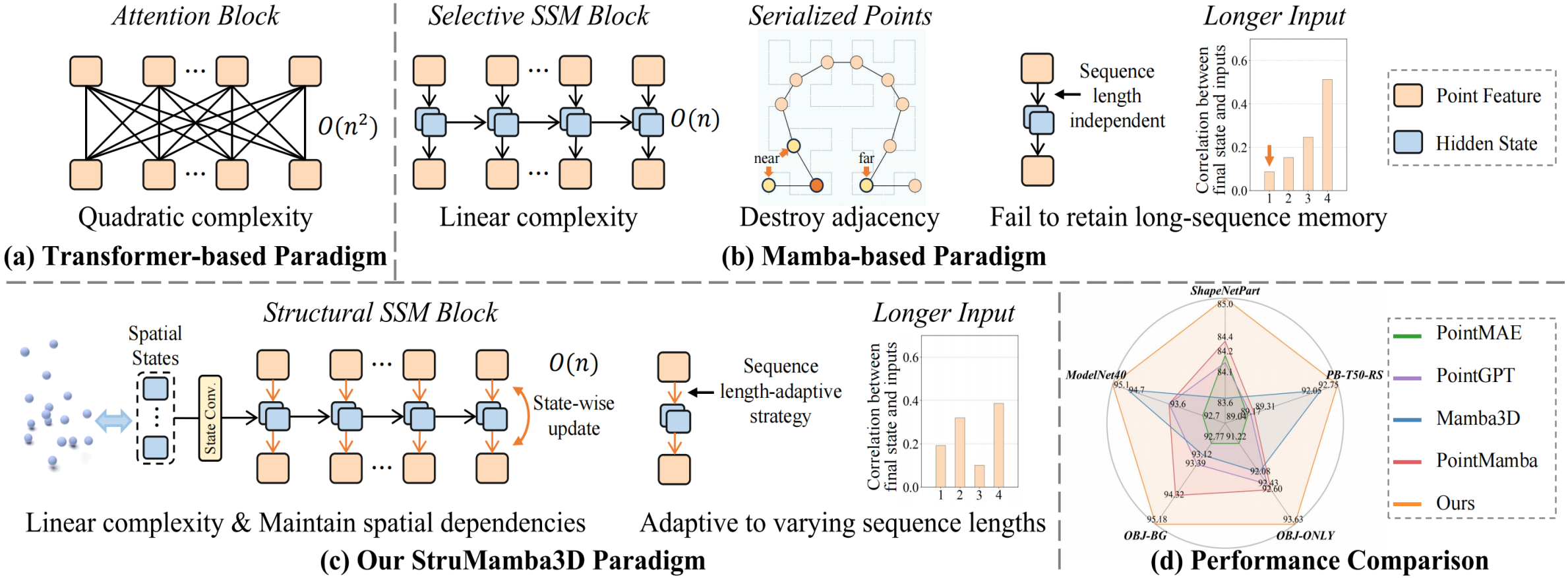
StruMamba3D: Exploring Structural Mamba for Self-supervised Point Cloud Representation Learning

Chuxin Wang[1,2], *Yixin Zha*[2], *Wenfei Yang*[1,2], *Tianzhu Zhang*[1,2]

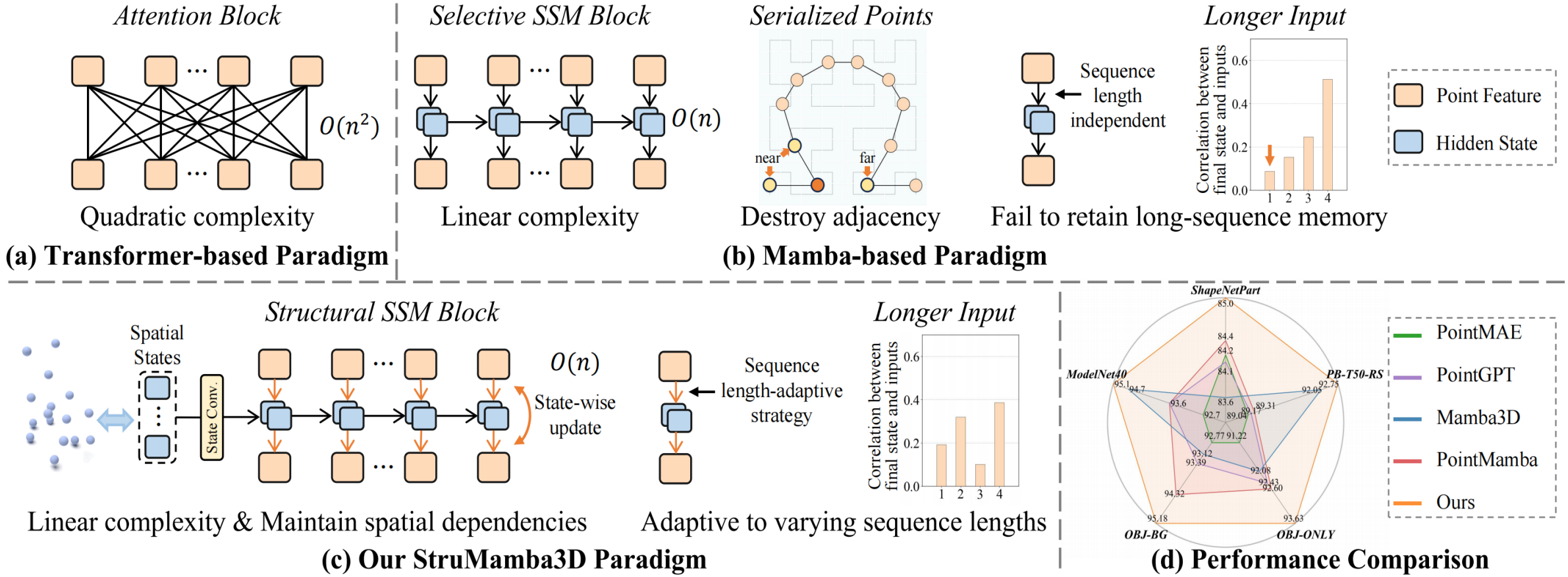
[1]University of Science and Technology of China

