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

milliMamba: Specular-Aware Human Pose Estimation via Dual mmWave Radar with Multi-Frame Mamba Fusion

Niraj Prakash Kini[†], Shiau-Rung Tsai[†], Guan-Hsun Lin[†],
Wen-Hsiao Peng[†], Ching-Wen Ma[†], Jenq-Neng Hwang[‡]
[†]National Yang Ming Chiao Tung University, Taiwan, [‡]University of Washington, USA
{nirajnycu.ee06, mick20001108.cs12, abc900203abc.cs12}@nycu.edu.tw,
wpeng@cs.nycu.edu.tw, machingwen@nycu.edu.tw, hwang@uw.edu

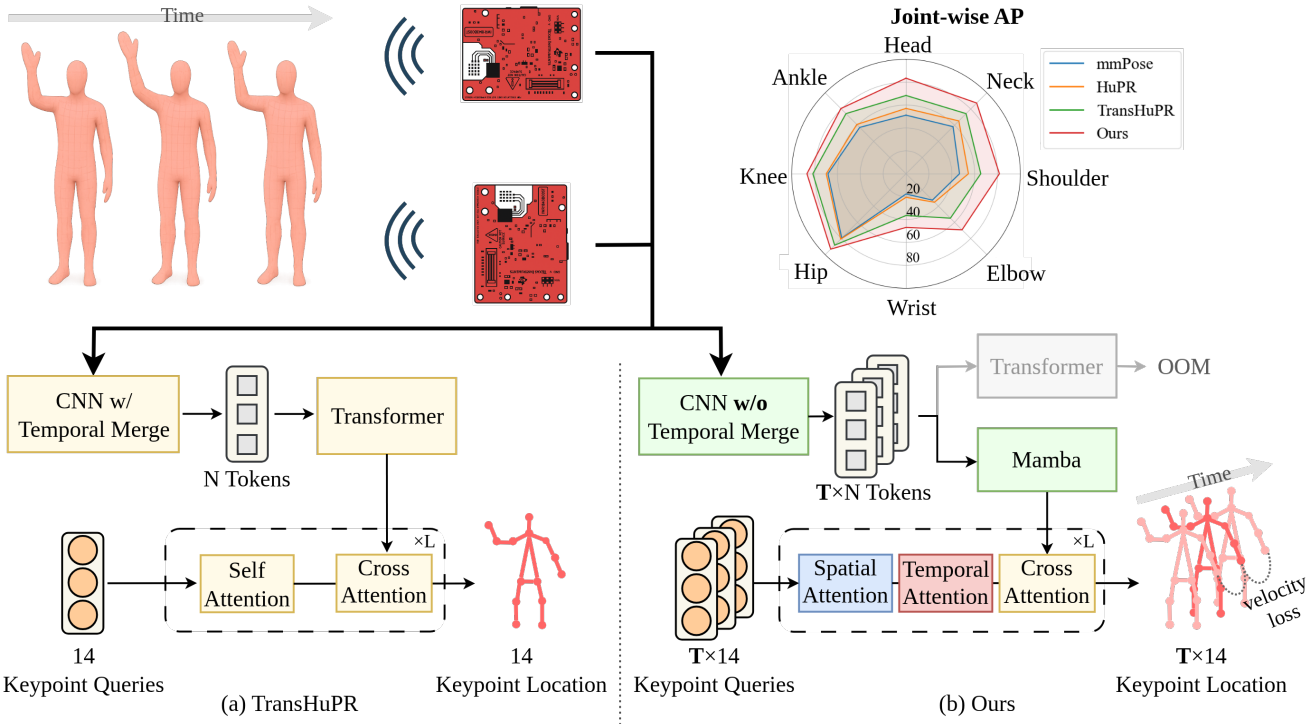


Figure 1. Our milliMamba performs spatio-temporal modeling across both the feature extraction and decoding stages, addressing a key limitation of TransHuPR [12], which models these dependencies only partially. This is made possible by milliMamba's ability to process a larger number of tokens with a comparable memory footprint, enabling richer temporal context and more accurate pose estimation.

