This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

UST-SSM: Unified Spatio-Temporal State Space Models for Point Cloud Video Modeling

Peiming Li[1] Ziyi Wang[1] Yulin Yuan[2] Hong Liu[1] Xiangming Meng[2] Junsong Yuan[3] Mengyuan Liu[1†]
[1]State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School
[2]The Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University
[3]State University of New York at Buffalo
{lipeiming1001,ziyiwang}@stu.pku.edu.cn yulinyuan@zju.edu.cn hongliu@pku.edu.cn
xiangmingmeng@intl.zju.edu.cn jsyuan@buffalo.edu liumengyuan@pku.edu.cn

Abstract

Point cloud videos capture dynamic 3D motion while reducing the effects of lighting and viewpoint variations, making them highly effective for recognizing subtle and continuous human actions. Although Selective State Space Models (SSMs) have shown good performance in sequence modeling with linear complexity, the spatio-temporal disorder of point cloud videos hinders their unidirectional modeling when directly unfolding the point cloud video into a 1D sequence through temporally sequential scanning. To address this challenge, we propose the Unified Spatio-Temporal State Space Model (UST-SSM), which extends the latest advancements in SSMs to point cloud videos. Specifically, we introduce Spatial-Temporal Selection Scanning (STSS), which reorganizes unordered points into semantic-aware sequences through prompt-guided clustering, thereby enabling the effective utilization of points that are spatially and temporally distant yet similar within the sequence. For missing 4D geometric and motion details, Spatio-Temporal Structure Aggregation (STSA) aggregates spatio-temporal features and compensates. To improve temporal interaction within the sampled sequence, Temporal Interaction Sampling (TIS) enhances fine-grained temporal dependencies through non-anchor frame utilization and expanded receptive fields. Experimental results on the MSR-Action3D, NTU RGB+D, and Synthia 4D datasets validate the effectiveness of our method. Our code is available at https://github.com/wangzy01/UST-SSM.

1. Introduction

Recent advances in depth sensing technologies have driven growing interest in point cloud videos analysis for robotics [17, 22, 30] and autonomous systems [2, 24, 36]. Point cloud videos inherently resist lighting/viewpoint variations while preserving 3D motion dynamics, making them particularly suitable for human action recognition.

Recent advances in point cloud video modeling [5, 6, 8, 9, 26, 37, 38] leverage CNNs or Transformers. However, CNNs [26] struggle with modeling long-term dependencies, and P4Transformer [6] and PST-Transformer [9] consume significant memory for long sequences. Recently, se-

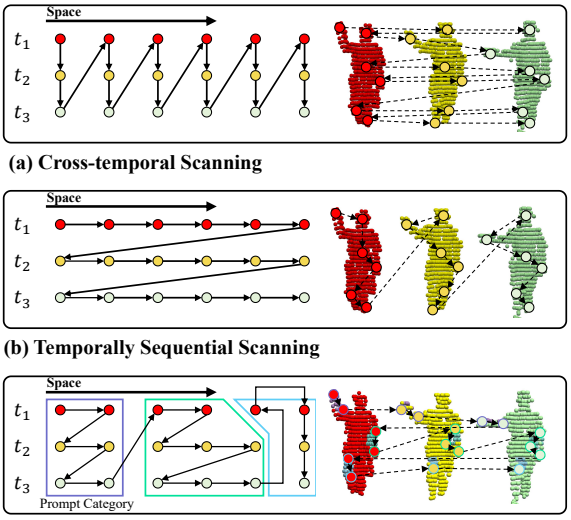
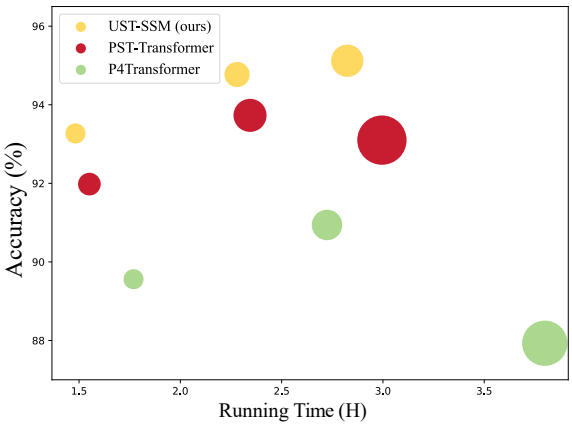
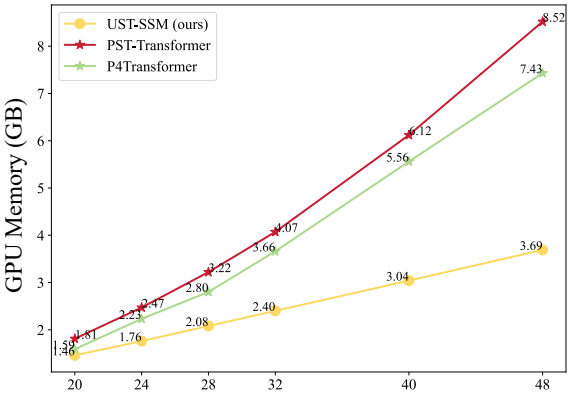


Figure 1. Illustration of Spatio-Temporal Selection Scanning and comparison with previous scanning strategies [15, 21, 25]. **(a) Cross-temporal Scanning:** Points across the frame are scanned according to coordinates. **(b) Temporally Sequential Scanning:** Points are first scanned within each frame, then along the temporal dimension. **(c) Spatio-Temporal Selection Scanning (Ours):** The scanning first occurs within semantic-aware sequences through prompt-guided clustering. Within each cluster, Hilbert sorting is performed to preserve local geometry, followed by the temporal dimension.

