

Article

RPT-Mamba: A Range-Aware Physical Token Mamba Network for Far-Field mmWave Radar Gesture Recognition

Yitong Shi ¹, Pei Peng ¹ and Zhiyuan Wang ^{2,*}¹ School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China² School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing 210023, China

* Correspondence: wangzy@nnu.edu.cn

Abstract

Millimeter-wave (mmWave) radar provides a privacy-preserving and illumination-robust sensing modality for contactless gesture recognition. However, sparse radar point clouds degrade substantially as sensing distance increases: the number of valid detections decreases, echo intensity attenuates, and Doppler-related motion cues become less reliable. Such range-induced degradation leads to a distribution shift between near-range training samples and far-field test samples, making it difficult for models trained at short distances to generalize to unseen longer distances. Existing point-cloud gesture recognition methods usually treat radar detections as generic sparse point sequences and rarely model distance-related point loss, echo attenuation, and physical-attribute unreliability explicitly. This work introduces RPT-Mamba, a range-aware physical token Mamba network for sparse mmWave radar point cloud sequences. RPT-Mamba constructs physical point tokens from spatial coordinates, Doppler velocity, echo intensity, point-level range, and sample-level range information. During training, a range-aware stochastic degradation strategy adaptively removes points and masks dynamic attributes according to the estimated sensing distance, while a context-guided attribute reconstruction objective recovers masked Doppler and intensity attributes from spatial and frame-level context. A bidirectional Mamba temporal encoder then models long-range gesture dynamics over frame tokens. On the public mTransSee dataset, RPT-Mamba achieves 92.09% accuracy and 92.04% Macro-F1 under the random split protocol, and 85.34% accuracy and 84.77% Macro-F1 under a challenging near-to-far protocol, exceeding point-cloud, radar-gesture, Transformer, and Mamba baselines.

Keywords: mmWave radar; gesture recognition; sparse point cloud; range generalization; Mamba; attribute reconstruction



Academic Editor: Athanasios Nikolaidis

Received: 17 May 2026

Revised: 15 June 2026

Accepted: 25 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Gesture recognition is an important natural human–computer interaction technique and has been widely studied for smart homes, in-vehicle interaction, wearable devices, augmented reality, and contactless control. Existing methods mainly rely on RGB cameras, depth cameras, inertial sensors, or wireless sensing devices. Vision-based methods can capture rich appearance and spatial information, but their performance is affected by illumination changes, occlusion, privacy concerns, and complex backgrounds. Wearable-sensor methods often achieve stable recognition accuracy, but require users to wear additional devices, which limits their convenience in natural interaction scenarios. In contrast, mmWave

radar is contactless, robust to lighting conditions, privacy preserving, and sensitive to subtle motion; it has therefore become an increasingly important sensing modality for gesture recognition.

An mmWave radar can estimate target range, radial velocity, angle, and echo intensity by transmitting and receiving high-frequency electromagnetic waves. Existing mmWave gesture recognition methods usually use one of two representations. One line of work converts radar returns into image-like features, such as range-Doppler maps, range-time maps, Doppler-time maps, angle-time maps, range-angle maps, or micro-Doppler spectrograms, and then applies conventional machine learning, convolutional neural networks (CNNs), recurrent networks, or Transformer-based backbones. Recent studies show that multifeature radar-map fusion improves robustness in complex scenes [1], and range-Doppler-map-based residual recurrent attention networks have also been explored for gesture recognition [2]. These dense representations are convenient for image backbones, and signal attenuation or weakened responses can still be reflected in radar maps. However, the gridding process aggregates individual detections into discretized cells, making point-level detection loss and per-point attribute reliability less explicit than in point-cloud representations.

The other line of work directly models sparse radar point clouds, where each detected point carries spatial coordinates and physical attributes such as Doppler velocity and echo intensity. Recent mmWave gesture recognition studies have used multiscale point-cloud feature extraction and sparse point-cloud spatio-temporal modeling [3,4]. This paper follows the point-cloud route not because it is simpler, but because it preserves the detection-level structure needed to analyze far-field degradation. In practical gesture recognition, the sensing distance between the user and the radar is often variable. As the distance increases, radar echoes become weaker, valid detections may disappear, and Doppler/intensity attributes become less reliable. These effects manifest in point-cloud representations as point loss, intensity attenuation, and degraded dynamic attributes.

The difficulty of modeling mmWave radar point clouds arises at both the signal and representation levels. At the signal level, recent 4D mmWave radar point-cloud studies report inherent sparsity, noise, limited spatial resolution, weak geometric detail, multipath artifacts, and false targets [5,6]. At the representation level, a gesture sample is an unordered and frame-dependent point sequence: the number of detections changes from frame to frame, while the discriminative motion pattern is distributed over time. Far-field sensing further intensifies this problem by reducing point density and weakening physical attributes. Therefore, robust far-field gesture recognition requires joint modeling of sparse point cloud structure, distance-related physical degradation, and long-range temporal dynamics.

Figure 1 illustrates the same gesture under near-field and far-field conditions. The near-field point cloud is dense and structurally clear, whereas the far-field point cloud is sparse and scattered. As shown in the example gesture in Figure 1, the number of valid points per frame decreases from 22.0 at 1.2 m to 8.5 at 4.8 m, approximately 39% of the near-field value. This pronounced density reduction suggests that sensing distance should be treated as an explicit nuisance factor rather than being left to implicit feature learning.

Existing mmWave radar gesture recognition methods can be broadly divided into image-like radar feature modeling and sparse point cloud modeling. Image-like methods make radar processing compatible with mature visual backbones, but gridding can obscure which detections are missing or physically unreliable. Point-cloud methods operate directly on radar detections using PointNet [7], PointNet++ [8], recurrent networks [9], Transformers [10], or state-space models. For example, Pantomime uses PointNet++ and long short-term memory (LSTM) for sparse mmWave point cloud sequences [11]; ST-PCT

introduces a point cloud Transformer for mmWave activity sensing [12]; and STPM adopts Mamba for mmWave radar point cloud activity recognition [13]. These methods mainly report in-distribution recognition accuracy, while the distance-induced shifts in point density, echo strength, and attribute reliability are less explicitly modeled.

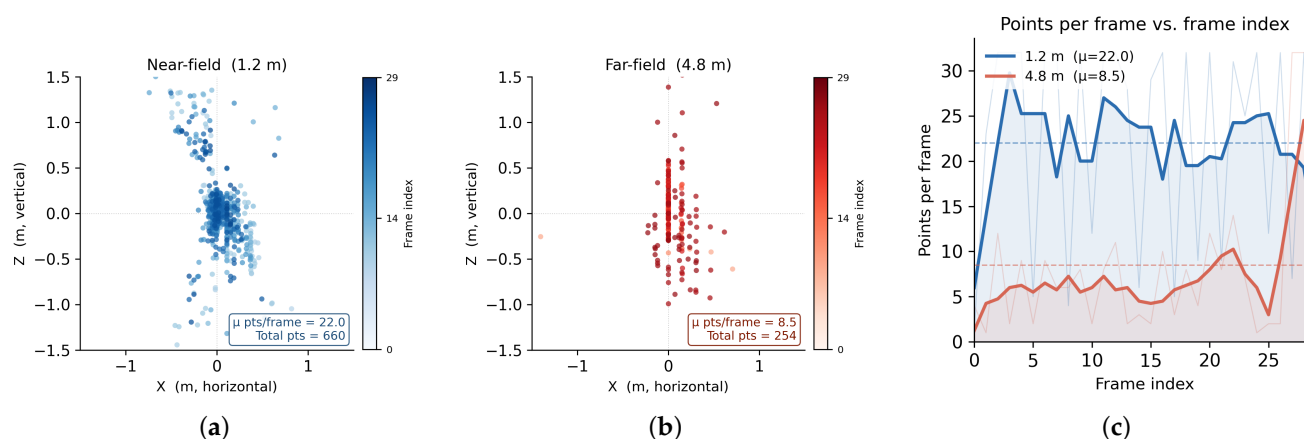


Figure 1. Comparison of the same PL gesture under near-field (1.2 m) and far-field (4.8 m) mmWave radar sensing. Each sample is preprocessed identically to the model input: every 10th raw frame is selected and at most 30 frames are retained (frame indices 0–29 on the color bar). The two point-cloud views share the same coordinate range. At 4.8 m, valid detections become much sparser and the gesture trajectory is less separable. (a) Near field (1.2 m). (b) Far field (4.8 m). (c) Valid points per frame.

RPT-Mamba is specifically designed to address this distance-induced distribution shift. Compared with radar-map-based range-robust methods, RPT-Mamba keeps the point-level physical attributes and explicitly models which detections and attributes become unreliable as range increases. Compared with radar point cloud Mamba models such as STPM, which mainly use Mamba as a spatio-temporal sequence encoder, RPT-Mamba differs in three aspects: it constructs physical tokens from spatial coordinates, Doppler velocity, echo intensity, point-level range, and sample-level range; it applies distance-adaptive point dropping and attribute masking as a training-time degradation process; and it uses context-guided reconstruction to regularize the bidirectional Mamba encoder under incomplete far-field observations. A bidirectional Mamba temporal encoder then processes the frame-token sequence, enabling the final classifier to incorporate both the physical structure of individual radar detections and the temporal evolution of the gesture.

The main contributions are as follows:

- C1. Physical point tokenization.** We encode spatial coordinates, Doppler velocity, echo intensity, point-level range, and sample-level range condition, allowing distance information to enter the model before temporal aggregation.
- C2. Range-aware degradation learning.** We link the perturbation strength to the estimated sensing distance. Together with Doppler/intensity reconstruction, this design converts far-field signal attenuation into an explicit training signal rather than a hidden source of distribution shift.
- C3. Near-to-far evaluation protocol.** We construct a near-to-far protocol on mTransSee. Under this setting, RPT-Mamba improves accuracy over STPM from 81.67% to 85.34% and raises the extreme far-field Bucket 3 accuracy from 73.24% to 78.56%.

