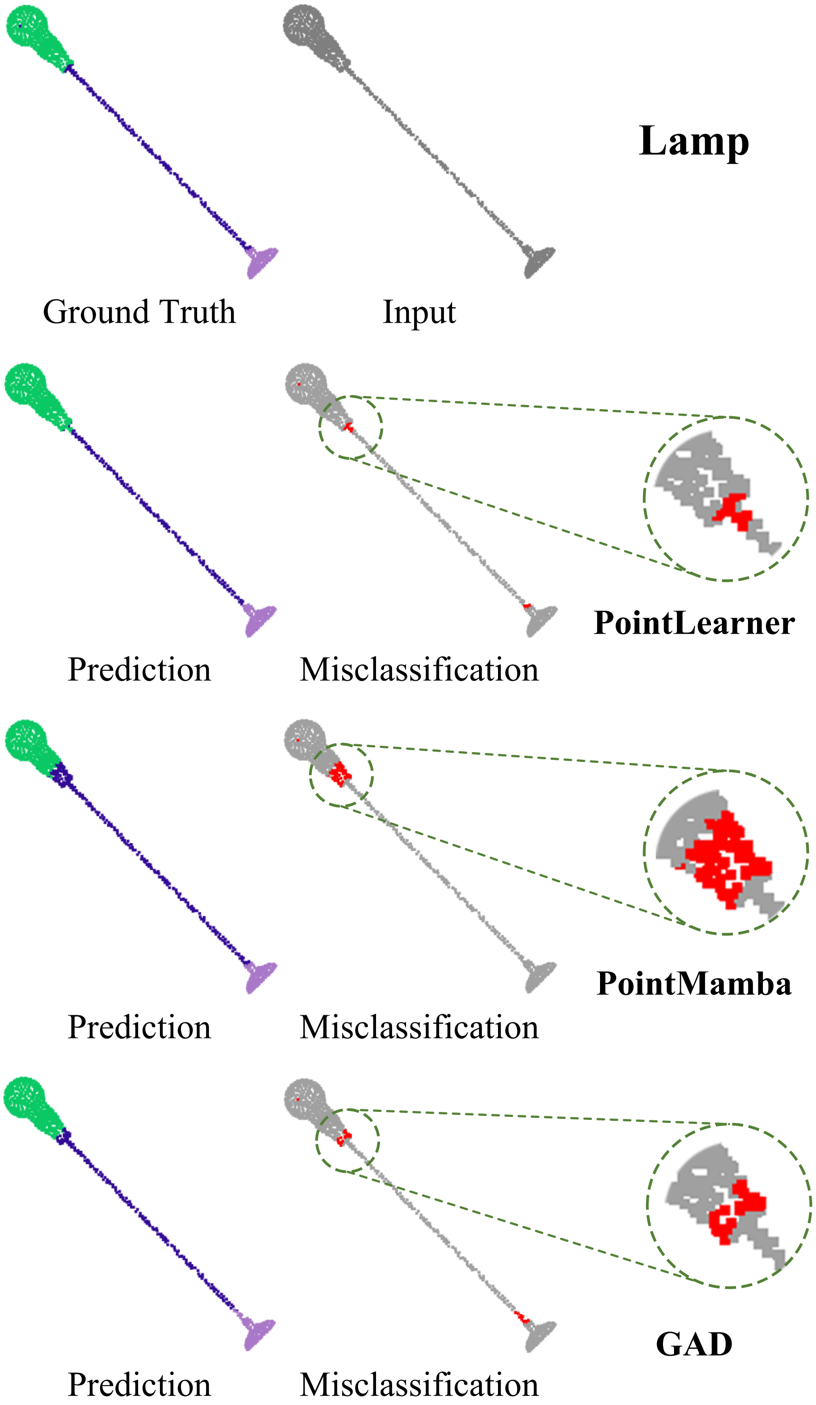
In our experimental comparisons, it is observed that most existing attention models employ pre-training methods to improve performance, with the self-supervised pre-training paradigm dominating. Self-supervised pre-training methods can leverage large amounts of unlabeled data to enhance feature modeling capabilities, as well as help Transformer models with large receptive fields achieve effective local or structural modeling by increasing data scale. Although self-supervised pre-training on large-scale point cloud datasets has been proven effective for improving the accuracy of Transformer models, the compatibility of existing Transformer self-supervised pre-training methods on hybrid architectures, as well as self-supervised pre-training strategies specifically tailored for hybrid architectures, remain underexplored. Hence, it is a promising direction for future research to collect more data and design self-supervised learning methods for hybrid models, as in PPT (Wu et al., 2024b) and Sonata (Wu et al., 2025) designed for Transformer models.

