Efficient Spiking Point Mamba for Point Cloud Analysis

Supplementary Material

In the supplementary material, we further explore the behavior of Farthest Point Sampling (FPS) and experimentally validate the theoretical foundation of HDE. We present more ablation studies on neurons and mask ratios, as well as comparing our SPM with additional models [41]. Next, we provide more implementation details of our SPM. Finally, we include additional visualizations across various tasks.

A. Farthest Point Sampling Behavior

Farthest Point Sampling (FPS) is a widely used technique for point cloud analysis, which select a subset of points from a larger point cloud in a way that maximizes the minimum distance between selected points. FPS can be divided into three key stages: early, middle, and late stages.

Early Stage. The early stage of FPS involves random initial selection, which can cause instability in the sampling process. The points selected in this stage are heavily influenced by the random choice, leading to a not so well distribution across the point cloud. This randomness results in high variance, making the early stage less reliable for capturing the overall structure of the point cloud. Together, the early stage is unstable due to random initial selection.

Middle Stage. As the algorithm progresses, the middle stage becomes more stable. FPS starts to cover important features of the geometry, improving the distribution of sampled points. The algorithm focuses on regions with significant features while maintaining a good spread across the point cloud. This stability enables a more meaningful representation of the point cloud. Together, the middle stage stabilizes and captures the skeletal structure.

Late Stage. In the late stage, FPS starts to experience diminishing returns. As the number of selected points increases, the algorithm starts to introduce redundancy or noise. The points selected in this stage are often located in areas that are already well-represented by previous selections, leading to overlapping regions.Together, the late stage may introduce redundancy or noise.

We use Chamfer Distance to measure the similarity between the early, middle, and late stages under different random seeds in Tab. 9, and visualize the sampling structures of the early, middle, and late stages under certain random seeds for quantitative comparison in Fig. 6, thereby demonstrating the rationale behind the hierarchical stages.

From Fig. 6, we can see that in the early stage under different random seeds, FPS is unstable and does not fully represent the object information, indicating a strong influence from the random initial points. In the middle stage, the points show little visual change and provide a stable representation of the entire point cloud. However, in the late stage, the point cloud no longer captures the overall structure, exhibiting a highly variable shape, and can only be considered as redundant points.

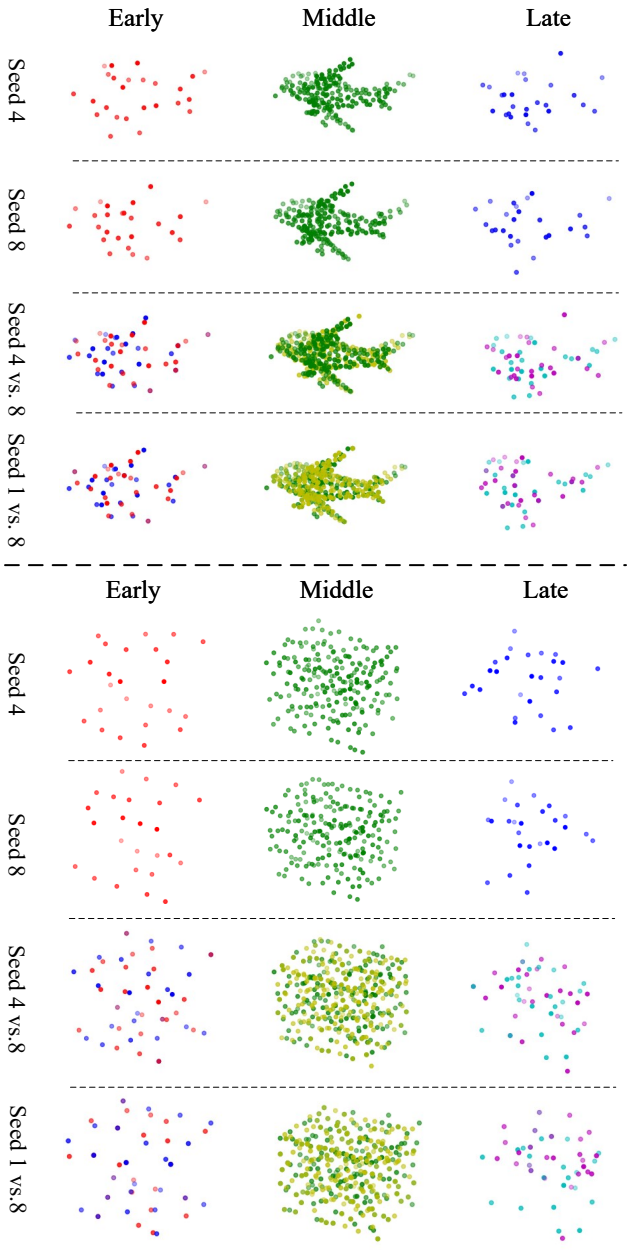

Figure 6. Quantitative comparison of the sampling structures of the early, middle, and late stages under different random seeds.