2. Related Work

2.1. mmWave Radar Gesture Recognition

Gesture recognition has been extensively studied for natural human–computer interaction. Compared with RGB or depth cameras and wearable inertial sensors, mmWave radar

offers contactless sensing, lighting robustness, and better privacy protection. Recent advances in high-frequency sensing and near-field array imaging further show the continuing development of radar-oriented signal processing techniques [14]. Early mmWave radar gesture recognition methods typically transform radar returns into range-Doppler maps, range-angle maps, micro-Doppler spectrograms, or other image-like features, followed by conventional machine learning or convolutional neural networks. Recent radar-map and signal-level methods further combine multiple feature views, cross-modal synthesis, or attention-based temporal modeling to improve robustness [1,2,15]. Such feature maps make radar processing compatible with image backbones, but they also aggregate individual detections, Doppler attributes, and frame-wise point sparsity into dense representations.

With the development of radar point cloud processing, an increasing number of studies have directly used sparse mmWave radar point clouds for gesture recognition. Point cloud sequences preserve spatial position, Doppler velocity, and echo intensity variations during hand motion. The mTransSee dataset and its baseline mSeeNet provide public data and an evaluation basis for environment-independent mmWave gesture recognition [16]. GesturePrint further studies user identification and cross-dataset generalization for mmWave gesture recognition [17]. Pantomime combines PointNet++ and LSTM for sparse radar point cloud gesture recognition [11]. Recent work has also explored traffic-police gestures, 3D point cloud spatiotemporal information, multiscale features, local-global Transformer representations, smart-home sparse point cloud modeling, and lightweight sparse radar point-cloud activity recognition [3,4,18–23]. These studies indicate a shift from handcrafted or image-like radar features toward direct point-level modeling.

However, most existing methods focus on conventional in-distribution evaluation and pay insufficient attention to point cloud sparsification, echo attenuation, and distribution shift induced by sensing distance. In practical deployments, the distance between the user and the radar is often variable. Far-field gestures usually produce sparser and noisier point clouds, resulting in measurable performance degradation. Improving model generalization to unseen far-field distances remains a key challenge.

2.2. mmWave Point Cloud Sequence Modeling

Point clouds are unordered, sparse, and irregular. For mmWave radar gesture recognition, each sample consists of consecutive frames of sparse detections; a model must extract spatial structure within each frame and dynamic dependencies across frames. In spatial modeling, PointNet first applies deep learning directly to unordered point sets using shared multilayer perceptrons and symmetric aggregation [7]. PointNet++ further introduces local grouping and hierarchical feature extraction [8]. Graph neural networks, point convolutions, and point cloud Transformers have also been used to model local relationships and global dependencies.

In temporal modeling, a common framework first extracts frame-level point cloud features and then uses LSTM [9], temporal CNNs, or Transformers [10] for sequence modeling. LSTM can capture temporal order but is limited in long-range dependency modeling and parallel training. Temporal CNNs are efficient but constrained by kernel sizes. Transformers offer global interaction, but their quadratic complexity can be costly for long sequences. Recent radar activity recognition studies have explored lightweight Transformers, residual attention structures, and point cloud video models to improve dynamic representation [2,24,25].

Point cloud video models and state-space models have further advanced dynamic point cloud recognition. P4Transformer treats point cloud sequences as 4D point cloud data and uses a Transformer to model spatial and temporal relationships [26]. Mamba and related state-space models provide linear-time long-sequence modeling and have been

applied to vision, video, and point cloud analysis [27–35]. Radar-oriented Mamba studies have also appeared in micro-Doppler activity recognition, mmWave point-cloud activity recognition, and 4D mmWave point-cloud enhancement [6,13,36]. Nevertheless, mmWave radar point clouds differ from LiDAR and depth-camera point clouds because they are much sparser and include physical attributes such as Doppler velocity and echo intensity. A generic point sequence encoder can model temporal order, but it does not explicitly account for the range-dependent reliability of point survival, intensity, or Doppler cues. RPT-Mamba introduces this range-conditioned modeling stage before temporal encoding.

The distinction from existing radar point cloud Mamba and range-robust gesture recognition methods is therefore methodological rather than only architectural. STPM adopts Mamba to model spatial-temporal dependencies in mmWave point clouds, but it does not explicitly encode sample-level sensing range, nor does it construct a distance-conditioned degradation and reconstruction objective. Range-robust radar gesture methods based on radar maps or multifeature fusion can improve robustness to environmental or signal variations, but their dense representations make point survival, Doppler reliability, and echo-intensity attenuation less explicit. In contrast, RPT-Mamba introduces a physical point tokenizer that separates spatial and dynamic attributes while injecting point-level and sample-level range cues; applies a range-aware stochastic degradation strategy that increases point dropping and attribute masking with sensing distance; and uses bidirectional Mamba after frame aggregation to exploit temporal context from both directions. This design makes the temporal encoder aware of range-induced point loss through the training objective, rather than treating Mamba as a generic sequence modeling block.

2.3. Auxiliary Reconstruction Learning and State-Space Models

Auxiliary reconstruction learning constructs recovery objectives beyond the main task. Masked autoencoding has been effective in vision and point cloud representation learning [37–41], and robustness-oriented learning objectives have also been studied in time-series anomaly detection and physics-informed learning [42,43]. Existing masking strategies usually include random masking, block masking, and tube masking. They primarily focus on geometric or contextual completion, but rarely consider that mmWave radar attributes have different failure modes: a point may disappear entirely, or it may remain spatially observable while Doppler and intensity become unreliable.

In mmWave radar gesture recognition, sensing distance affects point density, echo intensity, and observable motion trajectories. Conventional random masking treats near-field and far-field samples similarly, although their observation quality differs. Range-aware stochastic degradation instead increases the probability of point dropping and attribute masking with the estimated sample range, matching the empirical decline in valid detections and echo intensity.

Long-sequence modeling is also crucial for dynamic gesture recognition. Mamba is a selective state-space model that performs sequence modeling through input-dependent state updates [27]. Compared with Transformers, Mamba has lower long-sequence complexity; compared with recurrent networks, it supports more efficient parallel training and stronger long-range dependency modeling. Recent point Mamba studies indicate that tokenization, serialization, local geometry modeling, and bidirectional scanning can noticeably affect point cloud performance [30–33,35]. Efficient sequence and dynamical-system modeling has also been explored in neuromorphic computation and nonlinear multiagent systems [44–46]. These observations support the use of Mamba after radar-specific point token construction, rather than as a generic replacement for the temporal encoder.

3. Method

3.1. Overall Architecture

RPT-Mamba takes sparse frequency-modulated continuous-wave (FMCW) mmWave radar point cloud sequences directly as input. Compared with methods that convert radar observations into time-frequency images or heat maps, this design preserves point-level spatial positions, Doppler velocity, and echo intensity. Given a gesture sample, the input is a sequence of T sparse point clouds:

$$\mathcal{P} = \{P^1, P^2, \dots, P^T\}, \quad P^t = \{p_i^t\}_{i=1}^{N_t}, \quad (1)$$

where P^t denotes the point set of frame t , N_t is the number of detected valid points, and p_i^t is the i th point in frame t . Since different frames contain different numbers of points, each frame is padded to a maximum number N , and a padding mask $M^t \in \{0, 1\}^N$ is used to mark valid points.

At inference time, the main pipeline is expressed as:

$$\hat{y} = \Phi_{\text{cls}}(\Phi_{\text{temp}}(\Phi_{\text{agg}}(\Phi_{\text{tok}}(\mathcal{P})))), \quad (2)$$

where Φ_{tok} is the physical point tokenizer, Φ_{agg} is the frame aggregator, Φ_{temp} is the bidirectional Mamba temporal encoder, and Φ_{cls} is the classification head. During training, range-aware stochastic degradation and the attribute reconstruction decoder are used to construct an auxiliary loss. These components are removed at inference time and therefore do not increase inference cost. The overall architecture is shown in Figure 2.

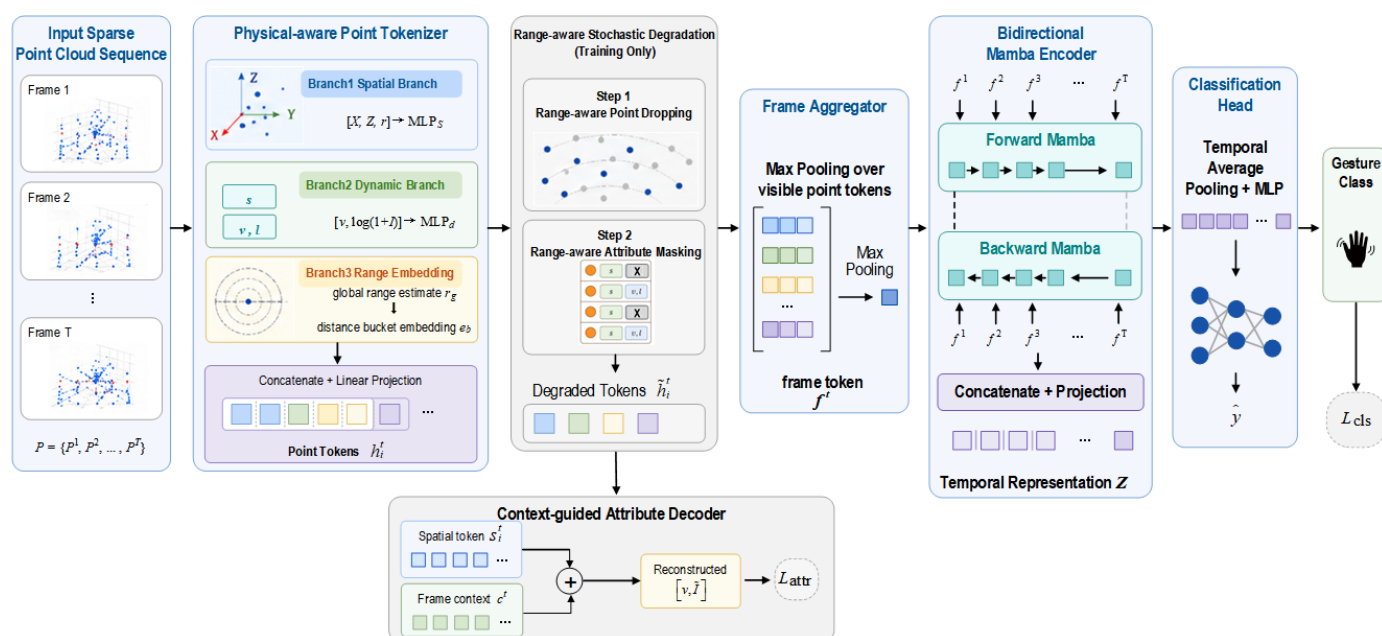


Figure 2. Overall architecture of RPT-Mamba. The framework first constructs physical point tokens, applies range-aware stochastic degradation during training, reconstructs masked Doppler and intensity attributes through a context-guided decoder, aggregates visible point tokens into frame tokens, and finally performs bidirectional temporal modeling with Mamba. Gray modules are used only during training.