† Corresponding author: Mengyuan Liu (liumengyuan@pku.edu.cn)



(a) Running Time and Accuracy Comparison



(b) GPU Memory Comparison

Figure 2. Comparison of UST-SSM with P4Transformer and PST-Transformer. **(a) Comparison on running time and accuracy (the circle size represents the GPU memory) and (b) comparison on GPU memory:** Our UST-SSM, with linear complexity, achieves optimal performance in terms of runtime, memory usage, and accuracy, as detailed in Sec. 4.2.

lective state space models (SSMs) have shown competitive performance compared with Transformers in NLP, offering linear complexity and significantly reduced memory usage. SSM-based methods [15, 21] have successfully applied SSMs to 3D point cloud analysis by serializing the spatial structure, but they focus solely on the 3D spatial dimension and neglect the temporal aspect. Mamba4D [25] replaces the classical transformer by employing a predefined temporally sequential scanning, unfolding the point cloud videos into 1D sequence. However, each point in Mamba is modeled based solely on its preceding point in the scanning sequence, which causes temporally sequential scanning to restrict the interaction range of query points to information from previous frames. Furthermore, previous research [14] has shown that Mamba experiences long-range attenuation in token interactions, meaning that interactions between tokens that are spatially and temporally distant are reduced. As a result, even distant but relevant points from earlier scans cannot be effectively utilized by the query points. In contrast, our UST-SSM employs Spatial-Temporal Selection Scanning, which clusters distant yet similar spatio-temporal points, overcoming the limitations of temporal sequential scanning while mitigating the long-range attenuation of similar points. Fig.1 highlights the key differences between previous scanning strategies and STSS. As demonstrated in Fig. 2, our method outperforms Transformer-based approaches in both performance and efficiency.

However, effectively modeling the spatio-temporal characteristics of point cloud videos presents unique challenges: (1) Non-unified spatio-temporal disorder: The absence of consistent spatio-temporal ordering in point cloud videos fundamentally conflicts with the unidirectional scanning paradigm of SSMs, as existing scanning strategies fail to achieve the desired spatio-temporal order. (2) Local geometry loss: Existing SSM-based methods' serialization strategies discard crucial geometric relationships between neighboring points. (3) Temporal interaction limitation: Traditional temporal sampling strategies create fragmented temporal contexts that hinder motion analysis.

Our proposed UST-SSM unifies spatio-temporal modeling for point cloud videos and addresses these issues through three key technical contributions: (1) Spatio-Temporal Selection Scanning (STSS): Point cloud videos lack consistent temporal alignment, and spatial disorder complicates temporal modeling. When extending 3D point cloud methods to 4D, previous methods typically perform spatial sorting followed by temporally sequential scanning. However, points that differ in temporal dimensions may actually represent the same point or exhibit higher similarity. After sorting, these points are placed far apart, leading to the problem of long-range attenuation of similar points in unidirectional modeling. STSS introduces a lightweight prompt network to cluster points based on semantic similarity rather than temporal proximity. This enables coherent sequence formation by grouping spatio-temporally distant but semantically related points, while preserving local geometry through Hilbert-curve spatial ordering within clusters. (2) Spatio-Temporal Structure Aggregation (STSA): The unidirectional sorting in SSMs relies solely on preceding points in the scan sequence, neglecting local geometric details. STSA actively recovers geometric details through spatio-temporal KNN neighborhood feature propagation, compensating for information loss during serialization. (3) Temporal Interaction Sampling (TIS): Traditional temporal sampling strategies usually ignore the fine-grained time dependence between frames, which limits the time interaction between frames. TIS innovatively leverages non-anchor frames through multi-stride sampling, expands the temporal receptive field, and improves temporal interaction within the sampled sequence. This spatio-temporal structure extends SSMs' sequential modeling ca-

pabilities to point cloud videos analysis, excelling in tasks that require long-term structure capture.

We evaluate our UST-SSM on the MSR-Action3D [20], NTU RGB+D [31], and Synthia 4D [4] datasets. Experimental results demonstrate the effectiveness of our approach. The contributions of this paper are threefold:

- We propose UST-SSM, an SSM-based model designed to unify spatio-temporal modeling for point cloud videos, in order to achieve a spatio-temporal ordering suitable for SSMs. Spatio-Temporal Selection Scanning (STSS) organizes unordered point cloud videos by clustering and enhancing interactions among spatio-temporally distant yet spatially similar points.
- We introduce Spatio-Temporal Structure Aggregation (STSA), which compensates for the spatial features overlooked during sorting by incorporating local geometric details. Temporal Interaction Sampling (TIS) extends the time receptive field and improves temporal information exchange and fine-grained feature capture.
- Extensive experiments demonstrate that the proposed UST-SSM extends SSMs' sequential modeling capabilities to point cloud videos analysis, enhancing computational efficiency and reducing resource consumption, particularly in long-term spatio-temporal modeling.

2. Related Work

Point Cloud Video Modeling. Early approaches convert point cloud videos into 3D voxel sequences [4, 29], though voxelization often leads to a loss of geometric detail. Most existing methods rely on convolutional networks or vision transformers for direct point cloud video modeling. MeteorNet [26] was the first to extend PointNet++ to the temporal dimension, allowing it to handle point cloud videos. To increase the receptive field, P4Transformer [6] leverages transformers to avoid explicit point tracking but does not encode the full spatio-temporal structure. PST-Transformer [9] addresses this by encoding features based on the spatio-temporal displacement between reference points and all points in the video. PPTr [38] introduces geometric planes as mid-level representations to improve efficiency, enhancing performance for dynamic scenes and longer video sequences. However, it does not mitigate the quadratic complexity of transformers. In this work, we propose a novel point cloud video backbone based on SSMs, addressing challenges in computational efficiency and long-term spatio-temporal modeling.

