ZigzagPointMamba: Spatial-Semantic Mamba for Point Cloud Understanding

Linshuang Diao1,2,Sensen Song1,2,* Yurong Qian1,2, Dayong Ren3,*

1 Key Laboratory of Signal Detection and Processing, Xinjiang University, Urumqi, 830046, China.

2 Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing. Xinjiang University,Urumqi Xinjiang 830046,China.

3 National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China.

Background

While State Space Models (SSMs) like PointMamba achieve efficient point cloud feature extraction with linear complexity O(n), existing methods face two critical limitations: (1) traditional scanning patterns (Random, Hilbert, Z-order) disrupt spatial continuity, producing disjointed token sequences that impair feature quality, and (2) random masking strategies rely solely on local neighbors for reconstruction, failing to capture global semantic dependencies crucial for segmentation and classification. To address these challenges, we propose ZigzagPointMamba, which introduces a novel zigzag scan path that preserves spatial proximity and a Semantic-Siamese Masking Strategy (SMS) that enables robust global semantic modeling.

Contribution

To address these challenges, our contributions are summarized as follows:

1. Proposes a simple yet effective zigzag scanning pattern that preserves spatial proximity during token sequencing, generating smoother and spatially coherent token sequences for enhanced feature representations.
2. Introduces a threshold-based masking approach that targets semantically similar tokens instead of random masking, enabling robust global semantic modeling and superior reconstruction quality.
3. Combines the advantages of zigzag scanning and SMS to significantly advance point cloud analysis, providing strong pre-trained weights for downstream tasks.

Methods

Architecture Overview

Point cloud analysis faces challenges from unstructured data and inefficient scanning patterns that disrupt spatial continuity, leading to suboptimal feature representations. Traditional scanning methods (random, Hilbert) create disjointed token sequences, while random masking strategies fail to capture global semantic dependencies. To address this, we propose ZigzagPointMamba, combining a novel zigzag scan path, Semantic-Siamese Masking Strategy (SMS), and Mamba-based MAE architecture. The zigzag scan preserves spatial proximity through coordinate-based layering across XY/XZ/YZ planes, while SMS identifies and masks semantically redundant tokens (threshold 0.8) to force global semantic learning. This design effectively balances spatial continuity and semantic modeling, achieving superior performance in downstream classification and segmentation tasks.