3.2. Physical-Aware Point Tokenizer

Each mmWave radar detection contains spatial position as well as physical measurements such as Doppler velocity and echo intensity. Under far-field conditions, these components degrade differently: detected spatial locations still provide geometric constraints,

while Doppler and intensity are more sensitive to weak echoes and low signal-to-noise ratios. We therefore explicitly encode spatial geometry and dynamic physical attributes through two multilayer perceptron (MLP) branches.

Range estimation. To avoid relying on manually annotated sensing distance at inference time, the sample-level range condition is estimated from the input point cloud itself. Because valid points in this dataset mainly lie on the X - Z plane and the Y coordinate varies less due to sensor placement and angular resolution, we approximate point-level range as:

$$r_i^t = \sqrt{(X_i^t)^2 + (Z_i^t)^2}. \quad (3)$$

The median distance over all valid points in a sample is used as the sample-level range estimate:

$$\bar{r} = \text{Median}\{r_i^t \mid M_i^t = 1, 1 \leq t \leq T\}. \quad (4)$$

It is normalized as:

$$\hat{r}_g = \text{clip}\left(\frac{\bar{r} - d_{\min}}{d_{\max} - d_{\min}}, 0, 1\right), \quad (5)$$

where $d_{\min} = 1.2$ m and $d_{\max} = 4.8$ m are the minimum and maximum sensing distances covered by the dataset. The operator $\text{clip}(x, 0, 1) = \min(\max(x, 0), 1)$ bounds the normalized range to $[0, 1]$. Thus, samples at the nearest supported distance are assigned a degradation severity close to 0, samples at the farthest supported distance are assigned a severity close to 1, and occasional estimates outside the calibrated range are saturated rather than extrapolated. We use the median over valid point-level ranges instead of the mean because sparse radar frames may contain isolated noisy detections or multipath outliers; the median provides a more stable sample-level range estimate for controlling the degradation strength. According to $\bar{r}$, each sample is assigned to one of four distance buckets $b \in \{0, 1, 2, 3\}$, corresponding to 1.2–1.8 m, 2.1–2.7 m, 3.0–3.9 m, and 4.2–4.8 m. Each bucket has a learnable range embedding.

Branch encoding. For the i th valid point in frame t , the spatial branch takes $[X_i^t, Z_i^t, r_i^t]$ as input:

$$\mathbf{s}_i^t = \text{MLP}_s([X_i^t, Z_i^t, r_i^t]) \in \mathbb{R}^{C/2}. \quad (6)$$

The dynamic branch takes Doppler velocity and log-transformed echo intensity:

$$\mathbf{u}_i^t = \text{MLP}_u([v_i^t, \tilde{I}_i^t]) \in \mathbb{R}^{C/2}, \quad \tilde{I}_i^t = \log(1 + I_i^t). \quad (7)$$

The logarithmic transform compresses the intensity dynamic range and alleviates the scale difference between strong near-field echoes and weak far-field echoes.

Token fusion. The two branch outputs are concatenated, projected into a C -dimensional token space, and augmented with the range-bucket embedding:

$$\mathbf{h}_i^t = W_{\text{proj}}[\mathbf{s}_i^t \parallel \mathbf{u}_i^t] + \mathbf{e}_b \in \mathbb{R}^C, \quad (8)$$

where $\parallel$ denotes feature concatenation and $\mathbf{e}_b$ is the learnable embedding of bucket b . This design allows a point token to jointly encode local spatial structure, dynamic radar attributes, and the sample-level range condition. Figure 3 summarizes the structure of the physical-aware point tokenizer.

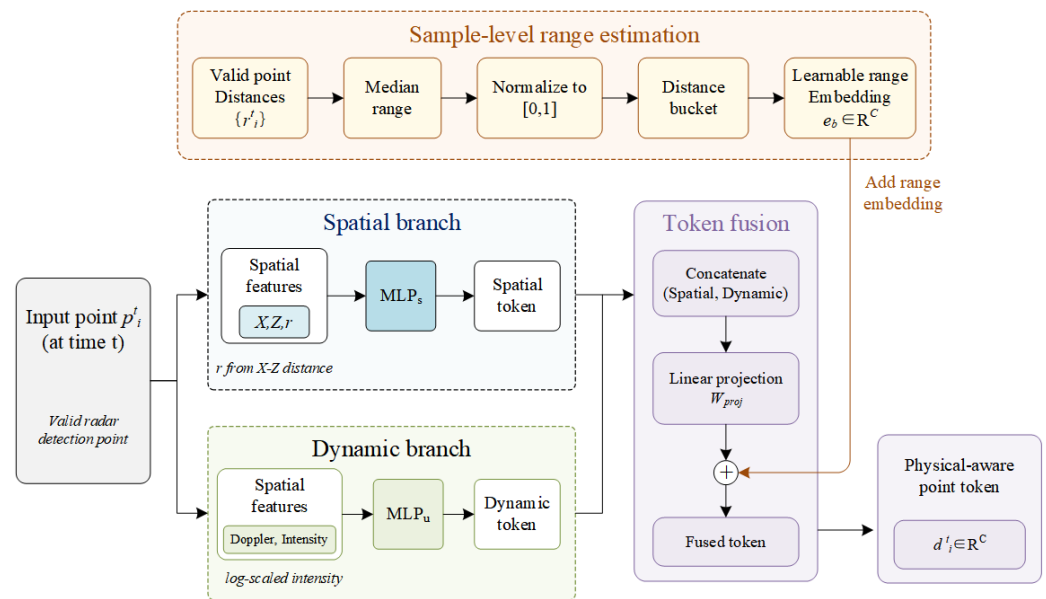


Figure 3. Structure of the Physical-aware Point Tokenizer. Spatial coordinates $[X, Z, r]$ and dynamic attributes $[v_D, \log(1 + I)]$ are encoded by two separate multilayer perceptron (MLP) branches. The two branch features are concatenated, linearly projected, and combined with a range-bucket embedding before temporal aggregation.

3.3. Range-Aware Stochastic Degradation

Far-field radar point clouds primarily exhibit two forms of degradation. First, echo energy attenuates with distance, and weak reflection points may fail to exceed the detection threshold, resulting in fewer valid detections. Second, for retained points, weaker echoes and lower signal-to-noise ratios reduce the reliability of Doppler velocity and echo intensity attributes. Figure 4 reports dataset statistics under different distances. The number of valid points per frame decreases with distance, and the average log intensity also attenuates, confirming substantial far-field degradation.

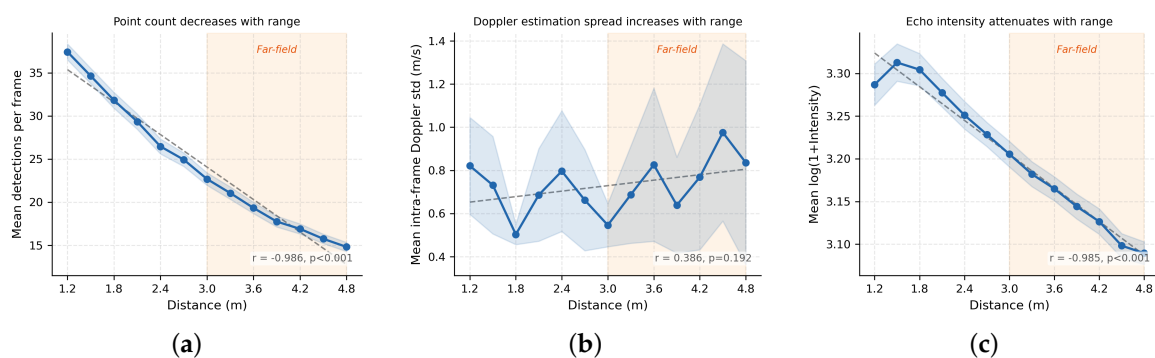


Figure 4. Statistical analysis of mmWave radar point cloud degradation under different sensing distances. The solid blue lines show distance-wise mean values, the blue shaded bands indicate variability around the mean, the gray dashed lines show fitted trends, and the orange shaded regions mark far-field distances. The average number of valid points and average log echo intensity decrease as distance increases, motivating the proposed range-aware stochastic degradation. (a) Valid points. (b) Doppler std. (c) Log intensity.

We formulate range-related degradation as two mutually exclusive training-time perturbations: range-aware point dropping and range-aware attribute masking.

Range-aware point dropping. For each valid point, a dropping variable is sampled as:

$$m_{\text{drop},i}^t \sim \text{Bernoulli}(\rho_{\text{drop}}(\hat{f}_g^t)), \quad (9)$$

where $m_{\text{drop},i}^t = 1$ means the point is removed from frame aggregation and attribute reconstruction. The dropping probability increases linearly with the normalized range:

$$\rho_{\text{drop}}(\hat{r}_g) = 0.15 + 0.35\hat{r}_g. \quad (10)$$

Thus, the dropping probability increases from 0.15 at the nearest range to 0.50 at the farthest range.