State Space Models. State Space Models (SSMs) [12, 13] have shown great promise in sequence modeling tasks. Early models like HiPPO [11] and LSSL [13] established foundational techniques for SSMs through projection and optimization methods. Structured State Space Sequence Models (S4) [12] broadened the applicability of SSMs in fields like speech and language, allowing SSMs to compete with transformers in long-sequence tasks. Advances such as S4D and S5 have introduced diagonal matrices and MIMO structures, making SSMs more efficient for parallel processing. Mamba [10] significantly enhances SSM efficiency in long-sequence modeling through selective state space mechanisms and hardware-aware optimizations. This advancement has inspired various applications in the visual domain, including image processing [28, 42], video understanding [3, 18], generative models [16], and 3D point clouds [15, 21]. However, research on applying SSMs to point cloud videos remains limited. Mamba4D [25] directly applies Mamba on top of P4Transformer. However, each point is modeled based solely on its preceding point in the scanning sequence, which restricts the interaction range of query points to information from the previous frame due to the temporally sequential scanning. In this work, we further investigate the potential of SSMs for point cloud video analysis and introduce Spatial-Temporal Selection Scanning, a method that clusters distant yet spatially similar points across time, effectively overcoming the constraints of sequential temporal scanning.

3. Method

3.1. Overview

Traditional SSMs face three fundamental challenges in point cloud video modeling: (1) *Non-unified spatio-temporal disorder* from unaligned dynamic points disrupts unidirectional scanning. (2) *Local geometry loss* caused by serialization compromises fine-grained details. (3) *Fragmented temporal context* in conventional sampling limits its motion understanding. Our UST-SSM addresses these through synergistic components:

- **Temporal Interaction Sampling (Sec. 3.2)**: Solves Challenge (3) by constructing continuous temporal contexts through multi-stride sampling.
- **Spatio-Temporal Selection Scanning (Sec. 3.3)**: Addresses Challenge (1) via prompt-guided clustering.
- **Spatio-Temporal Structure Aggregation (Sec. 3.4)**: Mitigates Challenge (2) through 4D KNN and spatio-temporal feature propagation.

The overall process of UST-SSM is shown in Fig. 3.

3.2. Temporal Interaction Sampling

Due to the unordered variations of high-frequency and low-frequency information in the temporal dimension of video sequences, traditional single-stride sampling strategies with small receptive fields tend to overlook fine-grained temporal dependencies between frames and fail to fully capture temporal context. This limitation results in insufficient interaction between frames. Temporal Interaction Sampling (TIS) addresses these issues through non-anchor frame utilization and expanded receptive fields, enabling better tem-

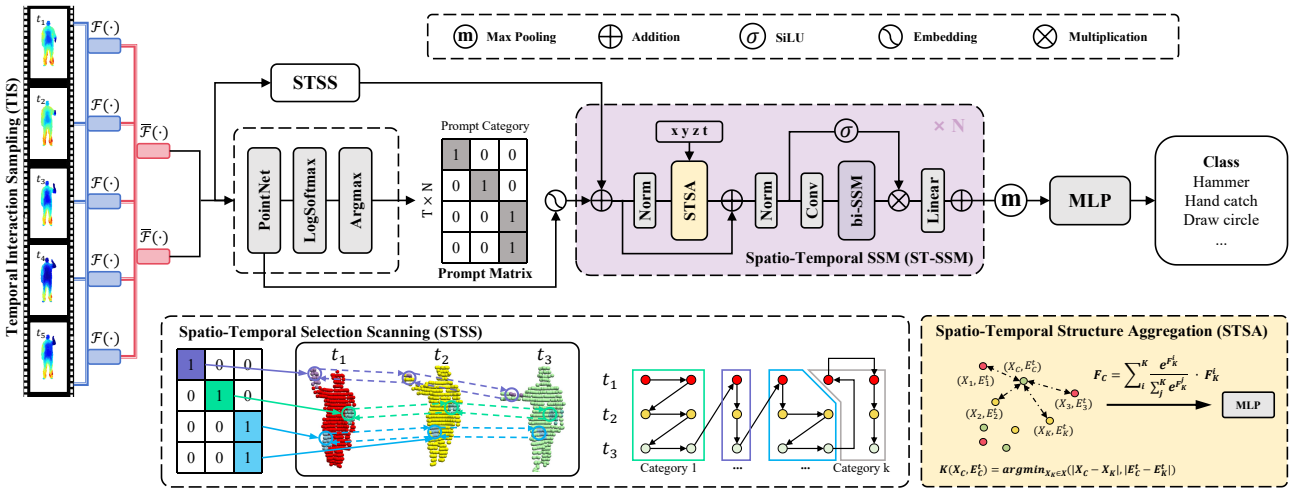


Figure 3. **Overview of UST-SSM:** First, Temporal Interaction Sampling is applied to the point cloud sequence. The downsampled sequence is then serialized using Spatio-Temporal Selection Scanning, followed by processing with ST-SSM and Spatio-Temporal Structure Aggregation. Finally, a Multi-Layer Perceptron is used for classification and prediction.

poral interaction across the sampled sequence.

Given a point cloud sequence $\mathbf{X} = \{T_1, T_2, \ldots, T_n\}$, single-stride sampling (*e.g.* stride=2) in previous methods creates temporal islands where anchor frames $\{T_2, T_4, ...\}$ only interact with immediate neighbors. This restrictive temporal window fails to exploit the full range of temporal dependencies between frames, formulated as:

$$\mathbf{Feat}_{T_{2i}} = \mathcal{F}(\mathcal{S}(T_{2i-1}, T_{2i}, T_{2i+1})), \tag{1}$$

where $i \in \{1, 2, \ldots, n/2\}$, $\mathcal{S}$ and $\mathcal{F}$ represents the spatial sampling and feature extraction operations. To enhance the utilization of non-anchor frames and expand the temporal receptive field, TIS first performs temporal sampling with a stride of 1, which aggregates information from consecutive frames. This can be formulated mathematically as:

$$\mathbf{Feat}'_{T_i} = \mathcal{F}(\mathcal{S}(T_{i-1}, T_i, T_{i+1})), \tag{2}$$