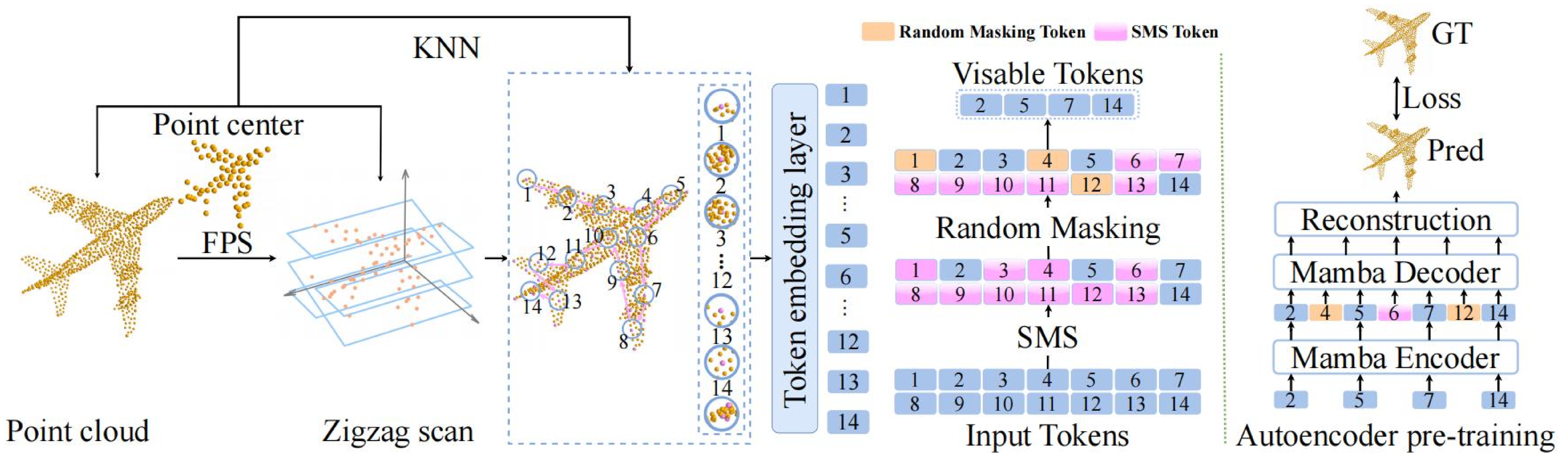
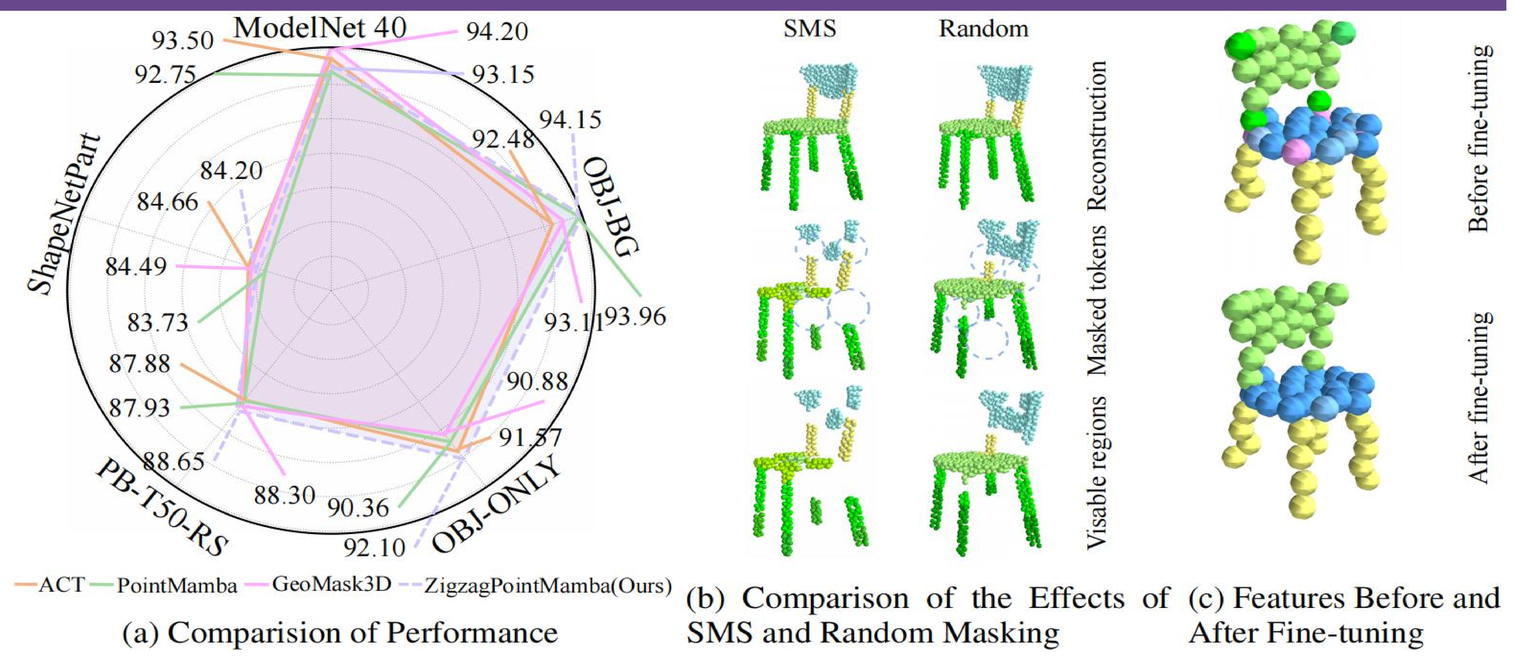


Figure 2: ZigzagPointMamba pre-training pipeline.



(a) Comparison of Performance (b) Comparison of the Effects of SMS and Random Masking (c) Features Before and After Fine-tuning

Figure 1: Performance Comparison and Feature Quality Analysis of ZigzagPointMamba.

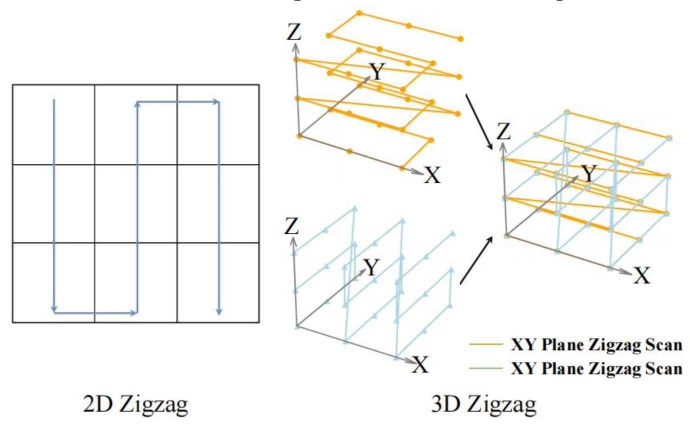


Figure 3: Comparison of 2D and 3D Zigzag Scan Paths.