[2]National Key Laboratory of Deep Space Exploration, Deep Space Exploration Laboratory

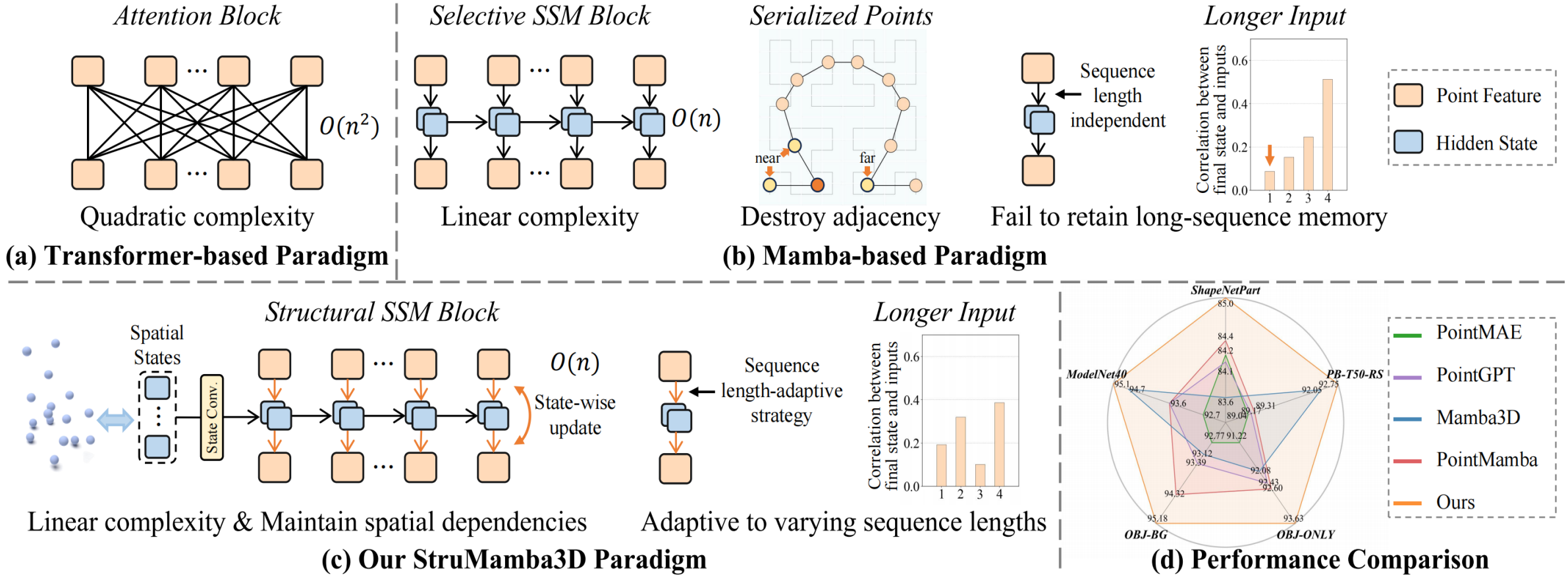
❖ **Mamba-based paradigm significantly reduces computational cost compared to the Transformer-based paradigm. This results in better scalability and efficiency for large-scale data processing.**