Range-aware attribute masking. For valid points that are not dropped, an attribute masking variable is sampled as:

$$m_{\text{attr},i}^t \sim \text{Bernoulli}(\rho_{\text{attr}}(\hat{r}_g)), \quad \text{if } m_{\text{drop},i}^t = 0, \quad (11)$$

with

$$\rho_{\text{attr}}(\hat{r}_g) = 0.20 + 0.30\hat{r}_g. \quad (12)$$

Attribute masking does not remove the detected point. Instead, only the dynamic-branch output is zeroed, while the spatial branch is retained:

$$\tilde{\mathbf{h}}_i^t = W_{\text{proj}}[\mathbf{s}_i^t \| \mathbf{0}_{C/2}] + \mathbf{e}_b. \quad (13)$$

This approximates the situation in which Doppler velocity and echo intensity become unreliable under weak far-field echoes.

The perturbed point token during training is therefore:

$$\tilde{\mathbf{h}}_i^t = \begin{cases} \text{invalid,} & m_{\text{drop},i}^t = 1, \\ W_{\text{proj}}[\mathbf{s}_i^t \| \mathbf{0}_{C/2}] + \mathbf{e}_b, & m_{\text{drop},i}^t = 0, m_{\text{attr},i}^t = 1, \\ \mathbf{h}_i^t, & m_{\text{drop},i}^t = 0, m_{\text{attr},i}^t = 0. \end{cases} \quad (14)$$

To avoid degenerate frames, at least one original valid point is kept in each non-empty frame during training. If a raw frame has no valid points, a learnable empty-frame token is used.

Figure 5 illustrates the training-time degradation and attribute reconstruction process.

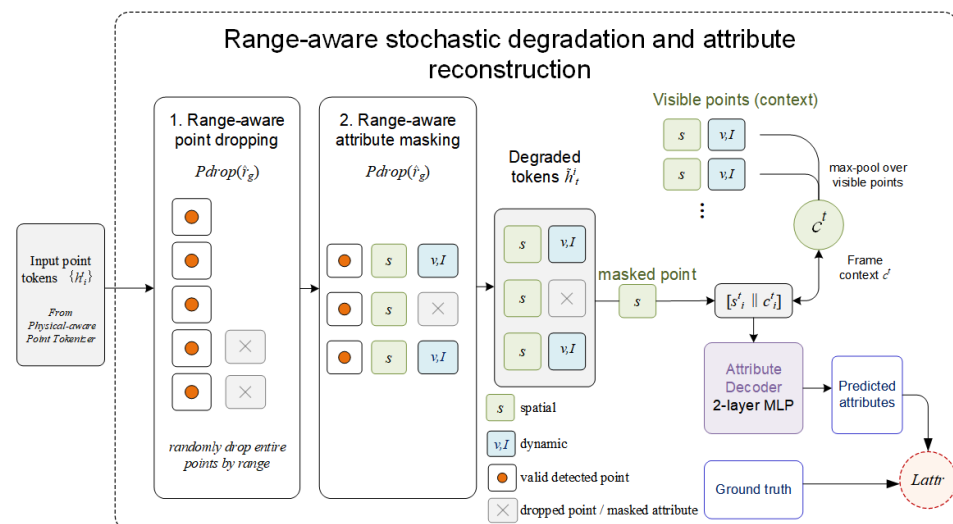


Figure 5. Training-time range-aware stochastic degradation and context-guided attribute reconstruction. As the estimated sample-level range increases, point dropping and attribute masking probabilities increase linearly.

3.4. Context-Guided Attribute Reconstruction

Range-aware attribute masking creates an auxiliary recovery task for Doppler velocity and echo intensity. The target set of masked attributes is:

$$\mathcal{M}_a = \{(t, i) \mid M_i^t = 1, m_{\text{drop},i}^t = 0, m_{\text{attr},i}^t = 1\}. \quad (15)$$

For $(t, i) \in \mathcal{M}_a$, the spatial branch $\mathbf{s}_i^t$ is retained while the dynamic branch $\mathbf{u}_i^t$ is zeroed. The model must reconstruct the original dynamic attributes from the point location and visible frame context.

The visible points in frame t are:

$$\Omega_t^{\text{vis}} = \{j \mid M_j^t = 1, m_{\text{drop},j}^t = 0, m_{\text{attr},j}^t = 0\}. \quad (16)$$

Their perturbed tokens are max-pooled to obtain the frame-level context:

$$\mathbf{c}^t = \max_{j \in \Omega_t^{\text{vis}}} \tilde{\mathbf{h}}_j^t \in \mathbb{R}^C. \quad (17)$$

For each masked point, the decoder predicts the standardized Doppler and log-intensity attributes:

$$\hat{\mathbf{a}}_i^t = \Phi_{\text{dec}}([\mathbf{s}_i^t \parallel \mathbf{c}^t]) \in \mathbb{R}^2, \quad (18)$$

where Φ_{dec} is a two-layer MLP. The mean and standard deviation used for standardizing Doppler velocity and log-intensity are computed only from the training split of each protocol and then applied to the validation and test sets. We let the standardized ground-truth attributes be

$$\mathbf{a}_i^{t,*} = [v_i^{t,*}, \tilde{I}_i^{t,*}]. \quad (19)$$

The reconstruction loss is:

$$\mathcal{L}_{\text{attr}} = \frac{1}{|\mathcal{M}_a|} \sum_{(t,i) \in \mathcal{M}_a} \|\hat{\mathbf{a}}_i^t - \mathbf{a}_i^{t,*}\|_2^2. \quad (20)$$

This auxiliary objective encourages the representation to capture the relationship among spatial position, frame-level context, and radar physical attributes. The decoder is used only during training.

3.5. Frame Aggregation and Bidirectional Temporal Mamba Encoding

3.5.1. Frame Aggregation

After point token construction and training-time degradation, each frame contains a set of visible point tokens. Given the limited number of valid points in far-field frames and the instability of local point structures at long range, max pooling is adopted as an order-invariant frame aggregation operation. For frame t , the retained visible set is:

$$\Omega_t = \{i \mid M_i^t = 1, m_{\text{drop},i}^t = 0\}. \quad (21)$$

The frame-level representation is:

$$\mathbf{f}^t = \begin{cases} \max_{i \in \Omega_t} \tilde{\mathbf{h}}_i^t, & |\Omega_t| > 0, \\ \mathbf{f}_{\text{empty}}, & |\Omega_t| = 0, \end{cases} \quad (22)$$

where $\mathbf{f}_{\text{empty}} \in \mathbb{R}^C$ is a learnable empty-frame token. The entire gesture is represented as a frame-token sequence:

$$\mathbf{F} = [\mathbf{f}^1, \mathbf{f}^2, \dots, \mathbf{f}^T] \in \mathbb{R}^{T \times C}. \quad (23)$$

3.5.2. Bidirectional Mamba Temporal Encoder

Gestures contain ordered phases such as initiation, primary motion, and completion, and discriminative frames may appear only locally in the sequence. Since the task is offline sequence classification, the full gesture sequence is available at inference time. We therefore adopt a bidirectional Mamba encoder to use both past and future motion context.

We let $\mathbf{F}^{(0)} = \mathbf{F}$. For the l th bidirectional Mamba layer, the input is first normalized:

$$\bar{\mathbf{F}}^{(l-1)} = \text{LN}(\mathbf{F}^{(l-1)}). \quad (24)$$

Forward and backward Mamba encoders are then applied:

$$\mathbf{Z}_f^{(l)} = \text{Mamba}_f^{(l)}(\bar{\mathbf{F}}^{(l-1)}) \in \mathbb{R}^{T \times C}, \quad (25)$$

$$\mathbf{Z}_b^{(l)} = \text{Reverse}\left(\text{Mamba}_b^{(l)}(\text{Reverse}(\bar{\mathbf{F}}^{(l-1)}))\right) \in \mathbb{R}^{T \times C}. \quad (26)$$

The two directions are concatenated and fused:

$$\mathbf{Z}^{(l)} = W_o^{(l)}[\mathbf{Z}_f^{(l)} \parallel \mathbf{Z}_b^{(l)}] \in \mathbb{R}^{T \times C}, \quad (27)$$

where $W_o^{(l)} \in \mathbb{R}^{C \times 2C}$. The layer output is obtained through a residual connection:

$$\mathbf{F}^{(l)} = \mathbf{F}^{(l-1)} + \mathbf{Z}^{(l)}, \quad l = 1, 2, \dots, L. \quad (28)$$

After L layers, the final temporal representation is:

$$\mathbf{Z}^{\text{out}} = \mathbf{F}^{(L)} = [\mathbf{z}_1^{\text{out}}, \mathbf{z}_2^{\text{out}}, \dots, \mathbf{z}_T^{\text{out}}] \in \mathbb{R}^{T \times C}. \quad (29)$$

3.6. Classification Head and Training Objective

The sequence-level representation is obtained by temporal average pooling:

$$\mathbf{z}_{\text{seq}} = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t^{\text{out}} \in \mathbb{R}^C. \quad (30)$$

The classification head maps it to gesture logits:

$$\hat{y} = \Phi_{\text{cls}}(\mathbf{z}_{\text{seq}}) \in \mathbb{R}^K, \quad (31)$$

where $K = 5$ in our experiments. The classification loss is cross entropy:

$$\mathcal{L}_{\text{cls}} = - \sum_{k=1}^K y_k \log(\text{Softmax}(\hat{y})_k). \quad (32)$$

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{attr}}, \quad (33)$$

where λ controls the relative contribution of the auxiliary attribute reconstruction objective. Since the primary task is gesture classification, the reconstruction term is used as a regularizer rather than as an equally weighted objective. We set $\lambda = 0.1$ for all experiments based on validation-set model selection: this value keeps the Doppler/intensity reconstruction signal active while preventing the auxiliary mean-squared error from dominating the cross-entropy classification loss, especially in early training when reconstructed attributes are still inaccurate. The same value is used across all protocols and ablation experiments. The model is trained end-to-end with AdamW, an initial learning rate of 1×10^{-3} , cosine learning-rate scheduling, and 100 epochs.