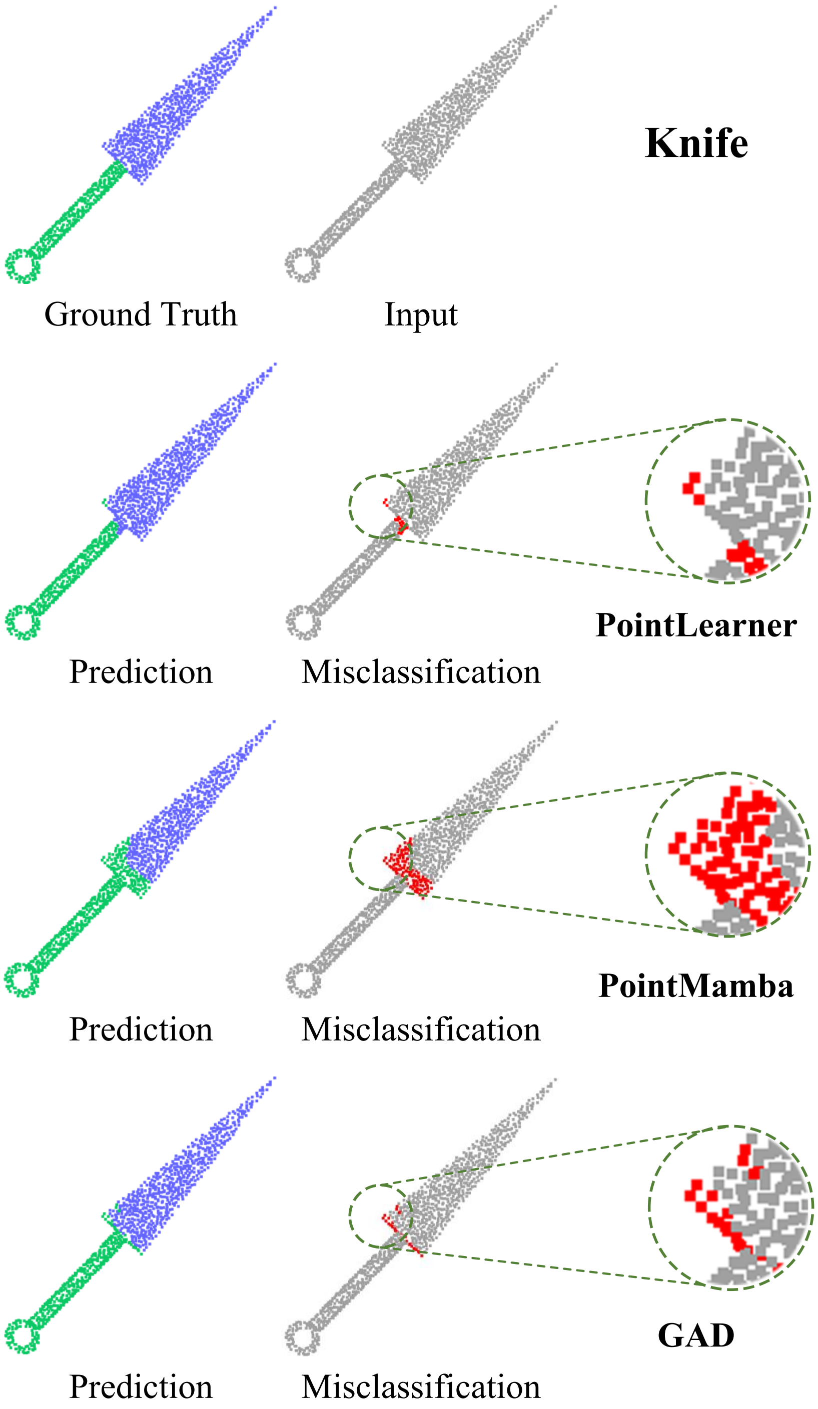
G THE USE OF LARGE LANGUAGE MODELS (LLMS)