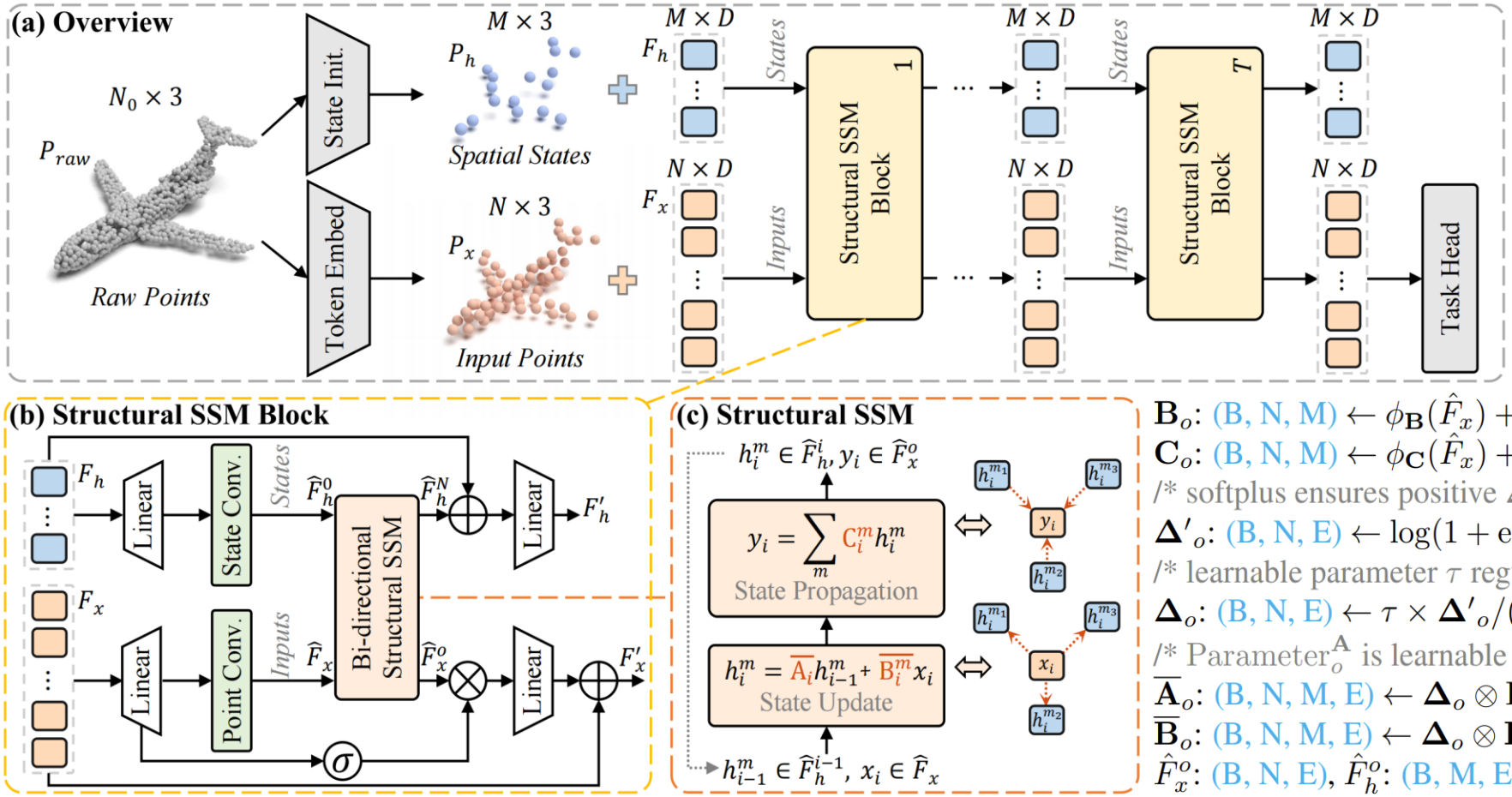
❖ **Mamba-based methods face challenges, such as serialized points destroying the adjacency of 3D points, and its pre-trained selection mechanism struggles to retain long-sequence memory.**



❖ **Our StruMamba3D paradigm uses the structural SSM to maintain the spatial dependencies among points and the sequence length-adaptive strategy to retain long-sequence memory.**



- **Hidden states are assigned real spatial positions, and their update is conditioned on parameters generated from spatial relationships, enabling preservation of spatial structures and effective feature modeling.**



$\mathbf{B}_o$: (B, N, M) $\leftarrow \phi_{\mathbf{B}}(\hat{F}_x) + \mathrm{MLP}_{\mathbf{B}}(\triangle P)$
$\mathbf{C}_o$: (B, N, M) $\leftarrow \phi_{\mathbf{C}}(\hat{F}_x) + \mathrm{MLP}_{\mathbf{C}}(\triangle P)$
/* softplus ensures positive $\boldsymbol{\Delta}_o$ */
$\boldsymbol{\Delta}'_o$: (B, N, E) $\leftarrow \log(1 + \exp(\phi_{\boldsymbol{\Delta}}(\hat{F}_x)))$
/* learnable parameter τ regulates the total sampling time $\boldsymbol{\Delta}_{all}$ */
$\boldsymbol{\Delta}_o$: (B, N, E) $\leftarrow \tau \times \boldsymbol{\Delta}'_o / (\sum_{i=1}^{N} \boldsymbol{\Delta}'^{i}_o)$
/* $\mathrm{Parameter}_o^{\mathbf{A}}$ is learnable parameter: (M, E) */
$\overline{\mathbf{A}}_o$: (B, N, M, E) $\leftarrow \boldsymbol{\Delta}_o \otimes \mathrm{Parameter}_o^{\mathbf{A}}$
$\overline{\mathbf{B}}_o$: (B, N, M, E) $\leftarrow \boldsymbol{\Delta}_o \otimes \mathbf{B}_o$
$\hat{F}_x^o$: (B, N, E), $\hat{F}_h^o$: (B, M, E) $\leftarrow \mathbf{SSM}(\overline{\mathbf{A}}_o, \overline{\mathbf{B}}_o, \mathbf{C}_o)(\hat{F}_x, \hat{F}_h)$