Configuration	Pre-training	Classification		Segmentation
	ShapeNet	ModelNet40	ScanObjectNN	ShapeNetPart
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	1e-3	1e-3	1e-3	2e-4
Weight decay	5e-2	5e-2	5e-2	5e-2
Learning rate scheduler	cosine	cosine	cosine	cosine
Training epochs	300	300	300	300
Warmup epochs	20	30	30	20
Batch size	128	72	36	36
Num. of encoder layers N	12	12	12	12
Num. of decoder layers N_d	4	-	-	-
Input points M	1024	1024	2048	2048
Num. of patches n	128	128	256	256
Patch size k	32	32	32	32
Augmentation	Scale&Trans	Scale&Trans	Rotation	-

Table 8. Implementation details for pre-training and downstream tasks such as classfication and segmentation tasks.

From Tab. 9, we can observe that the Chamfer Distance between the early, middle, and late stage point clouds under different random seeds can also be understood as their similarity. Multiple experiments indicate that the early and late stage points exhibit more unstable distributions compared to the middle stage points, which can be intuitively seen from the visualization in Fig. 6. Furthermore, the middle stage points show a high similarity across different random seeds, confirming that they reliably represent the skeletal structure of the entire point cloud. Additionally, we observe that the similarity between the late and middle stage points is also high, suggesting an overlap between these two stages and further supporting the redundancy characteristic of the late stage points.

Seed	Chamfer Distance			
	$D^{Early,Early}$	$D^{Mid,Mid}$	$D^{Mid,Late}$	$D^{Late,Late}$
1 vs. 2	0.18	0.09	0.11	0.24
4 vs. 7	0.17	0.07	0.09	0.20
6 vs. 25	0.16	0.06	0.08	0.22
16 vs. 32	0.19	0.06	0.10	0.25
44 vs. 42	0.18	0.06	0.10	0.22
123 vs. 321	0.19	0.06	0.07	0.23
Mean	0.18	0.07	0.09	0.23

Table 9. The similarity of the early, middle, and late stages under different random seeds. (a vs. b, $D^{A,B}$) denotes the Chamfer distance between stage A under random seed a and stage B under random seed b.

B. Ablation Study

The ablation study mainly supplements performance comparison experiments for SPM with different neurons, as well as comparisons with other models [41] using ILIF [29] and its training strategies. Additionally, we conduct an ablation study on the pre-training mask ratio.

Ablation on different neurons. In Tab. 10, we conduct a detailed ablation study using different neurons on the OBJ-BG, OBJ-ONLY, PB-T50-RS and ModelNet40 datasets. It can be observed that the accuracy of traditional IF neurons and their variants tends to fluctuate only slightly, whereas ILIF significantly improves the model's performance. This improvement is due to its integer-based training and spike-based inference mechanism. Therefore, a direct comparison with the traditional LIF model is not entirely reasonable. Here, we present the experimental results of SPM using ILIF, where accuracy reaches 91.5%, 91.2%, and 85.2% on OBJ-BG, OBJ-ONLY and PB-T50-RS datasets respectively, and 93.0% on the ModelNet40 dataset.

Method	Neurons	ScanObjectNN			ModelNet40
		OBJ-BG	OBJ-ONLY	PB-T50-RS	
E-3DSNN	ILIF [29]	86.5*	86.0*	80.4*	91.7
SPM	IF [3]	89.8	89.0	84.1	92.0
	LIF [9]	90.2	89.5	84.2	92.3
	EIF [1]	89.9	89.2	84.1	92.1
	PLIF [7]	90.5	89.6	84.1	92.5
	ILIF [29]	**91.5**	**91.2**	**85.2**	**93.0**
	Improvement	+5.0	+5.2	+4.8	+1.3

Table 10. Ablation study on different neurons with 4 time steps for IF, LIF, EIF and PLIF and 1×4 for ILIF.

Comparison with other models. In Tab. 10, we also make an additional comparison between SPM and other ILIF-based models such as E-3DSNN. To ensure a fair comparison, we used a configuration of $T \times D$ as 1×4, which follows the ILIF training strategy, allowing the model to convert into 4 time steps for spike-based inference during the inference phase. In the case of using ILIF equally, we

Masking ratio	Loss	PB-T50-RS	ModelNet40
0.4	1.66	86.1	92.6
0.6	**1.57**	**86.5**	**93.1**
0.8	2.03	85.9	82.7
0.9	2.00	86.0	92.5

Table 11. Ablation study on masking strategy. The pre-training loss (× 1000) along with fine-tuning accuracy (%) are reported on PB-T50-RS and ModelNet40.

observe that our SPM significantly outperforms benchmark models such as E-3DSNN, with overall accuracy reaching 91.5%, 91.2%, 85.2%, and 93.0% on OBJ-BG, OBJ-ONLY, PB-T50-RS, and ModelNet40, respectively.

