STPM: Spatial-Temporal Point Mamba for Activity Recognition Using mmWave Radar Point Clouds

Yingru Chen[†], Zhihao Guo[†], Haimin Zhang, Min Xu[*]
University of Technology Sydney, Sydney, NSW, Australia
{Yingru.Chen-1, Zhihao.Guo-1}@student.uts.edu.au, {Haimin.Zhang, Min.Xu}@uts.edu.au

***Abstract*—Human activity recognition using millimeter-wave radar point clouds has emerged as a promising visual privacy-preserving sensing paradigm, transitioning from multi-domain Doppler analysis to point cloud-based methods for richer spatial information. However, existing approaches face critical challenges in modeling temporal dependencies between consecutive frames and simultaneously capturing both local geometric structures and global spatial relationships. To address these challenges, we propose STPM (Spatial-Temporal Point Mamba), a novel framework that extends traditional State Space Models through three key innovations: (1) a bidirectional selective mechanism that captures comprehensive temporal dependencies while maintaining linear memory complexity, (2) a queue-based temporal processing strategy with theoretical guarantees for preventing error accumulation, and (3) a hierarchical grouping strategy that effectively models both local geometric details and global spatial contexts. Through extensive evaluations on RadHAR and MM-Fi datasets, STPM achieves state-of-the-art performance with 98.14% and 95.69% accuracy respectively, while reducing memory consumption by 35% compared to transformer-based alternatives. Extensive experiments demonstrate its effectiveness in distinguishing semantically similar but functionally distinct motions.**

***Index Terms*—Point Cloud, Human Activity Recognition, Millimeter-Wave (mmWave)**

I. Introduction

Human activity recognition using millimeter-wave (mmWave) radar has emerged as a promising visual privacy-aware sensing solution. The mmWave provides rich spatiotemporal information while enabling all-weather operation [1]. Recently, many researchers have provided advanced solutions, which have shifted from multi-domain Doppler analysis to point cloud-based methods for more comprehensive three-dimensional spatial information in complex motion analysis [2]–[5].

The evolution of point cloud processing has advanced through deep learning architectures and multi-modal fusion [6]–[8]. While Transformer-based architectures have shown promise in capturing global relationships in point-cloud sequences [9]–[11], they face significant computational challenges due to quadratic memory complexity when processing long sequences of radar point clouds. Recent progress in sequential modeling through State Space Models (SSMs) [12], [13] offers potential alternatives. They provide an important solution to save resources used. However, the evolution of point cloud processing for mmWave radar data faces two critical challenges in activity recognition. First, existing methods exhibit limited capability in modeling temporal relationships between consecutive radar frames, which significantly impacts their ability to distinguish subtle motion patterns that differ primarily in their temporal execution. Second, current approaches lack effective mechanisms to simultaneously capture both local geometric details and global spatial contexts, which is crucial for recognizing semantically similar actions with different functional purposes. These limitations particularly affect rehabilitation monitoring and daily activity assessment applications.

To address these challenges, we propose STPM, a novel framework that seamlessly integrates spatial and temporal modeling for mmWave radar point cloud sequences. Our approach extends the traditional SSM architecture through three key innovations that directly target the identified challenges: (1) a bidirectional selective mechanism that effectively captures temporal dependencies across consecutive frames, enabling detailed temporal pattern analysis, (2) the temporal processing strategy guarantees for preventing error accumulation in long sequences, crucial for maintaining stable temporal feature extraction, and (3) the grouping strategy that systematically captures both local geometric structures and global spatial relationships at multiple scales, facilitating the recognition of functionally distinct but structurally similar motions.

The primary contributions of this work can be summarized as follows:

- We introduce a novel spatial-temporal framework with theoretical guarantees, STPM, which demonstrates the importance of preserving temporal coherence through queue-based processing for model stability and performance.
- We demonstrate an architecture design that effectively distinguishes semantically similar actions across different domains by capturing both fine-grained kinematic patterns and high-level functional differences.
- We achieve state-of-the-art performance on both RadHAR and MM-Fi datasets, with significant improvements in distinguishing fine-grained motion patterns.