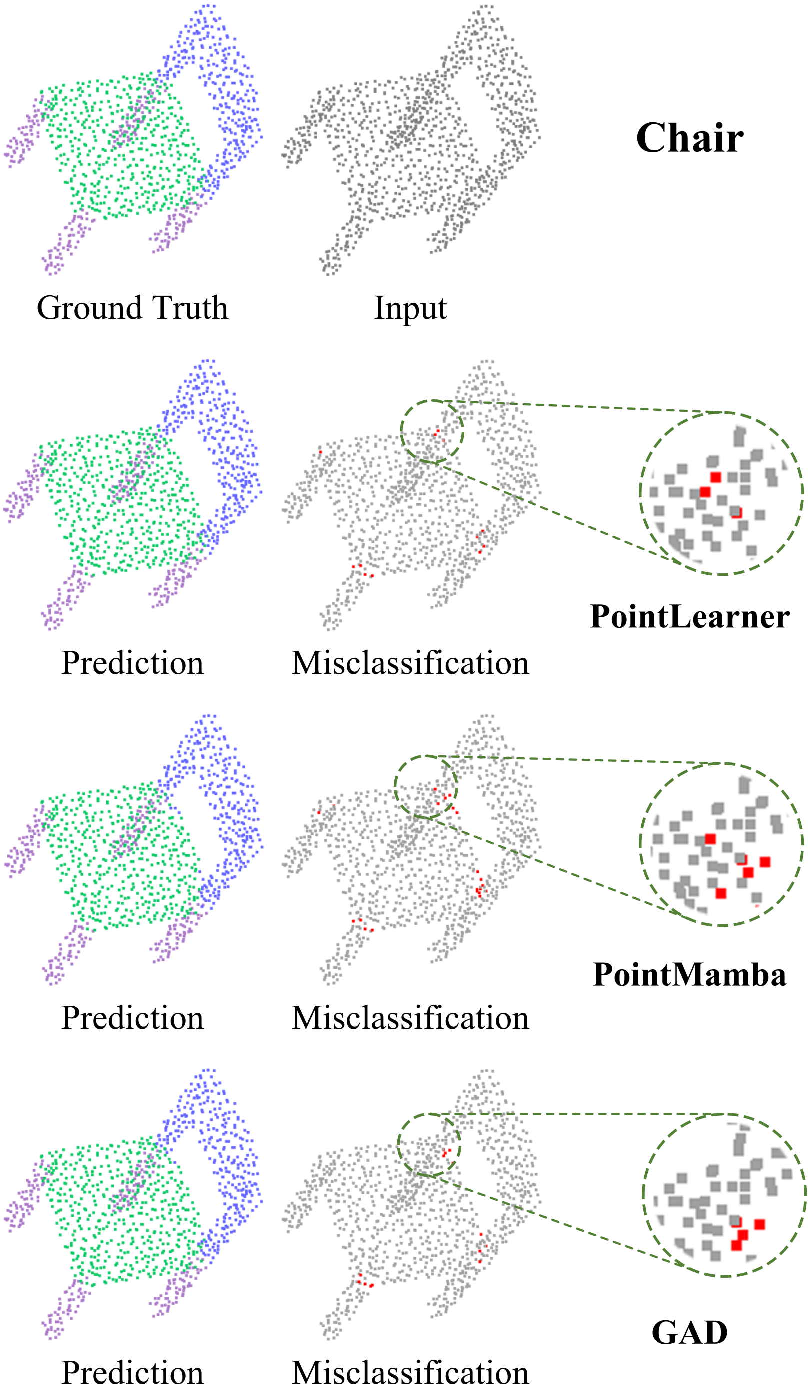
The authors utilized large language models (LLMs) to a limited extent for proofreading and improving grammatical correctness. All key aspects of the research, encompassing innovation, conceptual development, and literature discovery, were solely driven by the authors without LLM assistance.