Abstract

Millimeter-wave radar offers a privacy-preserving and lighting-invariant alternative to RGB sensors for Human Pose Estimation (HPE) task. However, the radar signals are often sparse due to specular reflection, making the extraction of robust features from radar signals highly challenging. To address this, we present milliMamba, a radar-based 2D human pose estimation framework that jointly models spatio-temporal dependencies across both the feature extraction and decoding stages. Specifically, given the high dimensionality of radar inputs, we adopt a Cross-View Fusion Mamba encoder to efficiently extract spatio-temporal features from longer sequences with linear complexity. A Spatio-Temporal-Cross Attention decoder then predicts joint coordinates across multiple frames. Together, this spatio-temporal modeling pipeline enables the model to leverage contextual cues from neighboring frames and

Table 1. Comparative analysis of mmWave radar-based HPE approaches.

Model		Input Representation	Prediction Strategy	Early Fuse Temporal	Multi-radar
RFMamba [35]	ICLR'25	Raw Signal	Many-to-one	✗	✗
TransHuPR [12]	BMVC'24	Point Cloud	Many-to-one	✓	✓
HuPR [13]	WACV'23	4D Heatmap	Many-to-one	✓	✓
CPFormer [2]	IEEE Sens.J.'25	3D Heatmap	Many-to-one	✓	✓
RF-Pose [36]	CVPR'18	2D Heatmap	Many-to-many	✗	WiFi
Ours (milliMamba)		3D Heatmap	Many-to-many	✗	✓

joints to infer missing joints caused by specular reflections. To reinforce motion smoothness, we incorporate a velocity loss alongside the standard keypoint loss during training. Experiments on the TransHuPR and HuPR datasets demonstrate that our method achieves significant performance improvements, exceeding the baselines by 11.0 AP and 14.6 AP, respectively, while maintaining reasonable complexity. Code: https://github.com/NYCU-MAPL/milliMamba

1. Introduction

Millimeter-wave (mmWave) radar-based HPE [4, 5, 7, 8, 11, 14, 20–22, 24, 29–32, 39] has emerged as a compelling alternative to RGB-based pose estimation methods, offering a unique balance of privacy preservation, environmental robustness, and deployment practicality. Despite these advantages, mmWave radar-based HPE remains technically challenging. Due to the specular nature of radar sensing, only body surfaces that reflect signals directly back to the receiver are captured, while others especially small or obliquely oriented joints are often missing. This leads to incomplete observations, making full-body pose reconstruction from single-frame inputs difficult. Additionally, weak reflections from extremities, fluctuations that disrupt temporal consistency, and high sensitivity to subject orientation and sensor placement further hinder estimation accuracy.

To address this issue, we introduce *milliMamba*, an mmWave radar-based 2D HPE framework that incorporates spatial and temporal modeling into both the encoding and decoding stages. This spatio-temporal modeling pipeline enables the model to leverage contextual cues from neighboring frames to infer missing joints caused by specular reflections. Beyond the modeling pipeline itself, we also revisit the signal preprocessing stage to make spatio-temporal modeling more tractable. Instead of constructing computationally expensive 4D heatmaps [25], we apply a commonly used 3D Fast Fourier Transform (FFT) to convert raw radar signals into 3D heatmaps. This implementation not only reduces preprocessing overhead but also mitigates the explosion of token counts, making the radar heatmap data easier to handle for downstream modeling.

To this end, our framework integrates two key components: a Mamba-based encoder and an attention-based decoder. The encoder is designed to efficiently process the large token volumes inherent in longer radar sequences, a challenge for prior Transformer-based approaches [3, 9–11, 16, 18, 19, 27, 33, 38] due to their quadratic complexity. Some methods [2, 12, 13] attempt to mitigate this by collapsing the temporal dimension early, but such early fusion can compromise the model's ability to recover missing joints caused by specular reflections.

We adopt a Cross-View Fusion Mamba encoder (CV-Mamba) that models longer-range spatio-temporal dependencies with linear complexity and effectively fuses dual-radar inputs across frames. To our knowledge, this is the first adaptation of Mamba for cross-view fusion in the radar domain, achieved by modifying the sequential scanning strategy from Vision Mamba [15, 40]. This design leverages Mamba's ability to capture dependencies over longer sequences efficiently, making it well-suited for the larger temporal context and multi-view correlations present in dual-radar setups.