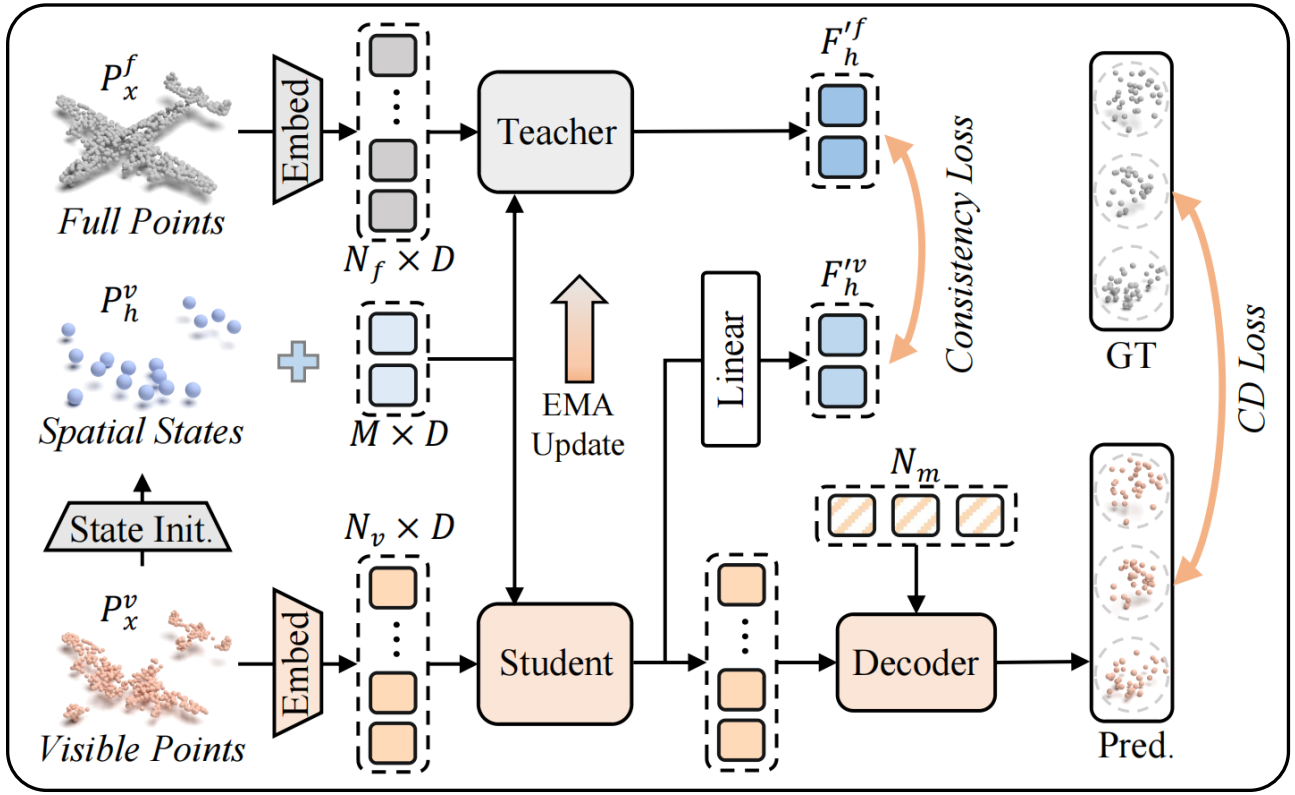
❖ **Our method incorporates a dynamic state update strategy and a spatial state consistency loss, which strengthen the model long-sequence memory capability and improve its robustness to variable input lengths.**

Discrete SSM:

$$h_t = \overline{\mathbf{A}} h_{t-1} + \overline{\mathbf{B}} x_t,\ y_t = \mathbf{C} h_t,$$
$$\overline{\mathbf{A}} = \exp(\mathbf{\Delta A}),\ \overline{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I})\mathbf{\Delta B},$$

Adaptive State Update:

$$\mathbf{\Delta}_i = \frac{\tau \times \mathbf{\Delta}_i}{\sum_{i=1}^{N} \mathbf{\Delta}_i}.$$