where $i \in \{1, 2, \ldots, n\}$. This is followed by sampling with a stride of 2 and feature update, allowing for the aggregation of information from a broader temporal context:

$$\begin{aligned} \mathbf{Feat}_{T_{2i}} =& \overline{\mathcal{F}}(\overline{\mathcal{S}}(\mathbf{Feat}'_{T_{2i-1}}, \mathbf{Feat}_{T_{2i}}, \mathbf{Feat}'_{T_{2i+1}})), \\ =& \overline{\mathcal{F}}(\overline{\mathcal{S}}(\mathcal{F}(\mathcal{S}(T_{2i-2}, T_{2i-1}, T_{2i})), \\ & \mathcal{F}(\mathcal{S}(T_{2i-1}, T_{2i}, T_{2i+1})), \\ & \mathcal{F}(\mathcal{S}(T_{2i}, T_{2i+1}, T_{2i+2})))), \end{aligned} \tag{3}$$

where $i \in \{1, 2, \ldots, n/2\}$. In this process, non-anchor frames are utilized multiple times, ensuring that the receptive field of each frame expands progressively. For instance, each anchor frame now accesses features from $\{T_{2n-1}, T_{2n+1}\}$ rather than $\{T_{2n}\}$, and each subsequent frame T_{2n} can leverage information from all preceding frames $\{T_1, T_2, ..., T_{2n-1}\}$. The richer temporal interactions and more comprehensive information flow improve the model's ability to capture sequential features and long-term dependencies, thus enhancing performance in tasks requiring extensive temporal context.

3.3. Spatio-Temporal Selection Scanning

Imposing a degree of order on the point cloud is essential to achieve a structured representation. Traditional methods typically employ a two-step approach: spatial sorting followed by temporal sorting. However, some points that are spatio-temporally distant may actually be the same point and exhibit higher similarity, but temporally sequential scanning separates semantically similar points across frames. As shown in Fig. 1, points from t_1 and t_3 are farther apart than random points after temporally sequential scanning. This leads to the issue of long-range attenuation of similar points in the unidirectional modeling of SSM.

In contrast to the conventional scanning, we point out STSS, which utilizes a prompt network to assign similarity-based categories to points, clustering those that are spatio-temporally distant but with high semantic similarity. A lightweight prompt network, such as PointNet, is used to generate a prompt matrix $\mathcal{P}$ with k prompt categories $(p_1, p_2, p_3, ..., p_k)$, clustering points by semantic similarity:

$$\begin{aligned} \mathcal{P} &= \text{PromptNetwork}(\mathbf{Feat}) \in \mathbb{R}^{N \times K}, \\ \mathcal{P} &= \{p_1, p_2, p_3, ..., p_k\}. \end{aligned} \tag{4}$$

STSS then clusters points that share the same prompt category and sorts them separately within each semantic-aware cluster. After clustering, points that are close in position exhibit higher correlation show higher correlation,

whereas using standard spatial filling curves that traverse the 3D space along the x, y, and z axes might result in adjacent points in the sequence having large spatial distances, leading to significant feature differences. Therefore, we use Intra-cluster Hilbert Sorting [39]. Within each cluster C_k, points are sorted using the 3D Hilbert curve:

$$\begin{aligned} \mathbf{X}_j &= \{x_i | \arg\max(p_i) = j\}, \\ \overline{\mathbf{X}}_j &= \text{HilbertSort}(\mathbf{X}_j). \end{aligned} \tag{5}$$

To maintain motion continuity, we organize points across frames in sequential order. As illustrated in the Fig. 1 and Fig. 3, we use Inter-cluster Chronological Ordering for temporal modeling within each cluster, which preserves the natural motion sequence (indicated by longitudinal or oblique arrows). Points within each frame are marked by the same color (*e.g.*, red, yellow, green). Given an input point cloud $\mathbf{X} \in \mathbb{R}^{\mathrm{T}\times\mathrm{N}\times 3}$ with features $\mathbf{Feat} \in \mathbb{R}^{\mathrm{T}\times\mathrm{N}\times\mathrm{C}}$, after processing with STSS, we obtain $\overline{\mathbf{X}} \in \mathbb{R}^{\mathrm{L}\times 3}$ and $\overline{\mathbf{Feat}} \in \mathbb{R}^{\mathrm{L}\times\mathrm{C}}$, where $\mathrm{L} = \mathrm{T} \times \mathrm{N}$, the overall process of STSS is formulated as follows::

$$\begin{aligned} \mathcal{P} &= \text{PromptNetwork}(\mathbf{Feat}) \in \mathbb{R}^{N\times K}, \\ \mathbf{X}_j, \mathbf{Feat}_j &= \{x_i, feat_i | \arg\max(p_i) = j\}, \\ \widehat{\mathbf{X}_j}, \widehat{\mathbf{Feat}_j} &= \text{HilbertSort}(\mathbf{X}_j, \mathbf{Feat}_j), \\ \overline{\mathbf{X}_j}, \overline{\mathbf{Feat}_j} &= \text{ChronologicalOrdering}(\widehat{\mathbf{X}_j}, \widehat{\mathbf{Feat}_j}), \\ \overline{\mathbf{X}} = \bigcup_{j=1}^{k} \overline{\mathbf{X}_j}, \overline{\mathbf{Feat}} &= \bigcup_{j=1}^{k} \overline{\mathbf{Feat}_j}. \end{aligned} \tag{6}$$

3.4. Spatio-Temporal Structure Aggregation

While STSS unifies spatio-temporal disorder, the unidirectional scanning inherent to SSMs inevitably disrupts local geometric relationships critical for fine-grained motion analysis. Conventional SSM-based methods [25] serialize points without preserving neighborhood contexts, sacrificing spatial coherence for computational efficiency. To overcome this limitation, we introduce Spatio-Temporal Structure Aggregation (STSA), a lightweight yet effective module that explicitly recovers geometric details while capturing dynamic motion patterns.