4. Experiments

4.1. Dataset and Evaluation Protocols

We evaluate RPT-Mamba on the public mTransSee mmWave radar gesture dataset [16]. The dataset contains five gesture classes, denoted as CR, KO, PL, PS, and UP. It was collected from 32 subjects under different sensing distances and environmental conditions. After segmenting continuous CSV point cloud sequences based on missing-frame boundaries, we obtain 58,491 gesture samples. The sensing distance covers 13 positions from 1.2 m to 4.8 m at 0.3 m intervals.

Each radar point is represented as:

$$(\text{FrameID}, N_{\text{obj}}, x, y, z, v, s), \quad (34)$$

where x , y , and z are coordinates, v is Doppler radial velocity, and s is echo intensity. In the model input, each point is represented as:

$$p_{t,i} = [x_{t,i}, z_{t,i}, v_{t,i}, \log(1 + s_{t,i}), r_{t,i}, d_{\text{norm}}], \quad (35)$$

where $r_{t,i} = \sqrt{x_{t,i}^2 + z_{t,i}^2}$ and

$$d_{\text{norm}} = \text{clip}\left(\frac{\bar{r} - 1.2}{4.8 - 1.2}, 0, 1\right). \quad (36)$$

Each sample is formed by selecting every 10th raw radar frame (stride = 10) and retaining at most $T = 30$ frames; each frame contains at most $N = 32$ points. Frames with more than N points are randomly subsampled, while frames with fewer points are zero padded with a padding mask. Each sample is therefore represented as $X \in \mathbb{R}^{T \times N \times 6}$.

We use three evaluation protocols. First, the random split protocol evaluates in-distribution performance. Samples are randomly split into training, validation, and test sets with a ratio of 70%, 10%, and 20%, respectively (Table 1).

Table 1. Random split protocol on mTransSee. Samples are randomly divided into training, validation, and test sets for in-distribution evaluation.

Set	Samples	Ratio	Usage
Training	40,944	70%	Model parameter learning
Validation	5849	10%	Model selection and hyperparameter tuning
Test	11,698	20%	In-distribution performance evaluation

Second, we introduce a subject-independent evaluation protocol to assess cross-subject generalization and reduce the risk of subject-level data leakage. Unlike sample-wise random splitting, this protocol partitions the dataset at the subject level. All samples from the same subject are assigned exclusively to the training, validation, or test set, ensuring that the subjects in the three subsets are mutually disjoint. Specifically, we perform 5-fold subject-independent cross-validation over the 32 subjects. The subjects are divided into five non-overlapping folds, and each fold is used once as the held-out test subject group. In each round, the remaining subjects are further divided into training and validation subjects without subject overlap. The final performance is reported as the mean and standard deviation of Accuracy and Macro-F1 over the five subject folds.

Third, we design a near-to-far protocol to evaluate far-field generalization. Samples are divided into four distance buckets (Table 2). In this protocol, models are trained and validated only on Bucket 0 and Bucket 1, and tested on the unseen far-field Bucket 2 and Bucket 3 (Table 3).

Table 2. Distance bucket definition used in the near-to-far protocol. Buckets 0–1 are used for training/validation, and Buckets 2–3 are held out for far-field testing.

Bucket	Actual Distances	Description
Bucket 0	1.2, 1.5, 1.8 m	Near field
Bucket 1	2.1, 2.4, 2.7 m	Middle field
Bucket 2	3.0, 3.3, 3.6, 3.9 m	Far field
Bucket 3	4.2, 4.5, 4.8 m	Extreme far field

Table 3. Near-to-far cross-distance protocol. Models are trained and validated on 1.2–2.7 m samples and tested only on unseen 3.0–4.8 m far-field samples.

Set	Distances	Samples	Usage
Training	1.2–2.7 m	24,225	Model parameter learning
Validation	1.2–2.7 m	2692	Model selection and hyperparameter tuning
Test	3.0–4.8 m	31,574	Far-field generalization evaluation

We report Accuracy and Macro-F1 as the main metrics. Unless otherwise specified, methods compared under the same protocol use identical preprocessing, input settings, and training configurations. For the random split and near-to-far protocols, each experiment is repeated three times, and average results are reported. For the subject-independent protocol, the mean and standard deviation are reported across the five subject folds; the standard deviation reflects fold-to-fold variation rather than repeated-run variation.

4.2. Comparison with Existing Methods

4.2.1. Compared Methods

The comparison covers three types of baselines. mSeeNet and GesturePrint are directly related to the mTransSee/mmWave gesture setting [16,17]. Pantomime and STPM represent sparse radar point cloud recognition with PointNet++-LSTM and Mamba-style temporal modeling, respectively [11,13]. PointNet++-LSTM, P4Transformer, and Vanilla Mamba are included to examine whether generic point cloud sequence encoders are sufficient after the input is adapted to mTransSee [8,9,26,27].

4.2.2. Random Split Protocol

Under random splitting, RPT-Mamba reaches 92.09% accuracy and 92.04% Macro-F1 (Table 4). STPM is the closest baseline at 90.34% accuracy, while Vanilla Mamba reaches 88.52%. The gap between RPT-Mamba and Vanilla Mamba suggests that the performance improvement cannot be attributed solely to the use of a Mamba temporal module; distance-aware token construction and training-time degradation also contribute to the final performance.

Table 4. Performance comparison under the random split protocol. Accuracy and Macro-F1 are reported on the held-out test set using the same preprocessing and input configuration. Bold values indicate the best performance.

Method	Main Structure	Temporal Modeling	Citation	Accuracy (%)	Macro-F1 (%)
PointNet++-LSTM	PointNet++	LSTM	[8,9]	85.62	85.14
P4Transformer	Point Transformer	Transformer	[26]	89.17	88.63
mSeeNet	Radar sequence model	Temporal modeling	[16]	87.83	87.31
GesturePrint	Radar feature learning	Sequence modeling	[17]	83.45	82.91
Pantomime	PointNet++	LSTM	[11]	84.71	84.23
STPM	Point encoder	Mamba	[13]	90.34	89.87
Vanilla Mamba	Point encoder	Mamba	[27]	88.52	88.04
RPT-Mamba	RPT-Mamba	Bidirectional Mamba	Ours	92.09	92.04

4.2.3. Subject-Independent Protocol

To further examine whether the model generalizes to unseen subjects, we conduct five-fold subject-independent cross-validation. As shown in Table 5, the evaluation is performed at the subject level, and no subject appears in more than one subset within the same fold. This setting is stricter than sample-wise random splitting because the test samples are collected from subjects that are not observed during training or validation.

Table 5. Subject-independent 5-fold cross-validation results. The mean and standard deviation are computed across the five subject folds.

Method	Accuracy (%), Mean $\pm$ Std	Macro-F1 (%), Mean $\pm$ Std
P4Transformer	84.38 $\pm$ 1.92	83.91 $\pm$ 2.05
Vanilla Mamba	85.21 $\pm$ 1.74	84.68 $\pm$ 1.89
STPM	86.73 $\pm$ 1.56	86.18 $\pm$ 1.68
RPT-Mamba	88.42 $\pm$ 1.35	88.01 $\pm$ 1.47

Compared with the random split protocol, the subject-independent results show lower recognition accuracy for all methods, indicating that sample-wise splitting may lead to optimistic performance estimates when subject identities are shared across the training and test sets. Nevertheless, RPT-Mamba still achieves the best average performance among the compared methods, with 88.42% Accuracy and 88.01% Macro-F1. The improvement over STPM suggests that the proposed range-aware physical tokenization and degradation-aware training remain effective when the test subjects are unseen during training. In addition, the relatively small standard deviation across the five folds indicates that the proposed method provides stable cross-subject generalization.

4.2.4. Near-to-Far Protocol

Under the near-to-far protocol, all methods show lower accuracy than under random splitting (Table 6). This performance reduction is consistent with the fact that the test samples are drawn from 3.0–4.8 m, a distance range entirely excluded from training. RPT-Mamba reaches 85.34% accuracy and 84.77% Macro-F1, improving over STPM by 3.67 percentage points in accuracy and over Vanilla Mamba by 7.88 percentage points.

Table 6. Performance comparison under the near-to-far protocol. All methods are trained on 1.2–2.7 m samples and evaluated on unseen 3.0–4.8 m samples. Bold values indicate the best performance.

Method	Training Distances	Test Distances	Citation	Accuracy (%)	Macro-F1 (%)
PointNet++-LSTM	1.2–2.7 m	3.0–4.8 m	[8,9]	74.31	73.87
P4Transformer	1.2–2.7 m	3.0–4.8 m	[26]	78.94	78.42
mSeeNet	1.2–2.7 m	3.0–4.8 m	[16]	76.58	76.09
GesturePrint	1.2–2.7 m	3.0–4.8 m	[17]	71.23	70.68
Pantomime	1.2–2.7 m	3.0–4.8 m	[11]	72.85	72.31
STPM	1.2–2.7 m	3.0–4.8 m	[13]	81.67	81.14
Vanilla Mamba	1.2–2.7 m	3.0–4.8 m	[27]	77.46	76.93
RPT-Mamba	1.2–2.7 m	3.0–4.8 m	Ours	85.34	84.77

4.2.5. Far-Field Distance Bucket Analysis

Table 7 separates the near-to-far test set into Bucket 2 and Bucket 3. Bucket 3 is the more challenging subset because the sensing distance reaches 4.2–4.8 m. In this bucket, RPT-Mamba achieves 78.56% accuracy, compared with 73.24% for STPM and 69.73% for Vanilla Mamba. This performance margin exceeds that observed under the random split

protocol, which is consistent with the range-conditioned design being particularly effective when the point cloud is severely degraded.