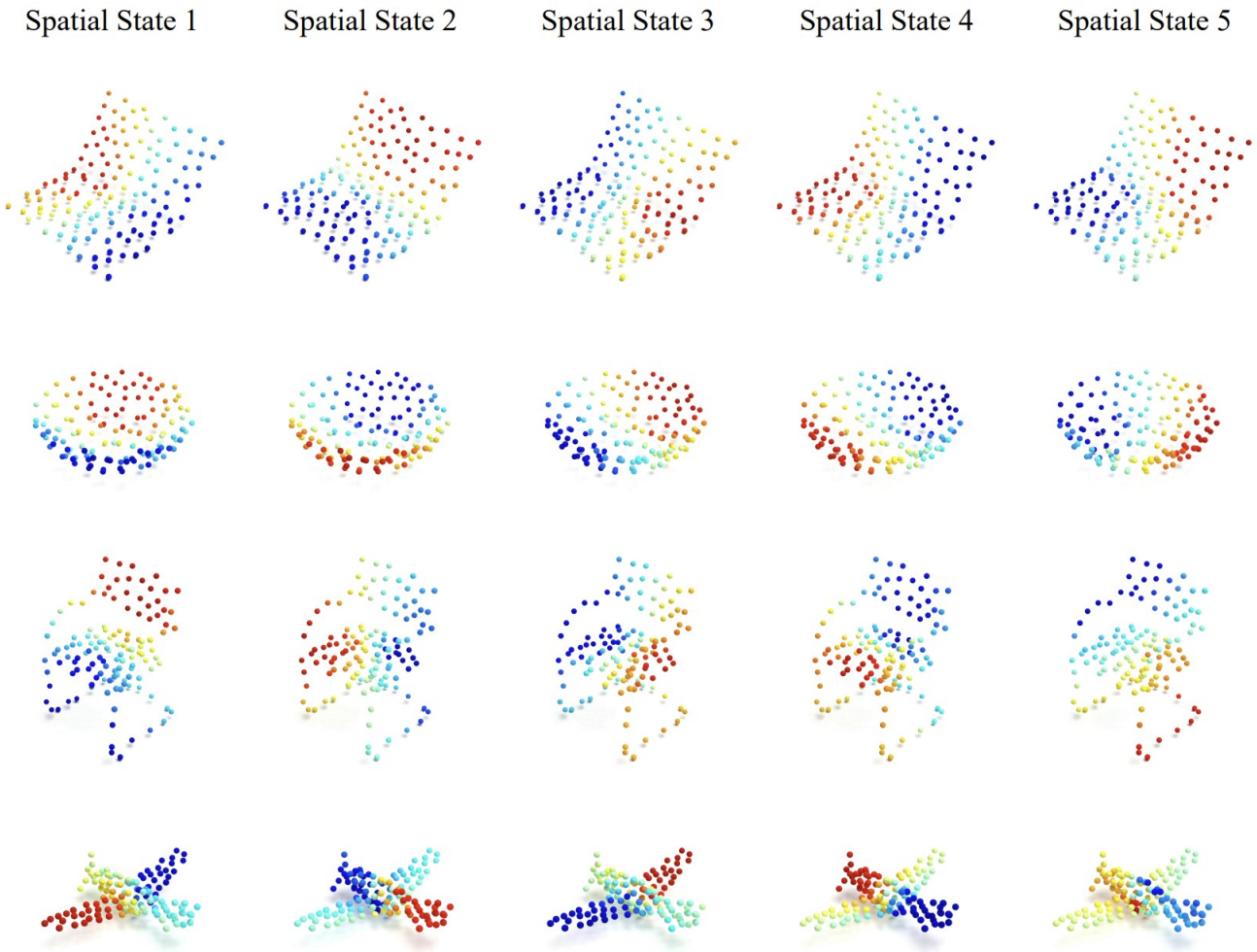


Spatial state modeling significantly enhances the representation of point cloud structural features without requiring sequence serialization.

Method	Overall Accuracy		mIoU_c
	ScanNN	MN40	SNPart
baseline	88.24	92.50	82.08
w/ Structural SSM	91.78	93.92	84.15
w/ Lightweight Conv. for h	92.22	94.65	84.62
w/ Lightweight Conv. for x	92.40	94.81	84.77
w/ Bidirectional Scanning	**92.75**	**95.06**	**84.96**

$\phi(x)$	SS	MLP($\triangle P$)	Overall Accuracy		mIoU_c
			ScanNN	MN40	SNPart
✓	✗	✗	90.94	93.84	83.62
✓	✓	✗	91.33	94.12	83.87
✓	✓	✓	**92.75**	**95.06**	**84.96**
✗	✓	✓	91.12	94.25	84.02

Backbone	Serialization	Overall Accuracy		mIoU_c
		ScanNN	MN40	SNPart
Mamba	✗	88.24	92.50	82.08
Mamba	Z-order	89.28	93.44	83.89
Mamba	Hilbert	89.87	93.68	84.07
StruMamba3D	✗	<u>92.75</u>	**95.06**	<u>84.96</u>
StruMamba3D	Z-order	92.47	<u>94.98</u>	84.86
StruMamba3D	Hilbert	**92.81**	94.89	**85.12**