The encoded features are passed to a Spatio-Temporal-Cross Attention (STCA) decoder, adapted from a DETR-style architecture [1] to support multi-frame pose prediction. While most prior radar-based HPE methods adopt a multi-frame to single-frame decoding scheme, STCA predicts poses for multiple frames simultaneously, as illustrated in Figure 1. STCA integrates both spatial attention and temporal attention, enabling it to model spatial relationships within each frame while capturing temporal dependencies across frames. This design offers two benefits: (1) richer supervision across time steps improves pose accuracy, and (2) the model better infers missing joints by leveraging contextual cues from neighboring frames and joints.

We evaluate our method on two mmWave radar-based 2D HPE datasets, TransHuPR [12] and HuPR [13]. Our method achieves significant performance improvements over the baselines, exceeding them by 11.0 AP and 14.6 AP, respectively, while maintaining an acceptable trade-off between accuracy and complexity.

This work makes the following contributions:

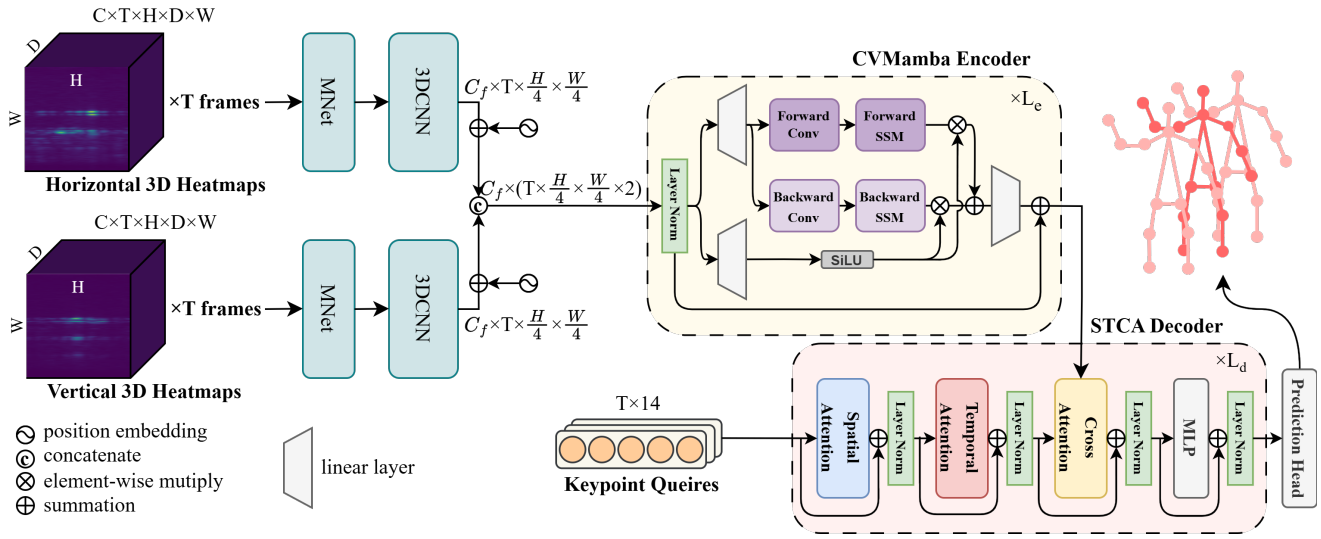


Figure 2. Overview of our milliMamba. The CVMamba encoder first extracts features from dual-view radar inputs. These features are then passed to the Multi-Pose STCA decoder, which progressively refines a set of keypoint queries to produce pose predictions.

- We adopt the CVMamba encoder to model longer-range spatio-temporal dependencies with linear complexity while performing cross-view fusion of dual-radar inputs.
- We propose the STCA decoder, a spatio-temporal cross-attention module that leverages multiple output frames to incorporate additional regularization enhancing pose accuracy and mitigating the effects of missing joints from specular reflections.
- Extensive experiments show that our proposed milliMamba serves as the new benchmark for the radar-based 2D HPE task on HuPR [13] and TransHuPR [12] datasets, with a significant leap in results over prior work.

2. Related Work

2.1. CNN- and Transformer-Based mmWave HPE

Deep learning approaches for mmWave-based human pose estimation (HPE) primarily leverage CNNs and Transformers. CNN-based methods [13, 22, 23] typically employ 2D or 3D convolutional blocks for feature extraction, followed by fully connected layers for pose prediction. These architectures are effective at capturing multiscale spatial and short-term temporal features, resulting in smooth pose estimations. However, they are often limited in their ability to fuse information from multiple radar sensors.