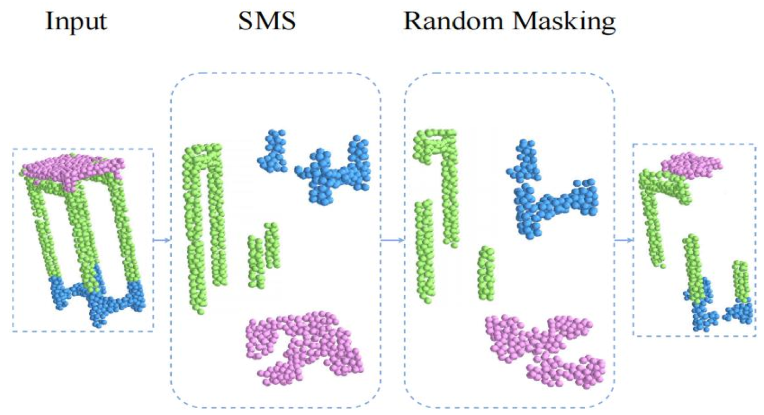


Figure 4: Two-Stage Masking Pipeline: SMS followed by Random Masking.

Algorithm 1 Semantic-Siames Masking Strategy

Input: *group_input_tokens*: point cloud feature tensor with shape B, G, C (batch_size, number of groups, feature_dimension).
threshold: SMS retention *threshold* (default 0.8), controlling the proportion of tokens to retain.
Output: *bool_masked_pos*: boolean mask tensor with shape B, G, C, indicating which tokens are masked.

1: $B, G, C \leftarrow$shape($group\_input\_tokens$) // Get tensor dimensions
2: $tokens\_norm \leftarrow$ F.normalize($group\_input\_tokens$, dim $= -1$) // Normalize feature vectors to unit length
3: $similarity\_matrix \leftarrow$ torch.bmm($tokens\_norm, tokens\_norm^T$).clamp(0, 1) // Compute cosine similarity matrix and clamp to [0,1]
4: $redundancy\_score \leftarrow \sum_{\text{dim}=-1}(similarity\_matrix)$ // Calculate redundancy score for each token
5: $k \leftarrow \max(1, \lfloor threshold \times G \rfloor)$ // Determine number of tokens to retain (at least 1)
6: **if** $k = 0$
return torch.zeros($[B, G]$, dtype = torch.bool)
7: $thresholds \leftarrow$ torch.topk($redundancy\_score, k = k$, largest = torch.False).$values[:, -1]$ // Get k-th smallest redundancy score as threshold
8: $bool\_masked\_pos \leftarrow redundancy\_score > thresholds$ // Generate mask (tokens with higher redundancy are masked)
9: **return** $bool\_masked\_pos$

Module 1: Zigzag Scan Path

Purpose

Preserve spatial continuity during point cloud serialization, addressing the limitation of traditional scanning methods (random, Hilbert) that disrupt local geometric coherence.

How It Works

FPS Sampling: Select M representative keypoints from input point cloud
Multi-Plane Layering: Divide points into layers along X, Y, Z axes
Sequential Connection: Merge layered paths into spatially coherent token sequences

Advantages

- Maintains proximity of spatially adjacent points
- Generates smoother token sequences (vs. random/Hilbert)
- Improves feature representation quality

Module 2: Semantic-Siamese Masking Strategy (SMS)

Purpose

Force global semantic learning by masking semantically redundant tokens instead of random selection.

How It Works

Compute cosine similarity between token pairs
Calculate redundancy score per token via similarity aggregation
Mask tokens exceeding threshold τ=0.8 + random mask (ratio 0.6)

Advantages

- Targets semantically similar regions (e.g., complete object parts)
- Preserves topological integrity during masking
- Superior reconstruction quality (vs. random masking)

Results

Datasets:

ScanObjectNN: The ScanObjectNN dataset contains 2902 real-world 3D object scans from indoor scenes, covering 15 categories, with three difficulty variants: OBJ-BG, OBJ-ONLY, and PB-T50-RS. Each object is represented as a 1024-point surface-sampled point cloud, providing object classification labels under occluded and noisy realistic conditions.