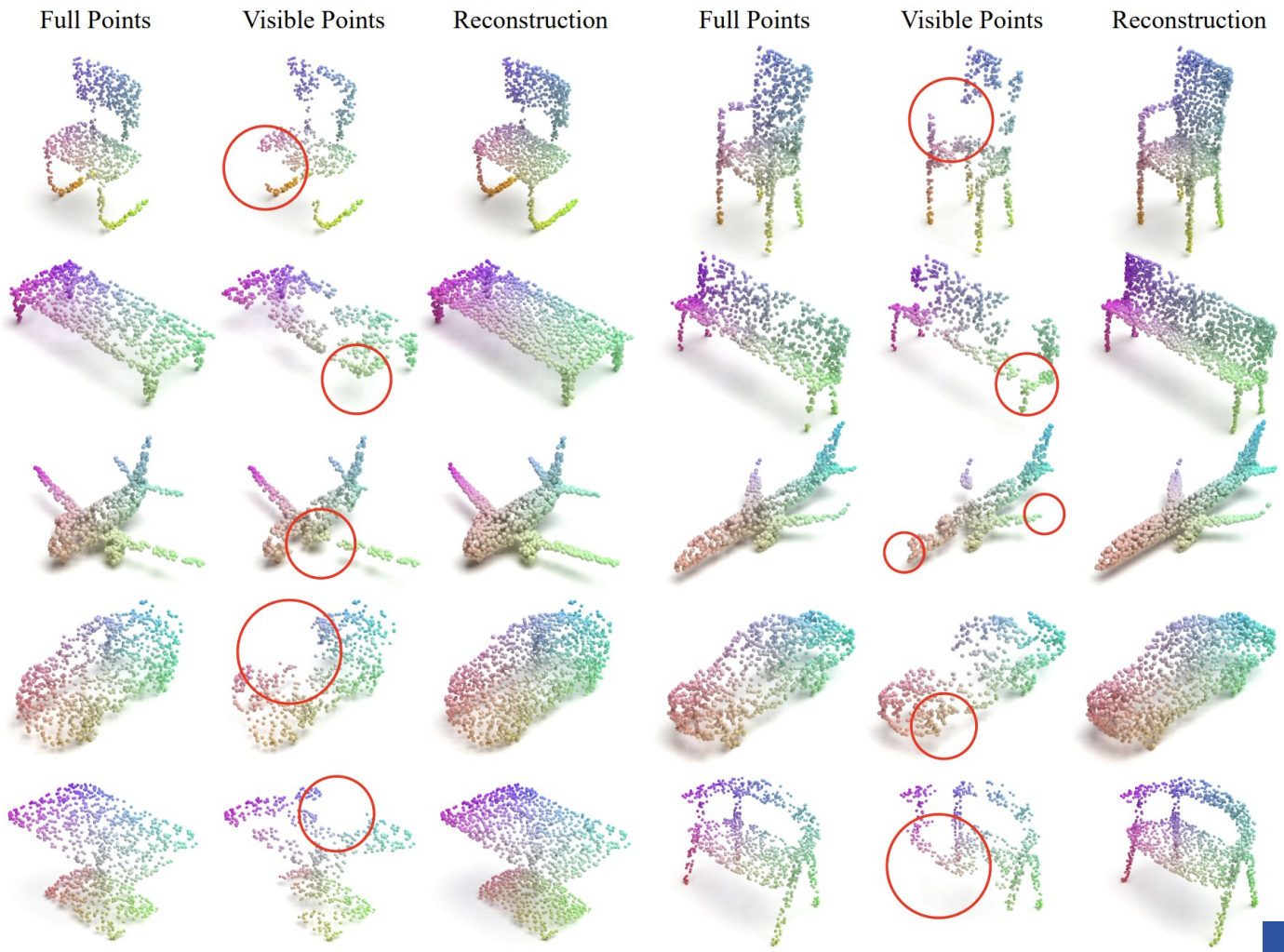


The sequence length adaptation strategy enhances the long-sequence memory capability of the pretrained model in downstream tasks.

Structural SSM Block	SLAS	Overall Accuracy		mIoU_c
		ScanNN	MN40	SNPart
✗	✗	87.23	91.86	81.56
✓	✗	92.09	94.45	84.49
✓	✓	**92.75**	**95.06**	**84.96**

ASUM	$\mathcal{L}_{ssc}$	Overall Accuracy		
		ScanNN	MN40	SNPart
✗	✗	92.09	94.45	84.49
✓	✗	92.26	94.57	84.56
✓	$\lambda = 1$	92.47	94.94	84.81
✓	$\lambda = 2$	**92.75**	**95.06**	**84.96**
✓	$\lambda = 5$	92.57	94.89	84.77
✗	$\lambda = 2$	92.16	94.65	84.56

Method	ScanObjectNN mOA	ModelNet40 mOA	ShapeNetPart mIoU_c
w/o Pretraining	91.33	93.68	83.96
MPM Pretraining	92.09	94.45	84.49
Our Pretraining	**92.75**	**95.06**	**84.96**