As illustrated in Fig. 3, STSA operates within the ST-SSM block alongside bi-SSM, forming a symbiotic relationship: STSA concentrates on localized spatio-temporal interactions while SSMs handle global sequential dependencies. This division of labor ensures efficient modeling of both neighborhood structures and long-range dynamics. The STSA process comprises two stages:

Neighborhood Construction via 4D KNN: Traditional spatial KNN fails to account for temporal continuity, treating each frame as isolated. Unlike previous Spatio-Temporal Structure Encoding, STSA places all points across frames into a unified coordinate space ($L = T \times N$) and extends KNN to 4D by incorporating temporal embeddings $\mathbf{E}^t$ into the distance metric for neighbor selection. This not only enhances filtering efficiency but also further enables unified spatio-temporal modeling:

$$K = \underset{\mathbf{X}_K \in X}{\text{argmin}} \left(|\mathbf{X}_C - \mathbf{X}_K| + |\mathbf{E}^t_C - \mathbf{E}^t_K| \right). \tag{7}$$

This constructs neighborhoods that preserve both geometric cohesion (through spatial coordinates $\mathbf{X}$) and motion trajectories (via learned temporal embeddings $\mathbf{E}^t$).

Spatio-temporal Feature Propagation: For each center point C, we normalize neighbor features $\mathbf{F}_K$ relative to the center while retaining absolute positional context:

$$\mathbf{F}'_K = \frac{\mathbf{F}_K - \mathbf{F}_C}{|\mathbf{F}_K - \mathbf{F}_C|_2 + \epsilon} \oplus \mathbf{F}_C \in \mathbb{R}^{\mathrm{L}\times k\times 2C}, \tag{8}$$

where $\oplus$ denotes concatenation. This dual representation captures both relative geometric variations and absolute feature characteristics, enabling nuanced local interactions. To dynamically prioritize informative neighbors, we replace conventional spatio-temporal convolution with attention-inspired exponential pooling:

$$\mathbf{F}'_C = \text{MLP}\left(\sum_{i=1}^{K} \frac{e^{\mathbf{F}'^i_K}}{\sum_j e^{\mathbf{F}'^j_K}} \cdot \mathbf{F}'^i_K \right) \in \mathbb{R}^{\mathrm{L}\times k\times C}. \tag{9}$$

This differentiable operation amplifies salient features while adaptively focusing on kinetically significant points without additional parameters.

STSA's design directly addresses the geometric information loss caused by serialization in three key ways: (1) The 4D KNN explicitly reconstructs spatio-temporal neighborhoods disrupted during scanning. (2) Relative feature normalization preserves displacement-sensitive motion cues. (3) Adaptive pooling maintains efficiency. By interlacing STSA with SSM layers, our model alternately focuses on micro-interactions and macro-dynamics, achieving comprehensive 4D understanding with linear complexity.

4. Experiment

4.1. Experimental Details

Datasets: The MSR-Action3D dataset [20] consists of 567 depth videos captured with Kinect v1, spanning 20 action categories and containing a total of 23K frames. We follow the protocol in P4Transformer [6] for training/testing split. The NTU RGB+D dataset [31] is a large-scale benchmark for action recognition, comprising 56,880 videos of 60 actions performed by 40 subjects across 80 camera views. For NTU RGB+D dataset, we adopt the cross-subject and cross-view evaluation protocols. Synthia 4D [4] leverages

Table 1. Action recognition accuracy on MSR-Action3D dataset.

Backbone	Method	Venue	Frames	Acc (%)
CNN	MeteorNet [26]	ICCV'19	16	88.21
			24	88.50
	PSTNet [6]	ICLR'21	16	89.90
			24	91.20
			36	88.21
	Kinet [41]	CVPR'22	16	91.92
			24	93.27
	PointCMP [32]	ICCV'23	16	92.26
			24	93.27
	3DinAction [1]	CVPR'24	16	90.57
			24	92.23
Transformer	P4Transformer [6]	CVPR'21	16	89.56
			24	90.94
			36	82.81
	PPTr [38]	ECCV'22	16	90.31
			24	92.33
	PST-Transformer [9]	TPAMI'23	16	91.98
			24	93.73
			36	91.15
	PointCPSC [33]	ICCV'23	16	92.26
			24	92.68
	LeaF [27]	ICCV'23	16	91.50
			24	93.84
SSM	MAMBA4D [25]	CVPR'25	24	92.68
			36	93.23
	UST-SSM (Ours)	-	16	**93.27**
			24	**94.77**
			36	**95.12**

Table 2. Action recognition accuracy of various methods on NTU RGB+D dataset. CS and CV represent the cross-subject and cross-view evaluation protocols.

Method	Venue	CS	CV
AS-GCN [19]	CVPR'19	86.8	94.2
AGC-LSTM [35]	CVPR'19	89.2	95.0
VA-fusion [40]	TPAMI'19	89.4	95.0
DGNN [34]	CVPR'19	89.9	96.1
GCA-LSTM [23]	TPAMI'20	74.4	82.8
P4Transformer [6]	CVPR'21	90.2	96.4
PSTNet [7]	ICLR'21	90.5	96.5
PointCPSC [33]	ICCV'23	88.0	-
PointCMP [32]	CVPR'23	88.5	-
PST-Transformer [9]	TPAMI'23	91.0	96.4
3DinAction [1]	CVPR'24	89.3	-
UST-SSM (Ours)	-	**92.1**	**97.8**

Table 3. Efficiency and accuracy comparison on MSR-Action3D dataset with 36 frames.