Table 7. Performance comparison on far-field distance buckets. Bucket 2 covers 3.0–3.9 m and Bucket 3 covers 4.2–4.8 m. Bold values indicate the best performance.

Method	Bucket 2 Acc. (%)	Bucket 2 Macro-F1 (%)	Bucket 3 Acc. (%)	Bucket 3 Macro-F1 (%)
PointNet++-LSTM	80.34	79.83	66.51	66.02
P4Transformer	84.87	84.35	71.16	70.62
mSeeNet	82.61	82.07	68.42	67.89
GesturePrint	77.42	76.89	62.93	62.41
Pantomime	79.08	78.55	64.87	64.34
STPM	87.93	87.41	73.24	72.68
Vanilla Mamba	83.45	82.91	69.73	69.18
RPT-Mamba	91.62	91.65	78.56	77.66

4.3. Computational Complexity and Efficiency Analysis

To further evaluate the deployability of RPT-Mamba, we compare its computational complexity with representative sequence-modeling baselines. All methods are profiled under the same input setting, with an input tensor size of $1 \times 30 \times 32 \times 6$. The parameter count and floating-point operations (FLOPs) are computed for a single forward pass under the same implementation and profiling environment, and inference latency is measured on an NVIDIA GeForce RTX 5090 GPU.

As shown in Table 8, RPT-Mamba achieves a favorable balance between recognition performance and computational cost. Compared with STPM, the proposed method reduces the parameter count, FLOPs, and inference latency while providing better near-to-far recognition accuracy. Compared with P4Transformer, RPT-Mamba has fewer parameters and lower latency, although its FLOPs are moderately higher due to the range-aware physical tokenization and bidirectional temporal modeling. Vanilla Mamba has the lowest latency, but it does not explicitly model distance-induced point loss and physical-attribute degradation. Together with the near-to-far recognition results, the complexity comparison suggests that RPT-Mamba remains practical for radar gesture recognition scenarios where both far-field robustness and inference efficiency are required, which is also relevant to reliability-sensitive and resource-constrained hardware systems [47].

Table 8. Computational complexity and latency comparison under the same input setting. Parameters, FLOPs, and single-sample inference latency are measured for an input tensor of $1 \times 30 \times 32 \times 6$.

Method	Input Size	Params (M)	FLOPs (G)	Latency (ms)
P4Transformer	$1 \times 30 \times 32 \times 6$	0.803	1.351	0.749
Vanilla Mamba	$1 \times 30 \times 32 \times 6$	0.404	3.394	0.469
STPM	$1 \times 30 \times 32 \times 6$	0.806	3.421	1.044
RPT-Mamba	$1 \times 30 \times 32 \times 6$	0.630	1.791	0.675

4.4. Ablation Study and Module Analysis

All ablation experiments are conducted under the near-to-far protocol, which better reveals the effect of each component on far-field generalization.

4.4.1. Key Module Ablation

Table 9 reports key module ablations. The baseline uses the same point encoder and mean pooling without attribute reconstruction, range-aware degradation, or bidirectional Mamba. Attribute reconstruction raises accuracy from 50.37% to 73.84%, indicating that

Doppler/intensity recovery provides an effective physical constraint. The third variant applies range-aware point dropping and attribute masking, but removes the reconstruction loss and still uses mean pooling. Its 68.62% accuracy is therefore not directly comparable to random attribute masking under identical controls. Instead, it indicates that range-aware degradation alone does not outperform reconstruction-based learning without temporal modeling. Adding bidirectional Mamba raises accuracy to 85.34%, indicating that the degradation objective requires a temporal encoder capable of exploiting the remaining motion context.

Table 9. Key module ablation results under the near-to-far protocol. The table isolates the effects of attribute reconstruction, range-aware degradation, and bidirectional Mamba. Bold values indicate the best performance.

Variant	Auxiliary Strategy	Temporal Aggregation	Accuracy (%)	Macro-F1 (%)
Baseline	None	Mean pooling	50.37	42.63
+Attribute Reconstruction	Random masking + reconstruction	Mean pooling	73.84	72.16
+Degradation Only	Range-aware degradation, no reconstruction	Mean pooling	68.62	65.31
+Bidirectional Mamba	Range-aware degradation + reconstruction	Bidirectional Mamba	85.34	84.77

These ablation results indicate that RPT-Mamba is sensitive to drastic architectural simplifications because the proposed components address different aspects of far-field degradation. Attribute reconstruction provides a physical constraint for recovering unreliable Doppler and intensity cues, whereas range-aware degradation exposes the model to distance-dependent point loss and attribute corruption. However, degradation alone may remove useful observations without providing a recovery objective, which explains why the degradation-only variant does not achieve the best performance. Bidirectional Mamba further allows the model to exploit temporal context from the complete gesture sequence, which is particularly important when far-field frames contain sparse or missing detections. Therefore, the performance drop under simplified variants reflects the complementary roles of the proposed modules rather than an isolated dependence on a single component.

4.4.2. Masking and Degradation Strategies

Table 10 compares different masking and degradation strategies under the same full downstream setting: the context-guided reconstruction loss and bidirectional Mamba encoder are both enabled, and only the masking/degradation rule is changed. Random Attribute Masking reaches 79.53% accuracy. Frame-wise and Tube Masking improve this to 81.27% and 82.43%, respectively, by adding temporal structure to the missing pattern. Range-aware Degradation reaches 85.34%, indicating that matching the perturbation pattern to sensing distance is more effective than applying a distance-agnostic mask when the reconstruction and temporal modules are held fixed.

Table 10. Effect of different masking and degradation strategies under the near-to-far protocol. The reconstruction branch and bidirectional Mamba encoder are kept fixed. Bold values indicate the best performance.

Strategy	Accuracy (%)	Macro-F1 (%)
Random Attribute Masking	79.53	78.94
Frame-wise Masking	81.27	80.65
Tube Masking	82.43	81.82
Range-aware Degradation	85.34	84.77

4.4.3. Temporal Modeling Modules

Table 11 isolates the temporal module while keeping the point tokenizer, degradation strategy, and input features unchanged. Mean pooling reaches 68.62% accuracy, confirming that frame-level statistics are insufficient for far-field gestures. LSTM and Transformer improve the result to 73.47% and 79.23%. Vanilla Mamba reaches 82.61%, and bidirectional Mamba further reaches 85.34%, indicating that future context contributes to offline gesture sequences whose discriminative motion may occur near the middle or end of the sample.

Table 11. Effect of different temporal modeling modules under the near-to-far protocol. The point tokenizer, degradation strategy, and input features are kept unchanged. Bold values indicate the best performance.

Temporal Module	Accuracy (%)	Macro-F1 (%)
Mean Pooling	68.62	65.31
LSTM	73.47	72.81
Transformer	79.23	78.65
Vanilla Mamba	82.61	82.04
Bidirectional Mamba	85.34	84.77

4.4.4. Input Feature Ablation

Table 12 studies the contribution of physical input features. Coordinates alone achieve 61.43% accuracy. Adding Doppler raises accuracy to 72.57%, indicating the importance of radial motion. Echo intensity adds another 4.27 percentage points, and point-level range raises the result to 81.96%. The sample-level normalized range further improves the final accuracy to 85.34%, supporting the choice of including both local and global range cues.

Table 12. Input feature ablation results under the near-to-far protocol. Features are added progressively from spatial coordinates to the full range-aware physical token. Bold values indicate the best performance.

Input Feature Setting	Feature Dim.	Accuracy (%)	Macro-F1 (%)
Coordinates (x, z)	2	61.43	57.82
Coordinates + Doppler	3	72.57	71.24
Coordinates + Doppler + intensity	4	76.84	75.93
Coordinates + Doppler + intensity + point-level range	5	81.96	81.33
Full features (+sample-level normalized range)	6	85.34	84.77

4.5. Visualization Analysis

4.5.1. Qualitative Degradation and Reconstruction Visualization

Figure 6 provides qualitative visualizations of the proposed range-aware stochastic degradation and context-guided attribute reconstruction.

The original point clouds in panels (a–d) show representative samples from different distance ranges, while panels (e–h) present the corresponding degraded point clouds. As the distance increases, the degraded samples become progressively sparser, which is consistent with the distance-adaptive design of the degradation module. This process simulates the range-dependent sparsity and partial observation common in far-field mmWave radar sensing.

The third row of Figure 6 visualizes the point-wise reconstruction error of the masked dynamic attributes at the attribute-masked point locations. During training, these points retain their spatial coordinates, while their Doppler velocity and echo intensity are masked. The reconstruction head predicts the masked dynamic attributes from the retained spatial information and the visible frame context. Therefore, the reconstruction objective does not aim to recover dropped points or reconstruct complete point clouds during inference;

instead, it regularizes the encoder to learn context-aware and range-robust representations from incomplete physical attributes.