ModelNet40: The ModelNet40 benchmark includes 12311 CAD models from online repositories, spanning 40 categories (e.g., furniture, vehicles). It is split into 9843 training and 2468 testing samples, with each 3D model converted to a point cloud (1024/2048 points) for standard classification performance evaluation.

ShapeNetPart: The ShapeNetPart segmentation dataset has 16881 3D shapes from the ShapeNet repository, covering 16 categories (e.g., airplanes, chairs), with 50 fine-grained part labels. Each shape is a 2048-point sampled point cloud, offering dense semantic labels to assess part segmentation capabilities.

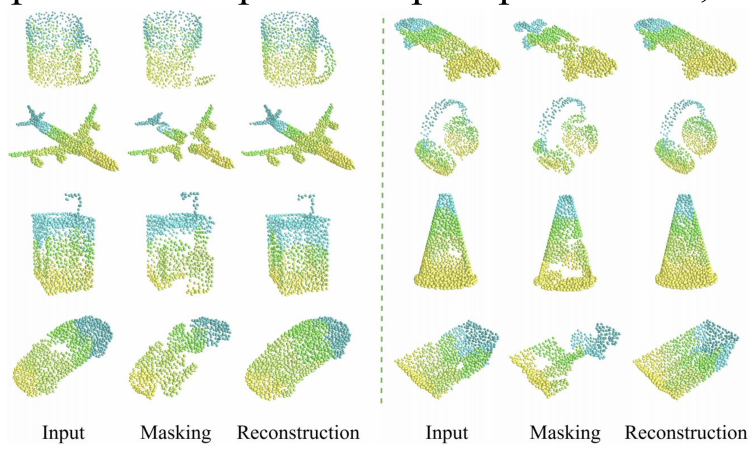


Fig 5: This figure shows the qualitative analysis results of the mask predictions made by the ZigzagPointMamba model on the ShapeNet validation set.

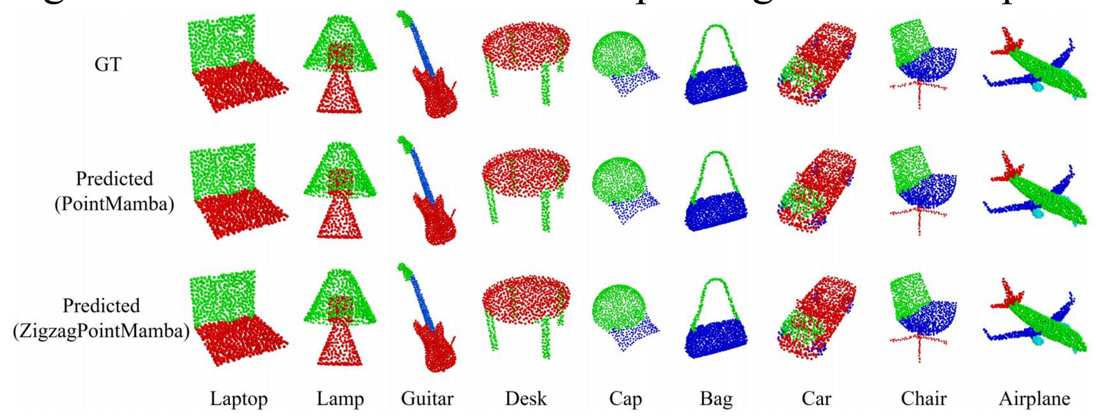


Fig 6: The qualitative outcomes of part segmentation achieved by our ZigzagPointMamba model on the ShapeNetPart dataset.

Table 6: The effect of the thresholds of different SMS.

Setting	OA(%) OBJ-ONLY	OA(%) PB-T50-RS
0.5	90.71	87.99
0.6	91.57	87.68
0.7	91.22	88.45
0.8	**92.08**	**88.65**
0.9	91.91	88.36

Table 7: The effect of different attention.

Setting	OA(%) OBJ-ONLY	OA(%) PB-T58-RS
Attention	91.22	88.17
Multi-Attention	90.53	87.79
SMS	**92.08**	**88.51**

Table 1: Object Classification on ScanObjectNN Dataset. We conducted experiments on three subsets of theScanObjectNN dataset: the OBJ-BG subset, OBJ-ONLY subset, and PB-T50-RS subset.