Method	Params (M)	Memory (GB)	Times (H)	Accuracy (%)
PSTNet	8.259	22.031	8.526	88.21
P4Transformer	0.834	3.667	3.799	82.81
PST-Transformer	0.289	4.072	2.995	91.15
UST-SSM	0.138	2.403	2.824	95.12

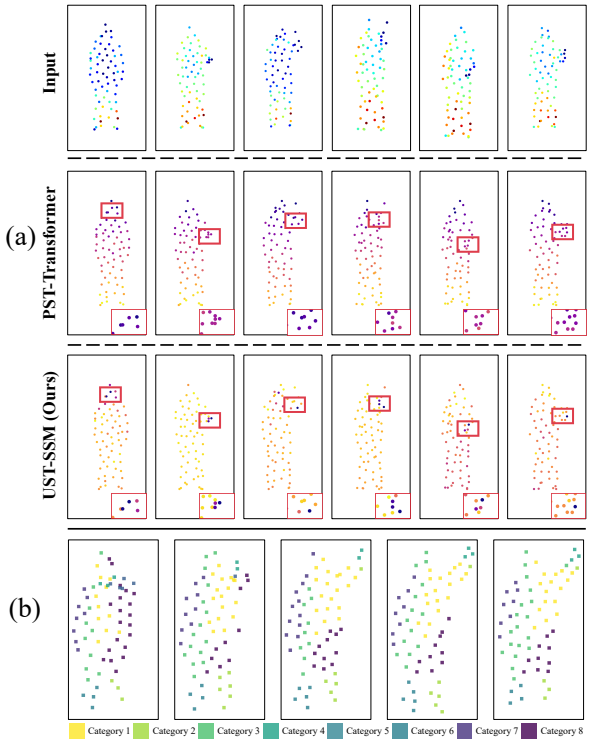


Figure 4. **(a) Feature visualization:** For the inputs, colors represent depth, while darker colors indicate higher feature weights. Compared to attention mechanisms, our method produces higher activations in key moving regions, demonstrating its ability to capture more informative cues for action reasoning. **(b) Prompt matrix visualization:** Semantically similar regions across distant frames are clustered into the same prompt category.

the Synthia dataset to generate 3D video sequences, covering six driving scenarios. Each scenario includes four stereo RGB-D images captured from the top of a moving vehicle. We use the same training/testing splits as [6, 7, 9].

Implementation Details: For the MSR-Action3D and Synthia 4D datasets, we set the number of ST-SSM to 1 and train the model for 50 epochs on an NVIDIA RTX 4090 GPU using the SGD optimizer with the initial learning rate of 0.01, decayed by a factor of 0.1 at the 20th and 30th epochs. For the NTU RGB+D dataset, we increase the number of ST-SSM to 3 and train the model for 15 epochs. The Prompt Network is composed of a T-Net for geometric transformation and a PointNet architecture for feature extraction, with both parts initialized using Xavier uniform distribution. For 4D semantic segmentation, we establish our segmentation framework based on recent representative transformer backbone P4Transformer. Specifically, we replace the transformer with ST-SSM and use STSS for both sorting and reverse sorting.

4.2. 3D Action Recognition

Quantitative Comparison on MSR-Action3D dataset. As shown in Tab. 1, we evaluate our model on the MSR-

Table 4. 4D semantic segmentation results on Synthia 4D dataset.

Method	Venue	Input	Frame	mIoU (%)
3D MinkNet14 [4]	CVPR'19	voxel	1	76.24
4D MinkNet14 [4]	CVPR'19	voxel	3	77.46
MeteorNet-m [26]	ICCV'19	point	2	81.47
MeteorNet-l [26]	ICCV'19	point	3	81.80
PSTNet [7]	ICLR'21	point	3	82.24
P4Transformer [6]	CVPR'21	point	1	82.41
P4Transformer [6]	CVPR'21	point	3	83.16
PST-Transformer [9]	TPAMI'22	point	1	82.92
PST-Transformer [9]	TPAMI'22	point	3	83.95
MAMBA4D [25]	CVPR'25	point	3	83.35
P4Transformer + Ours	-	point	3	**84.06**

Table 5. Impact of different scanning strategies and space-filling curves on MSR-Action3D dataset.

Scanning	Space-Filling	Accuracy
Cross-temporal	XYZ	85.02
	XZY	83.97
	YXZ	88.15
	YZX	87.46
	ZXY	85.02
	ZYX	86.41
	Hilbert Curve	89.20
Temporally Sequential	XYZ	93.03
	XZY	92.68
	YXZ	93.38
	YZX	91.99
	ZXY	92.68
	ZYX	93.38
	Hilbert Curve	94.08
Spatio-Temporal Selection	Hilbert Curve	**94.77**

Table 6. The robustness of key components on recent representative Transformer architecture on MSR-Action3D dataset.

Architecture	ST-SSM		STSS	TIS	Acc (%)
	Selective SSM	STSA			
P4Transformer + Ours	✗	✗	✗	✗	90.94
	✓	✗	✗	✗	92.33
	✓	✓	✗	✗	93.03
	✓	✓	✓	✗	94.43
	✓	✓	✓	✓	94.77

Table 7. Impact of different spatio-temporal sampling strategies on MSR-Action3D dataset.

Strategy	Stride	Params (M)	Memory (GB)	Accuracy (%)
Single-stride	1	0.046	3.079	94.77
	2	0.046	2.119	94.43
TIS	1+2	0.048	1.759	**94.77**

Table 8. Impact of STSS and STSA on NTU RGB+D dataset.

Spatio-Temporal Selection Sorting	STSA	Accuracy (%)
✗	✗	90.8
✗	✓	91.4
✓	✗	91.5
✓	✓	**92.1**

Action3D dataset and compare it against CNN-based methods [7, 26], Transformer-based methods [6, 9, 38] and SSM-based method [25]. Our approach achieves higher recognition accuracy than state-of-the-art methods across commonly used 16-frame, 24-frame, and longer 36-frame inputs. Notably, Transformer-based methods experience a substantial performance drop as sequence length increases, underscoring the challenges in modeling long-sequence 4D videos. In contrast, our method demonstrates a steady improvement in recognition accuracy with longer sequences.