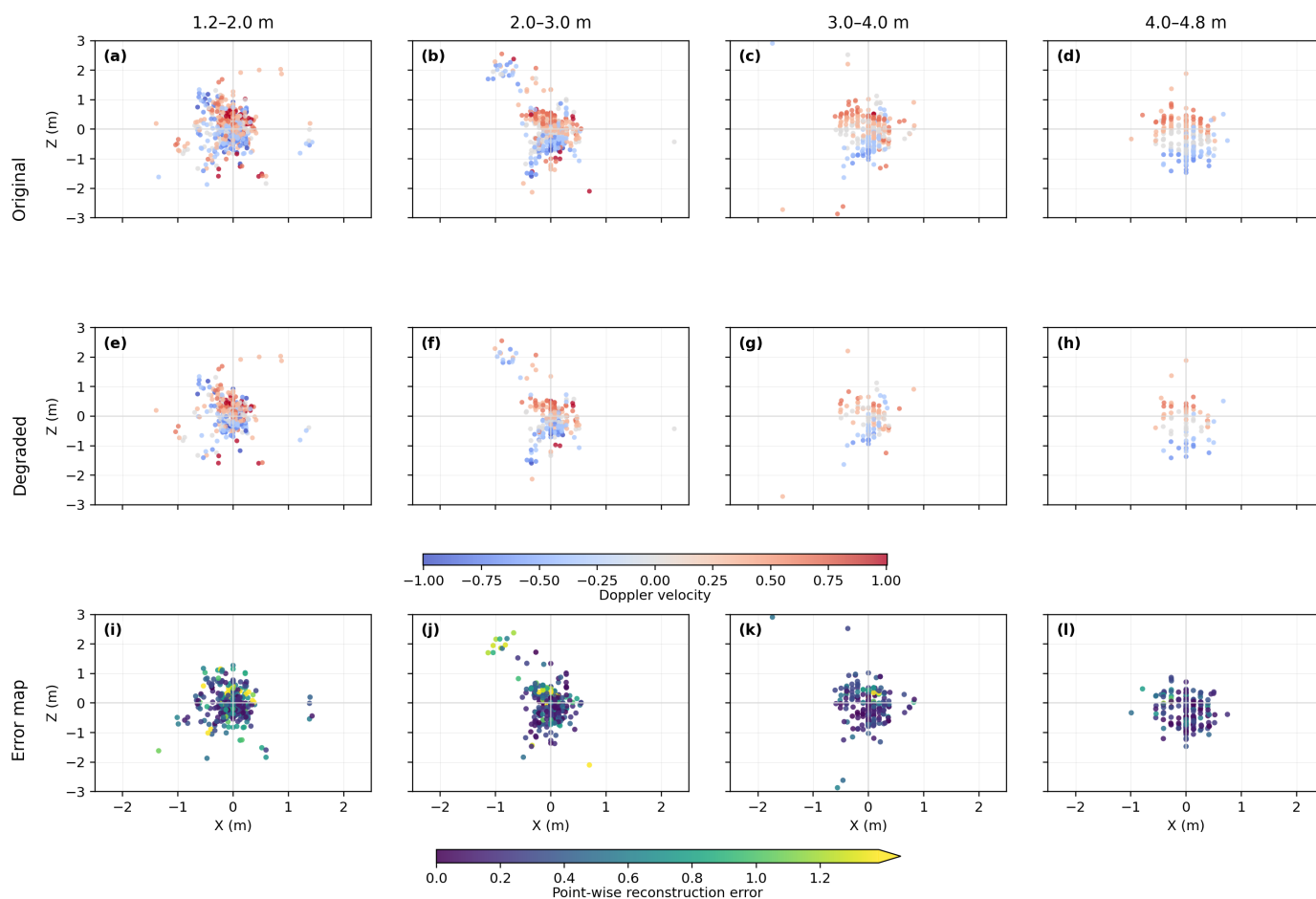


Figure 6. Qualitative visualization of range-aware stochastic degradation and context-guided attribute reconstruction. Columns correspond to four representative distance ranges from near-field to far-field: 1.2–2.0 m, 2.0–3.0 m, 3.0–4.0 m, and 4.0–4.8 m. Panels (a–d) show the original point clouds, where color indicates Doppler velocity. Panels (e–h) show the degraded point clouds after applying the proposed range-aware point dropping and attribute masking. Panels (i–l) visualize the point-wise reconstruction error of the masked dynamic attributes at the attribute-masked point locations. For these points, the spatial coordinates are retained, while the Doppler velocity and echo intensity are masked and predicted from the remaining visible frame context. The visualization illustrates that the proposed degradation strategy produces progressively sparser observations for farther distance ranges, while the reconstruction branch encourages the model to infer physical attributes from incomplete point-cloud context. The error maps are provided for qualitative interpretation rather than distance-wise performance ranking.

4.5.2. Confusion Matrix Analysis

Figure 7 compares the confusion matrices of the baseline and RPT-Mamba under the near-to-far protocol. The baseline shows substantial inter-class confusion under far-field conditions, especially among gestures with similar motion amplitudes and trajectories. RPT-Mamba produces more concentrated diagonal responses across all five classes, indicating that range-aware degradation, attribute reconstruction, and bidirectional temporal modeling reduce far-field misclassification.

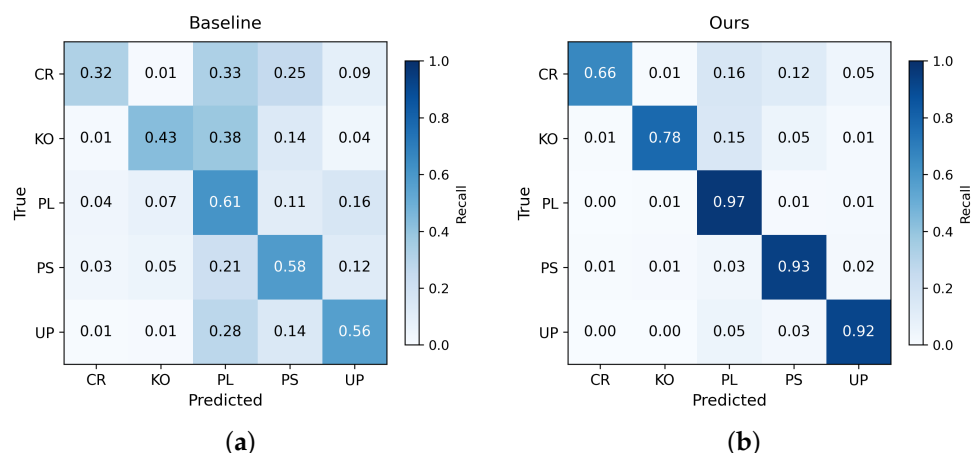


Figure 7. Confusion matrices of the baseline and RPT-Mamba under the near-to-far protocol. Gesture classes are ordered as CR, KO, PL, PS, and UP; rows denote ground-truth labels and columns denote predicted labels. (a) Baseline. (b) RPT-Mamba.

4.5.3. Distance Generalization Visualization

Figure 8 visualizes Accuracy and Macro-F1 over far-field test distances. Table 13 summarizes the bucket-level performance of key variants. All variants degrade from Bucket 2 to Bucket 3, but the degradation patterns differ. Attribute Reconstruction improves Bucket 3 accuracy from 39.18% to 66.85%. The Range-aware Degradation variant corresponds to the same intermediate setting as in Table 9: it introduces distance-conditioned dropping/masking but does not yet include the final bidirectional temporal encoder. It improves Bucket 2 accuracy to 83.67%, whereas Bucket 3 accuracy remains 48.26%. The full model reaches 91.62% on Bucket 2 and 78.56% on Bucket 3, suggesting that distance-aware perturbation and bidirectional temporal modeling should be used jointly to achieve substantial far-field gains.

Table 13. Recognition performance of variants on far-field distance buckets. B2 and B3 denote Bucket 2 and Bucket 3, respectively. Bold values indicate the best performance.

Variant	B2 Acc. (%)	B2 F1 (%)	B3 Acc. (%)	B3 F1 (%)
Baseline	58.42	47.25	39.18	35.63
+Attribute Reconstruction	79.42	77.53	66.85	65.18
+Degradation Only	83.67	78.92	48.26	46.57
+Full RPT-Mamba	91.62	91.65	78.56	77.66

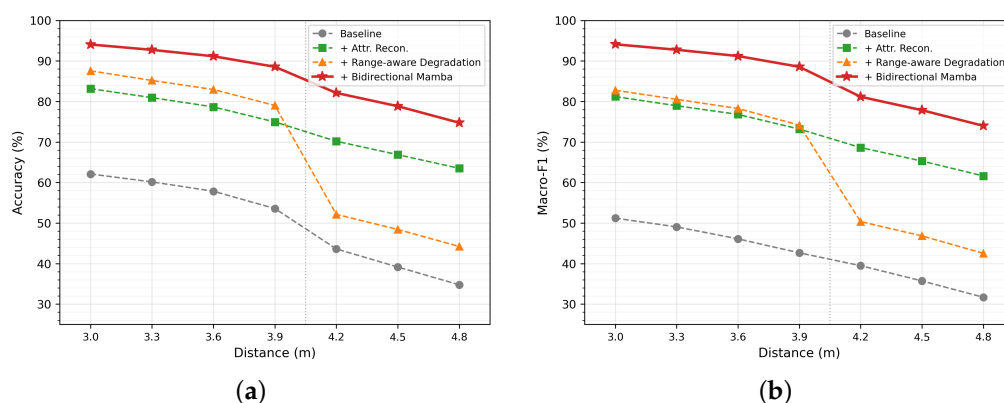


Figure 8. Recognition performance curves of different variants under far-field distances. The horizontal axis is the held-out test distance from 3.0 m to 4.8 m, and the vertical axes show Accuracy and Macro-F1. (a) Accuracy vs. distance. (b) Macro-F1 vs. distance.

5. Conclusions

RPT-Mamba targets the failure mode in which mmWave radar gesture point clouds become sparse and physically unreliable at longer sensing distances. It encodes each detection with spatial, Doppler, intensity, and range cues; uses range-conditioned point dropping and attribute masking to mimic distance-related degradation; reconstructs masked Doppler and intensity values as an auxiliary constraint; and applies bidirectional Mamba to the resulting frame-token sequence. On mTransSee, RPT-Mamba achieves 92.09% accuracy and 92.04% Macro-F1 under the random split protocol and maintains 88.42% accuracy under subject-independent 5-fold cross-validation. In the near-to-far setting, accuracy increases from 81.67% for STPM to 85.34%, and extreme far-field Bucket 3 accuracy increases from 73.24% to 78.56%. These results indicate that far-field performance benefits from explicit range-conditioned degradation modeling in addition to enhanced temporal encoding. Future work will consider broader deployment scenarios, additional radar configurations, and cross-device generalization.