Extensive experiments on RadHAR [2] and MM-Fi [14] demonstrate STPM superior accuracy and robustness in distinguishing semantically similar actions, while using memory more efficiently compared to Transformer-based alternatives. Our framework exhibits particular strength in challenging

[†]These authors contributed equally and [*] donates the corresponding author.

scenarios requiring precise motion pattern recognition, from daily activities to rehabilitation exercises, with comprehensive ablation studies validating each component's contribution.

II. Related Work

Point Cloud-based Human Activity Recognition: Point cloud-based methods [15]–[17] have emerged as a promising approach to human activity recognition, demonstrating superior performance in capturing fine-grained spatial information compared to traditional approaches [18], [19]. Pioneering works like PointNet++ [20] established fundamental architectures for processing raw point cloud data while maintaining permutation invariance. Recent advances have extended these foundations through various temporal modeling strategies, including RNN-based architectures [21] and 4D convolutions [22], to effectively capture spatio-temporal relationships. In the context of mmWave radar sensing, the evolution from traditional multi-domain Doppler analysis to point cloud-based methods represents a significant paradigm shift [1]. While early approaches [23]–[28] struggled to capture fine-grained spatial information necessary for complex motion analysis, recent point cloud-based methods [2], [4], [5], [29]–[32] have demonstrated promising capabilities in providing richer three-dimensional representations of human motion. Recent works [3], [11], [33] have shown that Transformer-based architectures excel at capturing global relationships in point-cloud sequences. However, these models face significant computational challenges due to their quadratic complexity in sequence length, particularly when processing long trajectories of radar point clouds.

Mamba for Point Cloud Sequences: Inspired by State Space Models (SSMs), researchers have developed architectures [34]–[36] that achieve linear computational scaling for sequence processing, addressing the quadratic complexity challenge faced by traditional attention-based approaches [9], [10]. While Mamba [12] architecture successfully extended SSMs to handle discrete and information-dense data through its selective mechanism, the absence of positional information due to the removal of attention mechanisms poses challenges for position-sensitive data processing. Vim [37] addressed this limitation in the visual domain by reconstructing self-attention through sequential injection of visual information. Similarly, PointMamba [13] established element interactions in unstructured point cloud data through spatial reordering, effectively combining selective state spaces with unordered data. While these approaches have shown promising results, they still lack the exploration of temporal coherence in sequential point cloud data. This limitation becomes particularly pronounced in mmWave radar applications, where maintaining temporal relationships between consecutive frames is crucial for accurate activity recognition. Direct application of existing SSM architectures to mmWave radar point cloud sequences presents unique challenges, especially in maintaining temporal consistency while processing sparse and noisy data streams.

III. Method

A. SSM with Temporal Dynamics

The Mamba [12] architecture leverages the SSM to capture temporal dependencies in sequential data. We extend the traditional SSM framework by incorporating a dynamic selective mechanism, making it particularly suitable for processing temporal point cloud sequences from mmWave radar. The fundamental continuous-time state space formulation of our model is given by:

$$\frac{dx(t)}{dt} = A(\Delta(t))x(t) + B(\Delta(t))u(t), \tag{1}$$

where $x(t) \in \mathbb{R}^d$ represents the state vector and $u(t)$ denotes the input vector. For a point cloud sequence $S = \{P_1, P_2, ..., P_n\}$ where $P_t \in \mathbb{R}^{M\times 3}$, the state matrices $A(\Delta(t))$ and $B(\Delta(t))x(t)$ are dynamically adjusted by the selective parameter $\Delta(t)$, computed as:

$$\Delta(t) = \sigma(W_\Delta \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} + b_\Delta), \tag{2}$$

where σ denotes the sigmoid function, and W_Δ and b_Δ are learnable parameters. This selective mechanism enables our model to focus on salient temporal dependencies while maintaining computational efficiency. For discrete-time implementation, we employ:

$$x_{t+1} = \bar{A}(\Delta_t)x_t + \bar{B}(\Delta_t)u_t. \tag{3}$$

Our theoretical analysis shows this formulation maintains bounded state transitions under the selective mechanism, with the accumulated state transformation satisfying: $\|\Phi_t\| \leq M\gamma^{t-1}$, where M is a bounded constant and $\gamma < 1$ ensures stability. Furthermore, we prove that for consecutive point clouds, the temporal coherence satisfies $\mathcal{D}(P_t, P_{t-1}) \leq \varepsilon$, where $\mathcal{D}$ represents the Chamfer distance and ε bounds the natural sampling variation, ensuring stable feature extraction.