Quantitative Comparison on NTU RGB+D dataset. As shown in Tab. 2, UST-SSM achieves higher recognition accuracy than state-of-the-art methods across all evaluation settings. Notably, in the cross-subject evaluation on the NTU RGB+D dataset, UST-SSM surpasses the leading Transformer-based method, PST-Transformer, by 1.1%.

Qualitative Results. As shown in Fig. 4, we compare the features obtained in ST-SSM with those derived from an attention mechanism. As expected, the Spatio-Temporal Structure Aggregation effectively focuses on the significant moving regions, supporting our intuition that ST-SSM can effectively capture the spatio-temporal structure of point cloud videos, serving as a viable alternative to attention mechanisms. As shown in the prompt matrix visualization in Fig. 4, similar regions across different frames are clustered into the same prompt category, validating the effectiveness of STSS in clustering similar points.

Efficiency Comparison. To validate the advantages of UST-SSM on long sequences, as shown in Tab. 3, with the setup of 36 frames, 2048 points, and a batch size of 32, UST-SSM significantly outperforms CNN-based and Transformer-based methods in terms of parameters, memory usage and training times on a single RTX 4090, while maintaining higher accuracy. As shown in Fig. 2, as sequence length grows, GPU memory usage for P4Transformer and PST-Transformer follows a quadratic growth, whereas our method scales linearly.

4.3. 4D Semantic Segmentation

Comparison on Synthia 4D dataset. To further demonstrate the versatility of UST-SSM for point cloud video modeling, we evaluate our method on 4D semantic segmentation. As shown in Tab. 4, our approach outperforms all CNN-based methods [26] and Transformer-based methods [6, 9, 38]. We present four segmentation results from the Synthia 4D dataset in Fig. 5. Our method achieves precise segmentation even in subtle regions, outperforming state-of-the-art Transformer-based methods.

4.4. Ablation Study

The STSS, STSA and TIS are critical components of our method. To highlight the importance of our customized strategies for extending point cloud videos in SSMs, we report the effects of various strategies used in UST-SSM in Tab. 6. The results show that the combination of STSS, STSA, and TIS contributes to improved action recognition accuracy, validating the effectiveness of unified spatio-

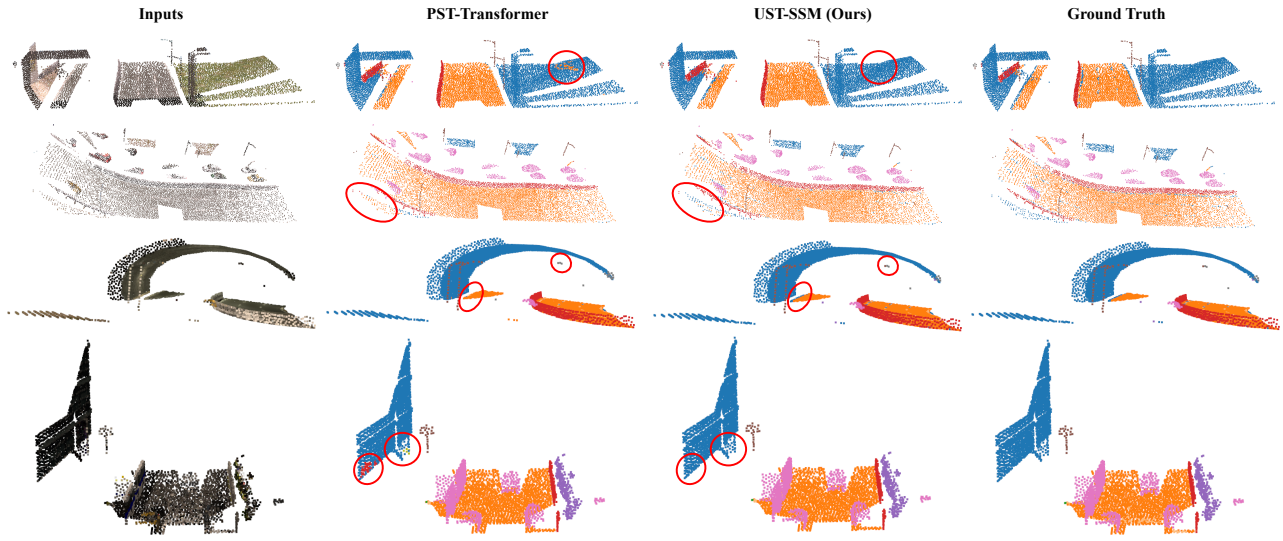


Figure 5. 4D semantic segmentation visualization on the Synthia 4D dataset. Our method achieves precise segmentation even in subtle regions, outperforming state-of-the-art Transformer-based methods.

temporal sorting and the impact of spatio-temporal feature interactions. Notably, the full UST-SSM yields a 3.83% improvement in recognition accuracy, underscoring the importance of extending SSMs to point cloud videos.

Evaluation of STSS: As shown in Tab. 6, applying STSS results in a 1.4% improvement in recognition accuracy. To further evaluate our STSS, we compare the performance of different scanning strategies, as presented in Tab. 5. Among these, the combination of the Hilbert space-filling curve and Spatio-Temporal Selection, adopted by UST-SSM, consistently outperforms all other configurations, demonstrating its superiority in unified spatio-temporal modeling.

Evaluation of STSA and ST-SSM: As shown in Tab. 6, adding STSA to the ST-SSM module results in a 0.7% improvement in recognition accuracy. Additionally, using the full ST-SSM module achieves a 2.09% improvement over P4Transformer. These results validate the effectiveness of STSA and its synergy with SSM.

Evaluation of TIS: We also conduct an ablation study on different spatio-temporal sampling strategies in Tab. 7, evaluating the memory consumption introduced by TIS. Single-stride refers to the original sampling strategy in previous methods. Our TIS compensates for the temporal field of view of each frame, achieving superior results with fewer sampled frames while reducing GPU memory usage.