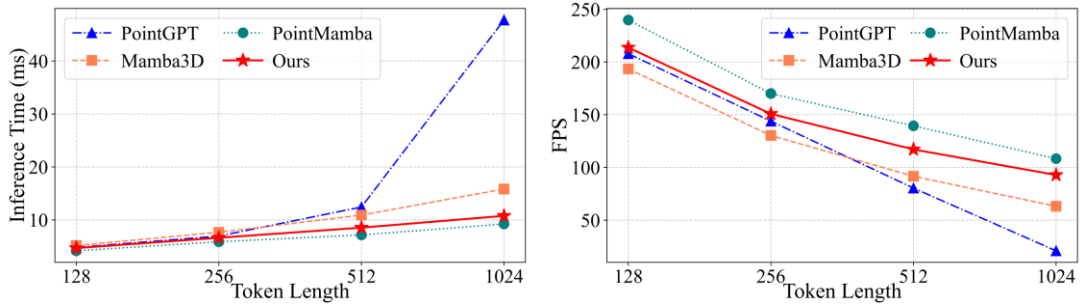


多项下游任务中刷新 SOTA，具备线性复杂度与良好的可扩展性

分类任务

Method	Backbone	Param. (M)	FLOPs (G)	ScanObjectNN			MODELNET40	
				OBJ-BG	OBJ-ONLY	PB-T50-RS	w/o Voting	w/ Voting
Supervised learning only								
PointNet [25]	-	3.5	0.5	73.3	79.2	68.0	89.2	-
PointNet++ [26]	-	1.5	1.7	82.3	84.3	77.9	90.7	-
PointCNN [18]	-	0.6	-	86.1	85.5	78.5	92.2	-
DGCNN [31]	-	1.8	2.4	82.8	86.2	78.1	92.9	-
PRANet [8]	-	2.3	-	-	-	81.0	93.7	-
PointNeXt [28]	-	1.4	3.6	-	-	87.7	94.0	-
PointMLP [22]	-	12.6	31.4	-	-	85.4	94.5	-
DeLA [2]	-	5.3	1.5	-	-	88.6	94.0	-
PCM [42]	-	34.2	45.0	-	-	88.1	93.4	-
Pre-training using single-modal information								
PointBERT [38]	Transformer	22.1	4.8	87.43	88.12	83.07	92.7	93.2
MaskPoint [21]	Transformer	22.1	4.8	89.30	88.10	84.30	-	93.8
PointM2AE [39]	Transformer	12.7	7.9	91.22	88.81	86.43	92.9	93.4
PointMAE† [23]	Transformer	22.1	4.8	92.77	91.22	89.04	92.7	93.8
PointGPT-S† [3]	Transformer	29.2	5.7	93.39	92.43	89.17	93.3	94.0
PointMamba† [20]	Mamba	12.3	3.1	94.32	92.60	89.31	93.6	94.1
Mamba3D† [15]	Mamba	16.9	3.9	93.12	92.08	92.05	94.7	95.1
Ours†	Structural SSM	15.8	4.0	**95.18**	**93.63**	**92.75**	**95.1**	**95.4**
Pre-training using cross-modal information								
ACT† [7]	Transformer	22.1	4.8	93.29	91.91	88.21	93.7	94.0
Joint-MAE [14]	Transformer	22.1	-	90.94	88.86	86.07	-	94.0
I2P-MAE† [41]	Transformer	15.3	-	94.15	91.57	90.11	93.7	94.1
ReCon† [27]	Transformer	43.6	5.3	95.18	93.29	90.63	94.5	94.7

计算开销对比



分割任务

Method	Architecture	$mIoU_c$	$mIoU_i$
MaskPoint [21]	Single-scale	84.6	86.0
PointBERT [38]	Single-scale	84.1	85.6
PointMAE [23]	Single-scale	84.2	86.1
PointM2AE [39]	Multi-scale	84.9	86.5
PointGPT-S [3]	Single-scale	84.1	86.2
PointMamba [20]	Single-scale	84.4	86.2
Mamba3D [15]	Single-scale	83.6	85.6
Ours	Single-scale	**85.0**	**86.7**

少样本分类任务

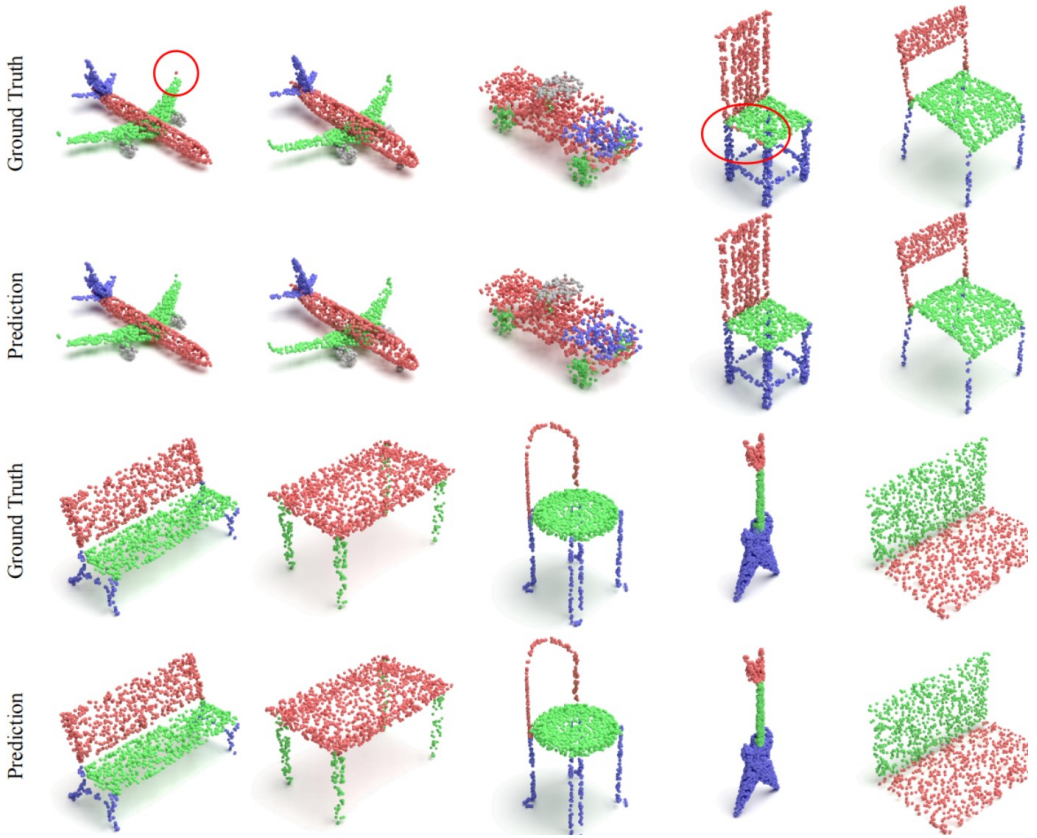
Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
PointBERT [38]	94.6±3.6	93.9±3.1	86.4±5.4	91.3±4.6
MaskPoint [21]	95.0±3.7	97.2±1.7	91.4±4.0	92.7±5.1
PointMAE [23]	96.3±2.5	97.8±1.8	92.6±4.1	93.4±3.5
PointM2AE [39]	96.8±1.8	98.3±1.4	92.3±4.5	95.0±3.0
PointGPT-S [3]	96.8±2.0	98.6±1.1	92.6±4.6	95.2±3.4
PointMamba [20]	96.9±2.0	99.0±1.1	93.0±4.4	95.6±3.2
Mamba3D [15]	96.4±2.2	98.2±1.2	92.4±4.1	95.2±2.9
Ours	**97.5**±2.3	**99.1**±1.4	**93.5**±3.7	**96.1**±3.5