The stability of our selective mechanism is further guaranteed through a Lipschitz analysis. For the feature extraction function $g(P)$ and selective mechanism Δ_t, we establish:

$$\mathcal{D}(g(P_t), g(P_{t-1})) \leq L_g \mathcal{D}(P_t, P_{t-1}), \tag{4}$$

$$\mathcal{D}(\Delta_t, \Delta_{t-1}) \leq L_\Delta \mathcal{D}(g(P_t), g(P_{t-1})), \tag{5}$$

where L_g and L_Δ are respective Lipschitz constants. These bounds, combined with our temporal coherence constraint, yield a compound Lipschitz constant $K = L_\Delta L_g$ that ensures the selective mechanism remains stable during sequence processing. This theoretical framework demonstrates that our approach effectively captures temporal dependencies while preventing error accumulation in the state space representation.

B. Our Method Framework

Data Processing: A key challenge in millimeter wave radar-based activity recognition is the sparsity and noise in individual point cloud frames. To address this, we propose a temporal aggregation strategy that combines consecutive frames while maintaining temporal coherence. As shown in Fig. 1(a), let

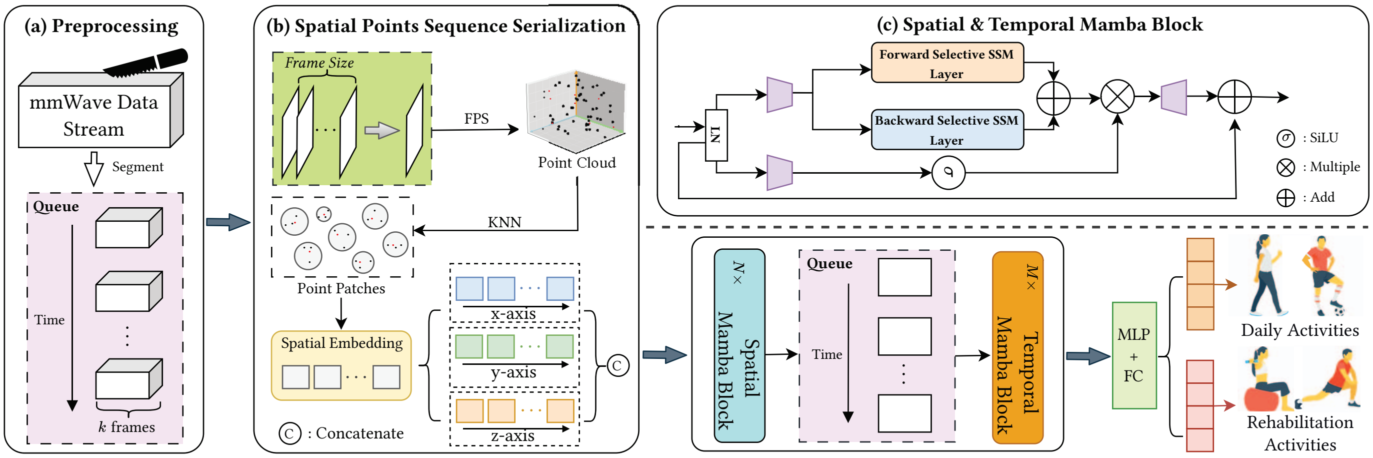

Fig. 1: Overview of the STPM for mmWave radar-based human activity recognition. (a) Preprocessing with FIFO queue-based temporal management for frame aggregation. (b) Spatial points sequence serialization through hierarchical grouping and multi-axis organization. (c) STPM architecture featuring parallel forward-backward selective SSM layers for temporal modeling. The framework processes mmWave point cloud sequences into discriminative features for human activity classification.

$W \in \mathbb{R}^{T\times P\times 3}$ denote the entire point cloud sequence, where T represents the temporal dimension and P the number of points per frame. Following the temporal coherence principle established in § III, we process this sequence through a queue Q that implements the first-in-first-out (FIFO) mechanism. Q manages segments $\{S_1, S_2, \cdots, S_i\}$, where each segment $S_j \in \mathbb{R}^{((k\cdot P)\times 3)}$ aggregates k consecutive frames into a unified point cloud representation.