Ablation on mask ratios In Tab. 11, we conduct a detailed ablation study on the masking ratio. It can be observed that when the masking ratio is set to 0.6, which is the ratio we ultimately selected, the accuracy of the fine-tuning task reaches their optimal performance, outperforming other settings. This suggests that for spike-based pre-training, too much masking can lead to excessive information loss, preventing the SNN encoder from learning meaningful feature representations. On the other hand, too little masking may reduce the difficulty of the reconstruction task, also causing SNN encoder to fail to learn strong feature representations.

C. Implement details

In this section, we provide more specific details about the training parameters for each dataset, as shown in Tab. 8. Different training hyperparameters were used for different datasets, but the backbone remained the same, consisting of a 12-layer stacked SPM, which facilitated the generalization across different datasets and the fine-tuning of pre-training models.

D. More Visualizations

The figure of main paper shows only a few sample visualizations. In this section, we provide more visualizations to quantitatively demonstrate the effectiveness and performance of our SPM, as shown in Fig. 7 and Fig. 8.

As can be seen from Fig. 7, the part segmentation results of SPM are almost identical to those of PointMamba across different classes. Slightly more complex classes may show a minor difference, but it does not significantly affect the overall performance. From Fig. 8, it can also be observed that our SPM performs excellently in the pre-trained reconstruction task. Despite the large masking range, it is still able to recover the general shape of the object.

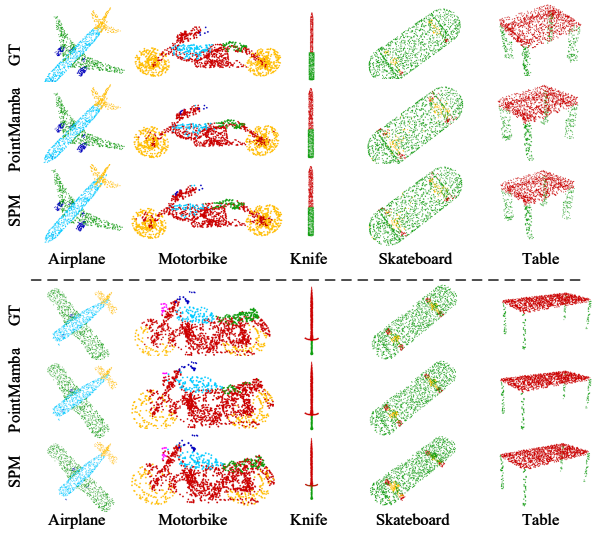

Figure 7. Qualitative results of part segmentation of our SPM and ANN counterpart (PointMamba) on ShapeNetPart.

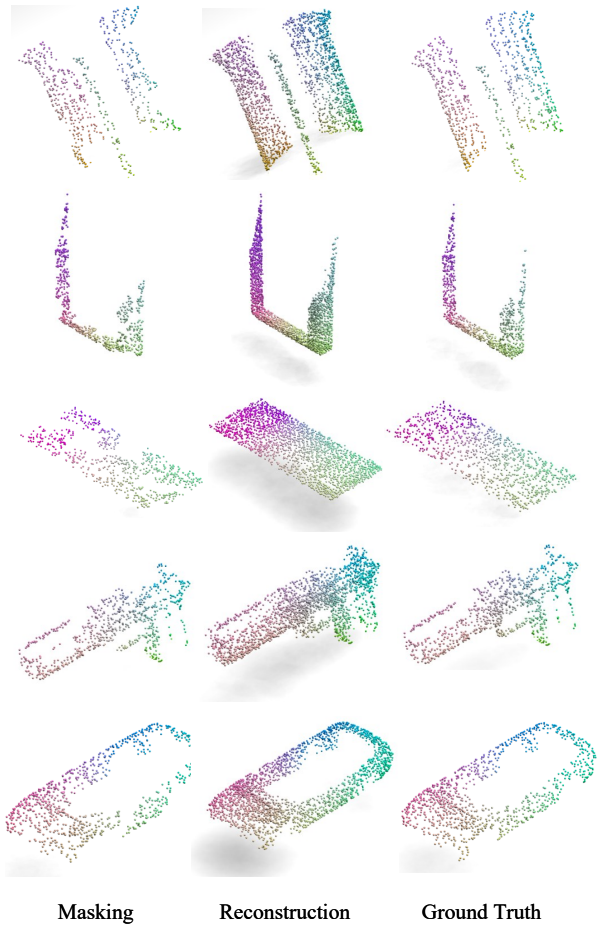

Figure 8. Qualitative results of reconstruction on ShapeNet.