(A) Classification on ScanObjectNN and ModelNet40 Datasets

Method	Backbone	Param. (M)	FLOPs (G)	ScanObjectNN			MODELNET40	
				OBJ-BG	OBJ-ONLY	PB-T50-RS	w/o Voting	w/ Voting
Supervised learning only								
PointNet [25]	-	3.5	0.5	73.3	79.2	68.0	89.2	-
PointNet++ [26]	-	1.5	1.7	82.3	84.3	77.9	90.7	-
PointCNN [18]	-	0.6	-	86.1	85.5	78.5	92.2	-
DGCNN [31]	-	1.8	2.4	82.8	86.2	78.1	92.9	-
PRANet [8]	-	2.3	-	-	-	81.0	93.7	-
PointNeXt [28]	-	1.4	3.6	-	-	87.7	94.0	-
PointMLP [22]	-	12.6	31.4	-	-	85.4	94.5	-
DeLA [2]	-	5.3	1.5	-	-	88.6	94.0	-
PCM [42]	-	34.2	45.0	-	-	88.1	93.4	-
Pre-training using single-modal information								
PointBERT [38]	Transformer	22.1	4.8	87.43	88.12	83.07	92.7	93.2
MaskPoint [21]	Transformer	22.1	4.8	89.30	88.10	84.30	-	93.8
PointM2AE [39]	Transformer	12.7	7.9	91.22	88.81	86.43	92.9	93.4
PointMAE† [23]	Transformer	22.1	4.8	92.77	91.22	89.04	92.7	93.8
PointGPT-S† [3]	Transformer	29.2	5.7	93.39	92.43	89.17	93.3	94.0
PointMamba† [20]	Mamba	12.3	3.1	94.32	92.60	89.31	93.6	94.1
Mamba3D† [15]	Mamba	16.9	3.9	93.12	92.08	92.05	94.7	95.1
Ours†	Structural SSM	15.8	4.0	**95.18**	**93.63**	**92.75**	**95.1**	**95.4**
Pre-training using cross-modal information								
ACT† [7]	Transformer	22.1	4.8	93.29	91.91	88.21	93.7	94.0
Joint-MAE [14]	Transformer	22.1	-	90.94	88.86	86.07	-	94.0
I2P-MAE† [41]	Transformer	15.3	-	94.15	91.57	90.11	93.7	94.1
ReCon† [27]	Transformer	43.6	5.3	95.18	93.29	90.63	94.5	94.7

(B) Segmentation on ShapeNetPart Dataset

Method	Architecture	$mIoU_c$	$mIoU_i$
Supervised learning only			
PointNet [25]	Single-scale	80.4	83.7
PointNet++ [26]	Multi-scale	81.9	85.1
APES [32]	Multi-scale	83.7	85.8
DeLA [2]	Multi-scale	85.8	87.0
PCM [42]	Multi-scale	85.3	87.0
Pre-training using single-modal information			
MaskPoint [21]	Single-scale	84.6	86.0
PointBERT [38]	Single-scale	84.1	85.6
PointMAE [23]	Single-scale	84.2	86.1
PointM2AE [39]	Multi-scale	84.9	86.5
PointGPT-S [3]	Single-scale	84.1	86.2
PointMamba [20]	Single-scale	84.4	86.2
Mamba3D [15]	Single-scale	83.6	85.6
Ours	Single-scale	**85.0**	**86.7**



(C) Linear Complexity of Our StruMamba3D