Spatial Points Sequence Serialization: Given an input point cloud sequence $P \in \mathbb{R}^{M\times 3}$ with M points, we propose a hierarchical grouping strategy to capture both local geometric structures and global spatial relationships. We partition the point cloud into n groups $\mathcal{G} = \{G_1, G_2, ..., G_n\}$ using Farthest Point Sampling (FPS) for center points selection and k-Nearest Neighbors (KNN) for local patch construction. For each group, we compute center embedding $\phi_c : \mathbb{R}^3 \rightarrow \mathbb{R}^d$ and position embedding $\phi_p : \mathbb{R}^{k\times 3} \rightarrow \mathbb{R}^{N\times d}$ to capture global and local information respectively. The final group representation combines these embeddings:

$$\mathcal{G}_i = \phi_c(c_i) + \text{MLP}(\phi_p(G_n^i)). \tag{6}$$

The grouped features are then serialized along each spatial axis and concatenated:

$$\mathcal{F} = \text{Concat}[\text{Sort}_x(\mathcal{G}), \text{Sort}_y(\mathcal{G}), \text{Sort}_z(\mathcal{G})], \tag{7}$$

where Concat is the function to concatenate.

Spatial-Temporal Point Mamba: integrates forward and backward selective SSM layers in a complementary architecture, specifically designed for the unique characteristics of radar point cloud sequences. As shown in Fig. 1(c), the STPM processes input features through three parallel pathways. These operations are combined through:

$$Y = \text{LN}(\mathcal{F} + \text{FFN}(\text{Forward}(\mathcal{F}) + \text{Backward}(\mathcal{F}))), \tag{8}$$

where Forward() and Backward() denote the selective SSM operations in respective directions, and FFN() is a feed-forward network with *SiLU* activation. This bidirectional structure offers two key advantages:

- **Comprehensive Temporal Modeling**: The forward SSM layer captures causal relationships, while the backward SSM layer models reverse temporal dependencies, providing a complete understanding of temporal dynamics.
- **Efficient Feature Integration**: The parallel processing nature maintains computational efficiency while expanding the model's receptive field.

Our comprehensive ablation studies (§ IV-D) demonstrate that each component of the STPM framework contributes meaningfully to the overall performance, with the complete architecture achieving optimal results through effective integration of spatial and temporal modeling.

IV. Experimental Results

A. Dataset and Implementation Details

Dataset: In order to comprehensively evaluate the performance and generalization ability of the proposed method, this study uses two representative mmWave radar point cloud datasets: RadHAR [2] and MM-Fi [14]. The RadHAR dataset comprises 71.6 minutes of training data (12,097 samples) and 21.4 minutes of testing data (3,538 samples), covering five fundamental human activities: boxing, jumping jacks, jumping, squatting and walking. Each sample contains a sequence of point cloud frames captured by a Texas Instruments IWR1443 mmWave radar sensor. The MM-Fi dataset presents a more challenging evaluation scenario due to its diverse activity types and complex environmental conditions. It contains more than 320,000 frames of multimodal data collected from 40 subjects in four different environments. It contains 27 different daily and rehabilitation movements, including stretching exercises, chest expansions, twisting motions, and lunging activities.