Earphone
Ground Truth
Input
PointLearner
Prediction
Misclassification
PointMamba
Prediction
Misclassification
GAD
Prediction
Misclassification


Lamp
Ground Truth
Input
PointLearner
Prediction
Misclassification
PointMamba
Prediction
Misclassification
GAD
Prediction
Misclassification


Knife
Ground Truth
Input
PointLearner
Prediction
Misclassification
PointMamba
Prediction
Misclassification
GAD
Prediction
Misclassification


Chair
Ground Truth
Input
Prediction
Misclassification
PointLearner
Prediction
Misclassification
PointMamba
Prediction
Misclassification
GAD

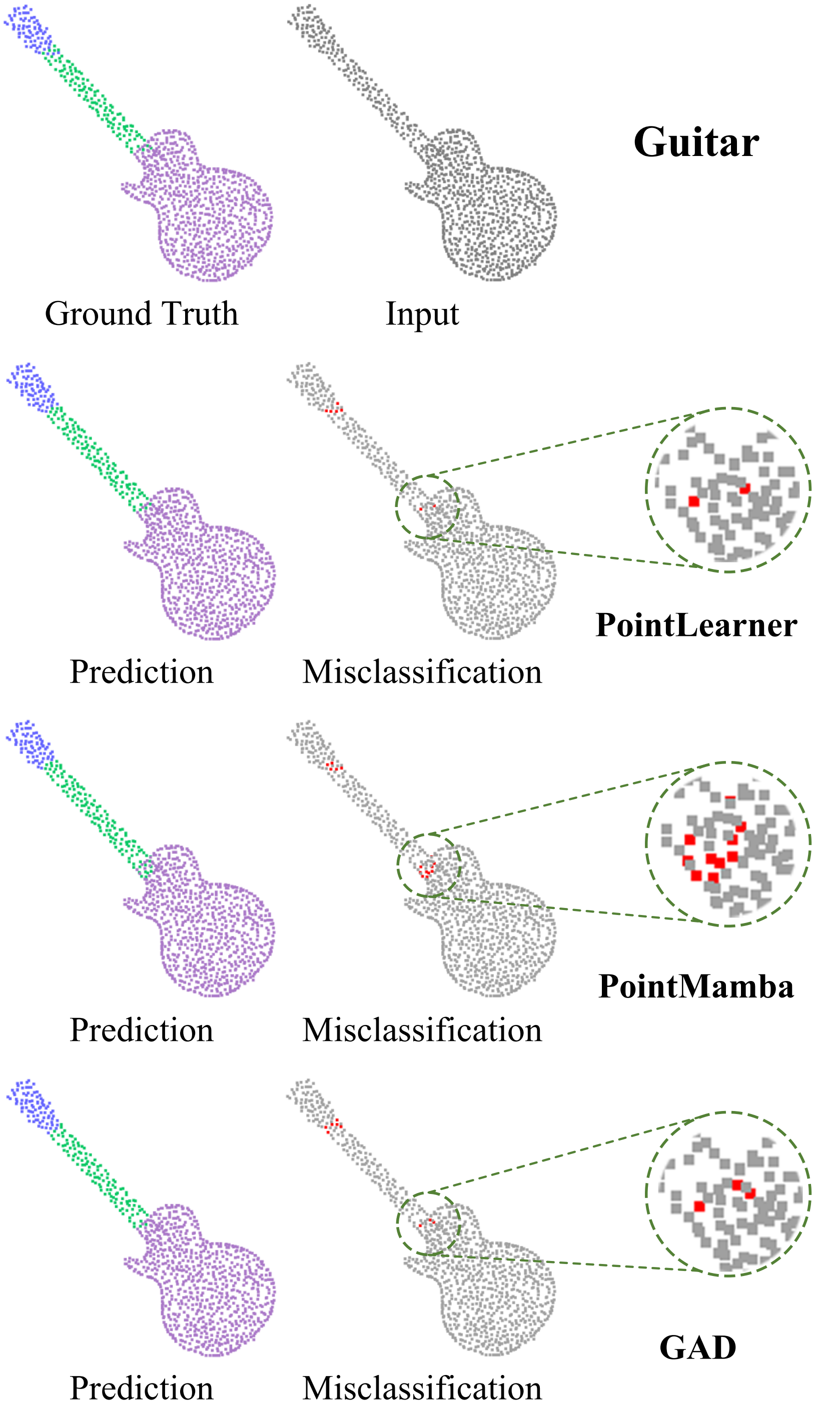


Figure 8: Visualization results of PointLearner, PointMamba, and GAD.