Author Contributions: Conceptualization, Y.S. and P.P.; methodology, Y.S. and P.P.; software, Y.S.; validation, Y.S.; formal analysis, Y.S.; investigation, Y.S.; resources, P.P. and Z.W.; data curation, Y.S.; writing—original draft preparation, Y.S.; writing—review and editing, P.P. and Z.W.; visualization, Y.S.; supervision, P.P. and Z.W.; project administration, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experiments are based on the public mTransSee dataset.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hao, Z.; Sun, Z.; Li, F.; Wang, R.; Peng, J. Millimeter-wave gesture recognition using multi-feature fusion models in complex scenes. *Sci. Rep.* **2024**, *14*, 13758. [PubMed]
2. Bai, W.; Chen, S.; Ma, J.; Wang, Y.; Han, C. Gesture recognition with residual LSTM attention using millimeter-wave radar. *Sensors* **2025**, *25*, 469. [CrossRef] [PubMed]
3. Li, W.; Guo, Z.; Han, Z. Millimeter-wave radar point cloud gesture recognition based on multiscale feature extraction. *Electronics* **2025**, *14*, 371. [CrossRef]
4. Chen, H.; Wang, X.; Hao, Z.; Lu, Y.; Li, J.; Zhang, H.; Xi, B. RaGeoSense for smart home gesture recognition using sparse millimeter-wave radar point clouds. *Sci. Rep.* **2025**, *15*, 15267. [PubMed]
5. Han, Z.; Jiang, J.; Ding, X.; Meng, Q.; Xu, S.; He, L.; Wang, J. DenserRadar: A 4D millimeter-wave radar point cloud detector based on dense LiDAR point clouds. In Proceedings of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems, Edmonton, AB, Canada, 24–27 September 2024; pp. 930–936.
6. Gao, H.; Xu, X.; Zhu, T.; Dong, X.; Bao, Y.; Zhang, M.L. Radar-Mamba: 4D millimeter-wave point cloud enhancement via state space models. In Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, 27–31 October 2025; pp. 1774–1782.
7. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
8. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5105–5114.
9. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
11. Palipana, S.; Salami, D.; Leiva, L.A.; Sigg, S. Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 1–27.

12. Kang, L.; Li, Z.; Zhao, X.; Zhao, Z.; Braun, T. ST-PCT: Spatial-temporal point cloud Transformer for sensing activity based on mmWave. *IEEE Internet Things J.* **2024**, *11*, 10979–10991.
13. Chen, Y.; Guo, Z.; Zhang, H.; Xu, M. STPM: Spatial-Temporal Point Mamba for activity recognition using mmWave radar point clouds. In Proceedings of the 2025 IEEE International Conference on Multimedia and Expo, Nantes, France, 30 June–4 July 2025; pp. 1–6.
14. Zhu, R.; Zhou, J.; Chen, S.; Ding, H. Pseudo-trapezoidal fast factorized backprojection algorithm for near-field sparse MIMO array 3-D imaging. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, *62*, 1–16.
15. Yang, H.; Han, M.; Li, X.; Duan, D.; Li, T.; Xu, W. iRadar: Synthesizing millimeter-waves from wearable inertial inputs for human gesture sensing. *arXiv* **2024**, arXiv:2412.15980.
16. Liu, H.; Cui, K.; Hu, K.; Wang, Y.; Zhou, A.; Liu, L.; Ma, H. mTransSee: Enabling environment-independent mmWave sensing based gesture recognition via transfer learning. *Proc. Interact. Mob. Wearable Ubiquitous ACM Technol.* **2022**, *6*, 1–28.
17. Xu, L.; Wang, K.; Gu, C.; Guo, X.; He, S.; Chen, J. GesturePrint: Enabling user identification for mmWave-based gesture recognition systems. In Proceedings of the 2024 IEEE 44th International Conference on Distributed Computing Systems, Jersey City, NJ, USA, 23–26 July 2024; pp. 1074–1085.
18. Dang, X.; Ke, W.; Hao, Z.; Jin, P.; Deng, H.; Sheng, Y. mm-TPG: Traffic policemen gesture recognition based on millimeter-wave radar point cloud. *Sensors* **2023**, *23*, 6816. [[PubMed](#)]
19. Dang, X.; Jin, P.; Hao, Z.; Ke, W.; Deng, H.; Wang, L. Human movement recognition based on 3D point cloud spatiotemporal information from millimeter-wave radar. *Sensors* **2023**, *23*, 9430. [[CrossRef](#)] [[PubMed](#)]
20. Zhang, H.; Liu, K.; Zhang, Y.; Lin, J. TRANS-CNN-based gesture recognition for mmWave radar. *Sensors* **2024**, *24*, 1800. [[PubMed](#)]
21. Li, S.; Li, H. Regional-to-local point-voxel Transformer for large-scale indoor 3D point cloud semantic segmentation. *Remote Sens.* **2023**, *15*, 4832.
22. Gu, Z.; He, X.; Fang, G.; Xu, C.; Xia, F.; Jia, W. Millimeter wave radar-based human activity recognition for healthcare monitoring robot. *arXiv* **2024**, arXiv:2405.01882.
23. Yan, J.; Xu, C.; Liu, D. OG-PCL: Efficient sparse point cloud processing for human activity recognition. *arXiv* **2025**, arXiv:2511.08910.
24. Huan, S.; Wang, Z.; Wang, X.; Wu, L.; Yang, X.; Huang, H.; Dai, G.E. A lightweight hybrid vision Transformer network for radar-based human activity recognition. *Sci. Rep.* **2023**, *13*, 17996. [[CrossRef](#)] [[PubMed](#)]
25. Ma, C.; Liu, Z. A novel spatial-temporal network for gait recognition using millimeter-wave radar point cloud videos. *Electronics* **2023**, *12*, 4785.
26. Fan, H.; Yang, Y.; Kankanhalli, M. Point 4D Transformer networks for spatio-temporal modeling in point cloud videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14204–14213.
27. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
28. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417.
29. Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; Qiao, Y. VideoMamba: State space model for efficient video understanding. In Proceedings of the European Conference on Computer Vision; Springer: Cham, Switzerland, 2024; pp. 237–255.
30. Liang, D.; Zhou, X.; Xu, W.; Zhu, X.; Zou, Z.; Ye, X.; Tan, X.; Bai, X. PointMamba: A simple state space model for point cloud analysis. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 32653–32677. [[CrossRef](#)]
31. Han, X.; Tang, Y.; Wang, Z.; Li, X. Mamba3D: Enhancing local features for 3D point cloud analysis via state space model. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, Australia, 28 October–1 November 2024; pp. 4995–5004.
32. Zhang, T.; Yuan, H.; Qi, L.; Zhang, J.; Zhou, Q.; Ji, S.; Yan, S.; Li, X. Point Cloud Mamba: Point cloud learning via state space model. *arXiv* **2024**, arXiv:2403.00762.
33. Wang, T.; Wen, W.; Zhai, J.; Xu, K.; Luo, H. Serialized Point Mamba: A serialized point cloud Mamba segmentation model. *arXiv* **2024**, arXiv:2407.12319.
34. Liu, J.; Han, J.; Liu, L.; Aviles-Rivero, A.I.; Jiang, C.; Liu, Z.; Wang, H. MAMBA4D: Efficient long-sequence point cloud video understanding with disentangled spatial-temporal state space models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 17626–17636.
35. Qu, K.; Gao, P.; Dai, Q.; Ye, Z.; Ye, R.; Sun, Y. CloudMamba: Grouped selective state spaces for point cloud analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2026; Volume 40, pp. 8659–8667.
36. Wu, Y.; Fioranelli, F.; Gao, C. RadMamba: Efficient human activity recognition through radar-based micro-Doppler-oriented Mamba state-space model. *IEEE Trans. Radar Syst.* **2026**, *4*, 261–272.
37. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollar, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.

38. Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; Lu, J. Point-BERT: Pre-training 3D point cloud Transformers with masked point modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19313–19322.
39. Pang, Y.; Tay, E.H.F.; Yuan, L.; Chen, Z. Masked autoencoders for 3D point cloud self-supervised learning. *World Sci. Annu. Rev. Artif. Intell.* **2023**, *1*, 2440001.
40. Guo, Z.; Zhang, R.; Qiu, L.; Li, X.; Heng, P.A. Joint-MAE: 2D-3D joint masked autoencoders for 3D point cloud pre-training. *arXiv* **2023**, arXiv:2302.14007.
41. Chen, G.; Wang, M.; Yang, Y.; Yu, K.; Yuan, L.; Yue, Y. PointGPT: Auto-regressively generative pre-training from point clouds. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 29667–29679.
42. Zhang, R.; Xu, H.; Jian, S.; Tan, Y.; Zhou, H.; Xu, R. Angel or devil: Discriminating hard samples and anomaly contaminations for unsupervised time series anomaly detection. *Neural Netw.* **2026**, *197*, 108532. [[CrossRef](#)] [[PubMed](#)]
43. Yang, T.; Qian, Z.; Hang, N.; Liu, M. S-PINN: Stabilized physics-informed neural networks for alleviating barriers between multi-level co-optimization. *Comput. Methods Appl. Mech. Eng.* **2025**, *447*, 118348.
44. Zhang, L.; Wu, T. Solving SAT by restricted spiking neural P systems with temporal division rules. *Theor. Comput. Sci.* **2026**, *1078*, 116009. [[CrossRef](#)]
45. Lv, J.; Wang, C.; Xie, L. Adaptive distributed observer design for nonlinear multiagent systems. *Automatica* **2026**, *183*, 112625.
46. Lv, J.; Wang, C.; Xie, L. Distributed output-feedback leader-following consensus of nonlinear multiagent systems. *Automatica* **2025**, *177*, 112281.
47. Ni, L.; Wu, S.; Pu, J.; Xin, S.; Wang, A.; Chen, X.; Lai, Z.; Meng, Y. Highly reliable RRAM-based physical unclonable function with auto-write technique. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2026**, *34*, 2000–2004.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.