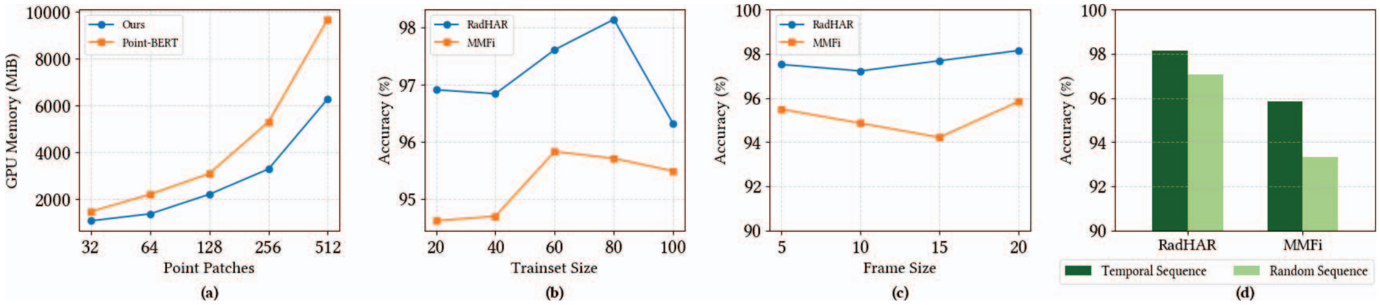

Fig. 2: Ablation studies and analysis of architectural design choices. (a) GPU memory usage comparison with Transformer-base model, showing the efficiency with increasing number of point patches. (b) Impact of training set size on model accuracy, showing robustness to reduced training data. (c) Performance with varying sequence sizes, demonstrating consistent accuracy across different frame configurations. (d) Impact of temporal coherence on recognition accuracy.

TABLE I: Comparison with state-of-the-art methods, where S. and T. denote spatial and temporal respectively.

Dataset	Method	Encoding Strategy	Accuracy (%)
RadHAR [2]	RadHAR [2]	S. & T.	90.47
	m-Activity [4]	S. & T.	93.25
	Combined DA [38]	S. & T.	93.29
	Point-BERT [9]	S.	96.90
	Dual GVPC [39]	S.	97.61
	STPM (Ours)	S. & T.	**98.14**
MM-Fi [14]	TENT [40]	S.	88.10
	STPM (Ours)	S. & T.	**95.69**

These datasets represent different application scenarios and environmental conditions, which can effectively verify the robustness of the model.

Implementation: This study was conducted on a Linux platform, leveraging Python for implementation and employing PyTorch as the framework for model development. The experimental setup featured 2× Intel Xeon Gold 5415+ processors alongside 2× NVIDIA RTX A5500 GPUs. STPM encoder comprises two independent Mamba blocks (12-layer spatial mamba block and 10-layer temporal mamba block), each featuring a 384-dimensional hidden layer. To enhance model robustness, a drop path rate of 0.3 was established.

B. Detailed Recognition Performance Analysis

1) Preference Comparison: Our comprehensive evaluation demonstrates the effectiveness of STPM across multiple metrics and datasets. It achieves state-of-the-art performance with 98.14% and 95.69% accuracy on the RadHAR and MM-Fi dataset respectively, substantially outperforming existing approaches across different encoding strategies in Table I. This improvement primarily stems from our bidirectional selective mechanism, which effectively captures both temporal dynamics and spatial features. The performance gain is particularly pronounced in complex motion recognition tasks, where conventional spatial-only approaches exhibit limitations in distinguishing subtle movement variations.

In terms of computational efficiency, our experimental results illustrated in Fig. 2 (a) demonstrate STPM's advantages over transformer-based architectures [9] when processing point cloud sequences. At 512 point patches, STPM achieves reducing memory consumption by 35% compared to transformer-based alternatives. This efficiency gain, though more modest than in previous state space models due to temporal dependency processing and sparse mmWave data characteristics, effectively balances computational efficiency with temporal coherence preservation necessary for accurate activity recognition.

2) Similarity Semantics: The confusion matrices depicted in Fig. 3 reveal STPM's superior capability in differentiating semantically similar actions. Notably, in the challenging MM-Fi rehabilitation exercise recognition task, our model achieves 97.3% accuracy in distinguishing between left and right-hand movements (A13/A14), demonstrating its effectiveness in capturing fine-grained motion patterns. This granular discrimination ability validates our temporal coherence preservation strategy and holds significant implications for practical rehabilitation monitoring applications where precise motion assessment is crucial.

3) Overview: This comprehensive analysis validates that our architecture achieves state-of-the-art performance in human activity recognition while maintaining efficient memory utilization. Quantitative evaluation through confusion matrices analysis reveals the model's effectiveness in discriminating semantically similar actions. The combination of superior accuracy, fine-grained motion discrimination, and computational efficiency positions STPM as a promising solution for visual privacy-aware human activity monitoring systems.

C. Architectural Design Analysis

To validate our theoretical framework and key innovations in temporal coherence modeling, we conduct a systematic analysis of our architecture design through three critical dimensions: data efficiency, temporal receptive field, and model scalability. The experiments provide empirical evidence supporting our architectural choices while revealing important insights about the model's behavior.