Evaluation of the robustness of our designs: Given that one of the key advantages of the selective state space models over Transformer-based models is its linear complexity, demonstrating its performance on large-scale datasets is essential to fully validate the scalability and practical benefits of the proposed method. Specifically, we evaluate the design on the NTU RGB+D dataset. The results shown in Tab. 8 highlight the effectiveness of our designs.

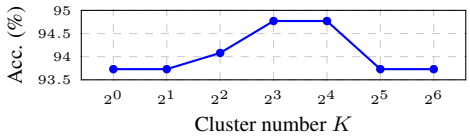


Figure 6. Ablation study on semantic similarity.

Evaluation of Semantic Similarity: An ablation study on the weighting for semantic similarity is conducted in the Fig. 6. When $K = 2^0$, excessively weighting spatial distance and ignoring semantic similarity reduces it to Temporally Sequential Scanning. When K equals the number of points 2^6, overemphasizing semantic similarity leads to semantic-only grouping, causing decreased performance. The best results are achieved when K is set to 2^3 or 2^4.

5. Conclusion

In this paper, we present UST-SSM, an SSM-based model for unified spatio-temporal encoding. UST-SSM clusters spatio-temporally distant but similar points through Spatio-Temporal Selection Scanning, enabling unified spatio-temporal sorting and long-sequence memory retention. Through Spatio-Temporal Structure Aggregation and Temporal Interaction Sampling, UST-SSM compensates for missing geometric and motion details while enhancing temporal interaction between frames. Extensive experiments validate the effectiveness of UST-SSM in point cloud video modeling. Our method transforms unordered point cloud videos into ordered sequences along the spatio-temporal dimension, effectively addressing the reliance of SSM models on structured point clouds and successfully extending SSMs to point cloud video modeling.

Acknowledgements This work was supported by National Natural Science Foundation of China (No. 62473007), Natural Science Foundation of Guangdong Province (No. 2024A1515012089), Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (No. RCYX20231211090248064).

References

[1] Yizhak Ben-Shabat, Oren Shrout, and Stephen Gould. 3dinaction: Understanding human actions in 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

[2] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1

[3] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding, 2024. 3

[4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 5, 7

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[6] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 3, 5, 6, 7

[7] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and MohanS. Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. *Cornell University - arXiv,Cornell University - arXiv*, 2022. 6, 7

[8] Hehe Fan, Xin Yu, Yi Yang, and Mohan Kankanhalli. Deep hierarchical representation of point cloud videos via spatio-temporal decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[9] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point spatio-temporal transformer networks for point cloud video modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 3, 6, 7

[10] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. 2023. 3

[11] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems*, 2020. 3

[12] Albert Gu, Karan Goel, and Christopher R'e. Efficiently modeling long sequences with structured state spaces. *ArXiv*, 2021. 3

[13] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. In *Advances in Neural Information Processing Systems*, 2021. 3

[14] Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. *ArXiv*, 2024. 2

[15] Xu Han, Yuan Tang, Zhaoxuan Wang, and Xianzhi Li. Mamba3d: Enhancing local features for 3d point cloud analysis via state space model. 2024. 1, 2, 3

[16] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Björn Ommer. Zigma: A dit-style zigzag mamba diffusion model, 2024. 3

[17] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1

[18] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding, 2024. 3

[19] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[20] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010. 3, 5

[21] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. 2024. 1, 2, 3

[22] Junchi Liang and Abdeslam Boularias. Learning category-level manipulation tasks from point clouds with dynamic graph cnns. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1

[23] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[24] Jiuming Liu, Guangming Wang, Chaokang Jiang, Zhe Liu, and Hesheng Wang. Translo: A window-based masked point transformer framework for large-scale lidar odometry. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1

[25] Jiuming Liu, Jinru Han, Lihao Liu, Angelica I. Aviles-Rivero, Chaokang Jiang, Zhe Liu, and Hesheng Wang. Mamba4d: Efficient long-sequence point cloud video understanding with disentangled spatial-temporal state space models, 2024. 1, 2, 3, 5, 6, 7

[26] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 6, 7

[27] Yunze Liu, Junyu Chen, Zekai Zhang, Jingwei Huang, and Li Yi. Leaf: Learning frames for 4d point cloud sequence understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 6

[28] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model, 2024. 3

[29] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 3

[30] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 1

[31] Amir Shahroudy, Jun Li, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. 2016. 3, 5

[32] Zhiqiang Shen, Xiaoxiao Sheng, Longguang Wang, Yulan Guo, Qiong Liu, and Xi Zhou. Pointcmp: Contrastive mask prediction for self-supervised learning on point cloud videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[33] Xiaoxiao Sheng, Zhiqiang Shen, Gang Xiao, Longguang Wang, Yu Kuen Guo, and Hehe Fan. Point contrastive prediction with semantic clustering for self-supervised learning on point cloud videos. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6

[34] Linjun Shi, Yifan Zhang, Jing Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. *IEEE Conference Proceedings,IEEE Conference Proceedings*, 2019. 6

[35] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[36] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Neural Information Processing Systems,Neural Information Processing Systems*, 2021. 1

[37] Yimin Wei, Hao Liu, Tingting Xie, Qiuhong Ke, and Yulan Guo. Spatial-temporal transformer for 3d point cloud sequences. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 1

[38] Hao Wen, Yunze Liu, Jingwei Huang, Bo Duan, and Li Yi. Point primitive transformer for long-term 4d point cloud video understanding. 2022. 1, 3, 6, 7

[39] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[40] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 6

[41] Jianqi Zhong, Kaichen Zhou, Qingyong Hu, Bing Wang, Niki Trigoni, and A. Markham. No pain, big gain: Classify dynamic point cloud sequences with static models by fitting feature-level space-time surfaces. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[42] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model, 2024. 3