Methods	Reference	Param.(M)	FLOPs(G)	OBJ-BG	OBJ-ONLY	PB-T50-RS
Point-Bert[39]	CVPR 22	22.1	4.8	87.3	88.12	83.07
MaskPoint[19]	CVPR 22	22.1	4.8	89.70	89.30	84.60
PointMAE[24]	ECCV 22	22.1	4.8	90.02	88.29	84.60
PointM2AE[41]	NeurIPS 22	15.3	3.6	91.22	88.81	86.43
ACT[8]	ICLR 23	22.1	4.8	93.29	91.91	88.21
ReCon[27]	ICML 23	43.6	5.3	**94.15**	**93.12**	**89.73**
GeoMask3D[2]	TMLR 25	-	-	93.11	90.36	88.30
PointMamba([18])(baseline)	NeurIPS 24	**12.3**	**3.1**	93.96	90.88	87.93
ZigzagPointMamba(Ours)		**12.3**	**3.1**	**94.15**	92.10	88.65

Table 2: Classification on ModelNet40 Dataset. We report the overall accuracy from 1024 points without voting

Methods	Reference	Param.(M)	FLOPs(G)	OA(%)
Point-Bert[39]	CVPR 22	22.1	4.8	92.7
MaskPoint[19]	CVPR 22	22.1	4.8	92.6
PointMAE[24]	ECCV 22	22.1	4.8	93.2
PointM2AE[41]	NeurIPS 22	15.3	3.6	93.4
ACT[8]	TCLR 23	22.1	4.8	93.6
GeoMask3D[2]	TMLR 25	-	-	**94.20**
PointMamba([18])(baseline)	NeurIPS 24	12.3	1.5	92.75
ZigzagPointMamba(Ours)		**12.3**	**1.5**	93.15

Table 3: Part Segmentation on ShapeNetPart Dataset. The mIoU of all classes (Cls.) and instances (Inst.) is reported.

Methods	Reference	Inst.mIoU	Cls.mIoU
Point-BERT [39]	TMLR 25	85.6	84.1
MaskPoint[19]	TMLR 25	86.0	84.4
PointMAE[24]	ECCV 22	86.1	84.1
PointM2AE[41]	NeurIPS 22	**86.51**	**84.86**
ACT[8]	ICLR 23	86.14	84.66
GeoMask3D[2]	TMLR 25	86.04	84.49
PointMamba([18])(baseline)	NeurIPS 24	85.28	82.57
ZigzagPointMamba(Ours)		85.78	84.16

Our method uses 17.36M parameters and 5.5G FLOPs.

Table 5: The effect of different scanning curves.

Scanning curve	OBJ-ONLY	PB-T58-RS
Random	92.60	90.18
Z-order and Trans-Z-order	**93.29**	90.36
Hilbert and Z-order	**93.29**	90.88
Trans-Hilert and Trans-Z-order	**93.29**	**91.91**
Hilbert and Trans-Hilbert	90.88	87.93
zigzag scan path (Ours)	92.10	88.65

Table 4: Few-shot learning on ModelNet40. A dedi_x0002_cated dataset for few-shot learning constructed based on ModelNet40.

Methods	Reference	5-way 10-shot	5-way 20-shot	10-way 10-shot	10-way 20-shot
Point-Bert[39]	CVPR 22	94.6±3.0	96.3±2.5	91.0±5.0	92.7±4.8
MaskPoint[19]	CVPR 22	95.0±3.7	97.2±1.5	91.4±4.5	93.4±3.2
PointMAE[24]	ECCV 22	96.3±3.1	97.8±1.8	92.6±4.0	95.0±2.8
PointM2AE[41]	NeurIPS 22	96.8±2.0	98.3±1.5	92.3±4.2	95.2±2.5
ACT[8]	ICLR 23	96.8±2.1	98.0±1.5	93.3±4.0	95.6±3.0
PointGPT-S[7]	NeurIPS 23	96.8±1.8	98.6±1.2	92.6±3.5	95.2±2.5
ReCon[27]	ICML 23	**97.3±1.8**	98.0±1.5	**93.3±4.3**	**95.8±2.8**
PointMamba([18])(baseline)	NeurIPS 24	96.0±2.0	**99.0±1.0**	88.5±2.4	93.8±1.2
ZigzagPointMamba(Ours)		96.0±2.1	**99.0±1.2**	90.0±2.2	94.2±1.0

Conclusion

In this paper, we introduced ZigzagPointMamba, an innovative state-space model that addresses critical limitations in existing PointMamba-based approaches for point cloud self-supervised learning. Extensive experiments demonstrate that our zigzag scan path preserves spatial continuity while the SMS helps the model focus on global structures, preventing over-reliance on local features. ZigzagPointMamba provides a powerful pre-trained backbone that effectively supports downstream point cloud analysis tasks.

Wetchat