On the data efficiency front, our model demonstrates strong data efficiency in Fig. 2 (b), achieving 95.83% accuracy on RadHAR with only 20% of training data and improving to 98.1% at 80%. The model shows different scaling characteristics on MM-Fi, reaching optimal performance of 95.69% at

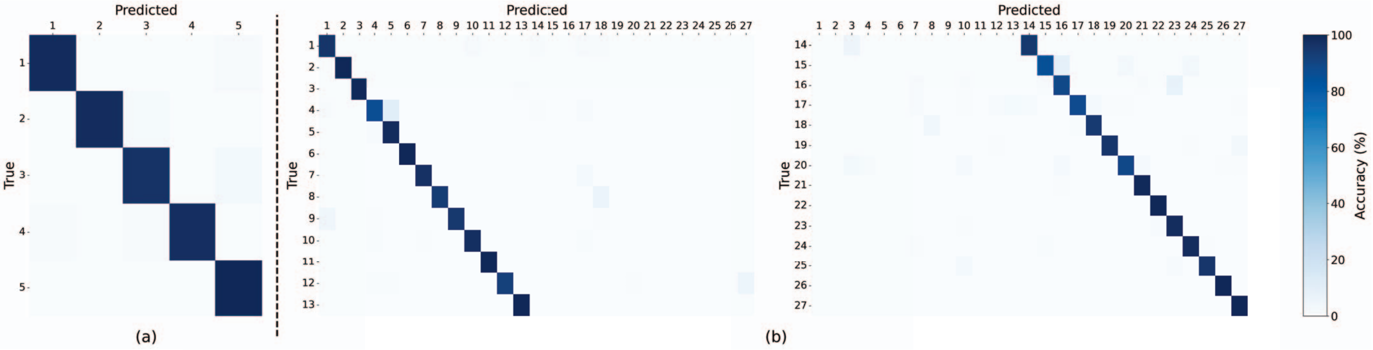

Fig. 3: Confusion Matrices of (a) RadHAR dataset, (b) MM-Fi dataset.

60% data utilization before experiencing slight performance degradation, likely due to the dataset's larger sample size and higher complexity compared to RadHAR. In addition to data efficiency, Fig. 2 (c) shows robustness to temporal block size variations, maintaining performance with accuracy variations within 2% across block sizes from 5 to 20 frames on both datasets. This stability to hyperparameter changes suggests the model's intrinsic robustness in temporal feature extraction.

The interaction between spatial and temporal components reveals optimal configurations aligning with our theoretical predictions. The diminishing returns beyond 12 spatial blocks confirm our analysis of bounded state transitions, while the sensitivity to temporal block count validates our emphasis on temporal coherence preservation. Most importantly, the superior performance of the complete architecture (98.14% on RadHAR) compared to spatial-only (90.03%) or temporal-only (89.33%) variants empirically proves the necessity of our integrated approach to spatiotemporal modeling. This architectural analysis not only validates our theoretical framework but also provides practical insights for deploying Mamba-based architectures in real-world activity recognition applications. The results demonstrate that our design choices effectively balance model capacity, computational efficiency, and recognition accuracy while maintaining the critical property of temporal coherence.

TABLE II: Ablation study on model components.

Spatial Block	Temporal Block	Random Sequence	Accuracy (%)
×	✓	×	89.33
✓	×	×	90.03
✓	✓	✓	97.05
✓	✓	×	**98.14**

D. Ablation Studies

To systematically validate the effectiveness of each proposed component, we conduct comprehensive ablation studies on the RadHAR dataset. Our first experiment examines the interaction between spatial and temporal blocks. As shown in Table II, using either spatial blocks (90.03%) or temporal blocks (89.33%) alone results in substantially degraded performance compared to our complete architecture (98.14%). This significant performance gap reveals the complementary nature of spatial-temporal feature extraction - spatial blocks capture local geometric patterns while temporal blocks model motion dynamics, and their integration enables holistic activity understanding. When shuffling sequence instead of our temporal coherence preservation strategy, the accuracy drops to 97.05%. This degradation suggests that maintaining temporal relationships is crucial for distinguishing subtle motion patterns that differ primarily in their execution order.

TABLE III: Performance of different serialization.

X-axis	Y-axis	Z-axis	Accuracy (%)
✓	×	×	97.64
✓	✓	×	97.15
×	✓	×	95.36
×	✓	✓	97.01
×	×	✓	96.45
✓	×	✓	96.91
✓	✓	✓	**98.14**

We further examine our multi-axis serialization strategy through different axis combinations, as presented in Table III. Single-axis serialization (X-axis: 97.64%, Y-axis: 95.36%, Z-axis: 96.45%), while providing reasonable performance, cannot fully capture the complete spatial relationships in 3D motion patterns. The optimal performance (98.14%) is achieved only when combining all three axes, empirically demonstrating that comprehensive spatial modeling across multiple perspectives is essential for accurate motion recognition. This finding aligns with our theoretical analysis that effective activity recognition requires both detailed spatial feature extraction and robust temporal coherence preservation.

V. Conclusion

This papaer presented STPM, a novel framework for mmWave radar point cloud sequence processing that effectively integrates spatial and temporal modeling through bidirectional selective mechanism and queue-based temporal processing. Our comprehensive experiments on RadHAR and MM-Fi datasets demonstrate superior performance in distinguishing semantically similar actions while maintaining computational efficiency. The theoretical guarantees and empirical results establish STPM as a promising solution for visual privacy-aware human activity recognition, though challenges remain in handling extremely sparse point clouds and multi-person scenarios. Future work could focus on extending the framework to address these limitations while maintaining its computational advantages.

REFERENCES

[1] J. Zhang and et al., "A survey of mmwave-based human sensing: Technology, platforms and applications," *Commun. Surveys Tuts.*, vol. 25, no. 4, p. 2052–2087, Jul. 2023. [Online]. Available: https://doi.org/10.1109/COMST.2023.3298300

[2] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 2019, pp. 51–56.

[3] M. A. U. Alam, M. M. Rahman, and J. Q. Widberg, "Palmar: Towards adaptive multi-inhabitant activity recognition in point-cloud technology," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[4] Y. Wang, H. Liu, K. Cui, A. Zhou, W. Li, and H. Ma, "m-activity: Accurate and real-time human activity recognition via millimeter wave radar," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8298–8302.

[5] S. Wang, D. Cao, R. Liu, W. Jiang, T. Yao, and C. X. Lu, "Human parsing with joint learning for dynamic mmwave radar point cloud," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–22, 2023.

[6] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas *et al.*, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th annual international conference on mobile computing and networking*, 2018, pp. 289–304.

[7] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Gan-based synthetic radar micro-doppler augmentations for improved human activity recognition," in *2019 IEEE Radar Conference (RadarConf)*. IEEE, 2019, pp. 1–5.

[8] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *CVPR*, 2018, pp. 7356–7365.

[9] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *CVPR*, 2022.

[10] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *CVPR*, 2021, pp. 16 259–16 268.

[11] L. Kang, Z. Li, X. Zhao, Z. Zhao, and T. Braun, "St-pct: Spatial-temporal point cloud transformer for sensing activity based on mmwave," *IEEE Internet of Things Journal*, 2023.

[12] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2024. [Online]. Available: https://arxiv.org/abs/2312.00752

[13] D. Liang, X. Zhou, W. Xu, X. Zhu, Z. Zou, X. Ye, X. Tan, and X. Bai, "Pointmamba: A simple state space model for point cloud analysis," in *Advances in Neural Information Processing Systems*, 2024.

[14] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: https://openreview.net/forum?id=1uAsASS1th

[15] Y. Wang, Y. Xiao, F. Xiong, W. Jiang, Z. Cao, J. T. Zhou, and J. Yuan, "3dv: 3d dynamic voxel for action recognition in depth video," in *CVPR*, 2020, pp. 511–520.

[16] X. Liu, M. Yan, and J. Bohg, "Meteornet: Deep learning on dynamic 3d point cloud sequences," in *CVPR*, 2019, pp. 9246–9255.

[17] Y. Min, Y. Zhang, X. Chai, and X. Chen, "An efficient pointlstm for point clouds based gesture recognition," in *CVPR*, 2020, pp. 5761–5770.

[18] V. Lepetit, P. Lagger, and P. Fua, "Randomized trees for real-time keypoint recognition," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 775–781.

[19] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Bmvc 2009-british machine vision conference*. BMVA Press, 2009, pp. 124–1.

[20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[21] X. Ye, J. Li, H. Huang, L. Du, and X. Zhang, "3d recurrent neural networks with context fusion for point cloud semantic segmentation," in *ECCV*, 2018, pp. 403–417.

[22] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *CVPR*, 2019, pp. 3075–3084.

[23] A. Khamis, B. Kusy, C. T. Chou, M.-L. McLaws, and W. Hu, "Rfwash: a weakly supervised tracking of hand hygiene technique," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, ser. SenSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 572–584. [Online]. Available: https://doi.org/10.1145/3384419.3430733

[24] E. Kurtoğlu, A. C. Gurbuz, E. A. Malaia, D. Griffin, C. Crawford, and S. Z. Gurbuz, "Asl trigger recognition in mixed activity/signing sequences for rf sensor-based user interfaces," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 699–712, 2021.

[25] A. J. Akbar, Z. Sheng, Q. Zhang, and D. Wang, "Cross-domain gesture sequence recognition for two-player exergames using cots mmwave radar," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. ISS, pp. 327–356, 2023.

[26] V. Polfliet, N. Knudde, B. Vandersmissen, I. Couckuyt, and T. Dhaene, "Structured inference networks using high-dimensional sensors for surveillance purposes," in *Engineering Applications of Neural Networks: 19th International Conference, EANN 2018, Bristol, UK, September 3-5, 2018, Proceedings 19*. Springer, 2018, pp. 16–27.

[27] L. Werthen-Brabants, G. Bhavanasi, I. Couckuyt, T. Dhaene, and D. Deschrijver, "Split birnn for real-time activity recognition using radar and deep learning," *Scientific Reports*, vol. 12, no. 1, p. 7436, 2022.

[28] J. Pegoraro, J. O. Lacruz, M. Rossi, and J. Widmer, "Sparcs: A sparse recovery approach for integrated communication and human sensing in mmwave systems," in *IPSN*. IEEE, 2022, pp. 79–91.

[29] H. Liu, K. Cui, K. Hu, Y. Wang, A. Zhou, L. Liu, and H. Ma, "mtranssee: Enabling environment-independent mmwave sensing based gesture recognition via transfer learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–28, 2022.

[30] F. Jin, A. Sengupta, and S. Cao, "mmfall: Fall detection using 4-d mmwave radar and a hybrid variational rnn autoencoder," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 2, pp. 1245–1257, 2020.

[31] Y.-H. Wu, H.-C. Chiang, S. Shirmohammadi, and C.-H. Hsu, "A dataset of food intake activities using sensors with heterogeneous privacy sensitivity levels," in *Proceedings of the 14th Conference on ACM Multimedia Systems*, 2023, pp. 416–422.

[32] F. Ding, Z. Luo, P. Zhao, and C. X. Lu, "milliflow: Scene flow estimation on mmwave radar point cloud for human motion sensing," in *ECCV*. Springer, 2025, pp. 202–221.

[33] K. Deng, D. Zhao, Q. Han, Z. Zhang, S. Wang, A. Zhou, and H. Ma, "Midas: Generating mmwave radar data from videos for training pervasive and privacy-preserving human sensing tasks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–26, 2023.

[34] 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," *Advances in neural information processing systems*, vol. 34, pp. 572–585, 2021.

[35] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *ICLR*, 2022.

[36] A. Orvieto, S. L. Smith, A. Gu, A. Fernando, C. Gulcehre, R. Pascanu, and S. De, "Resurrecting recurrent neural networks for long sequences," in *International Conference on Machine Learning*. PMLR, 2023, pp. 26 670–26 698.

[37] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[38] Z. Wang, D. Jiang, B. Sun, and Y. Wang, "A data augmentation method for human activity recognition based on mmwave radar point cloud," *IEEE Sensors Letters*, vol. 7, no. 5, pp. 1–4, 2023.

[39] C. Yu, Z. Xu, K. Yan, Y.-R. Chien, S.-H. Fang, and H.-C. Wu, "Noninvasive human activity recognition using millimeter-wave radar," *IEEE Systems Journal*, vol. 16, no. 2, pp. 3036–3047, 2022.

[40] Y. Zhou, J. Yang, H. Zou, and L. Xie, "Tent: Connect language models with iot sensors for zero-shot activity recognition," 2023. [Online]. Available: https://arxiv.org/abs/2311.08245