In contrast, Transformer-based methods [11, 12, 37] have demonstrated strong performance in fusing features from multiple radar sources. Their ability to model global dependencies across spatial and temporal dimensions leads to improved pose accuracy. Nevertheless, Transformers generally incur high computational costs, particularly in terms of memory usage and training time.

2.2. mmWave radar-based HPE

Table 1 compares representative mmWave radar-based HPE approaches across input representation, prediction strategy, temporal fusion, and multi-radar support. RF-Pose [36] is a WiFi-based HPE that relies on 2D heatmaps, while CPFormer [2] and HuPR [13] use higher-dimensional heatmaps, and TransHuPR [12] employs point cloud projections. RFMamba [35] using high resolution SFCW radar, uniquely processes raw signals with a many-to-one strategy, though without multi-radar support. Besides RF-Pose [36], most prior work adopts many-to-one prediction and relies on early temporal fusion. In contrast, our proposed milliMamba combines a many-to-many prediction strategy with 3D heatmap inputs and multi-radar support, striking a balance between spatial richness and efficiency. Its design inherently captures temporal dependencies without explicit early fusion, allowing superior robustness, efficiency, and accuracy over existing methods.

2.3. Mamba-Based mmWave HPE: RFMamba

While Transformers dominate the field, a recent effort has explored state space models (SSMs) for mmWave-based HPE. RFMamba [35] is, to date, the only method that applies SSMs to radar-based 3D HPE. Its architecture retains complex-valued representations throughout the network. While RFMamba presents a novel application of SSMs in this domain, its source code is not publicly available, which precludes direct comparison in our experiments.

In contrast, our method directly adopts a vision-domain Mamba architecture [6, 40] to process radar signals in the

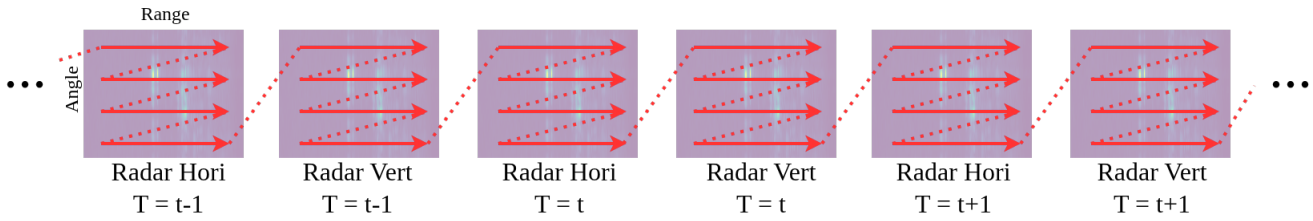


Figure 3. Our mamba scanning pattern. Only the forward direction is shown for clarity.

frequency domain. We treat the real and imaginary components of the complex-valued input as separate feature channels, allowing the model to operate entirely in the real domain. This design avoids complex-valued operations, simplifies implementation, and aligns more naturally with standard vision model architectures.

3. Proposed Method

Figures 2 illustrate the pipeline of our proposed framework, *milliMamba*, which transforms dual-view mmWave radar signals into temporally coherent 2D human poses. We adopt a sliding window strategy centered at the current frame, such that the input includes both past and future radar frames. The architecture consists of three sequential stages: **(i)** radar pre-processing via a simple 3D Fast Fourier Transform, **(ii)** a Cross-View Fusion Mamba encoder that fuses temporal dynamics and complementary spatial cues from horizontal and vertical views, and **(iii)** a Spatio-Temporal-Cross Attention decoder that predicts multi-frame pose sequences within the sliding window. Each component is detailed in the following sections.

3.1. Radar Pre-processing

Input Data Format. An FMCW radar produces complex-valued cubes $\mathbf{X} \in \mathbb{C}^{12\times128\times256}$, whose three dimensions correspond to 12 virtual-antenna pairs, 128 chirps, and 256 ADC samples. Two radars are mounted orthogonally to capture horizontal and vertical views, and we acquire T consecutive frames from both sensors in each sliding window [12, 13].

Clutter Removal and Chirp Sub-sampling. First, static clutter is removed by subtracting the mean across chirps. Then, the chirp dimension is uniformly subsampled to 8 chirps per frame to reduce computation while preserving Doppler resolution.

3D Fast Fourier Transform (FFT). Finally, we convert each radar cube into a 3D angle-doppler-range heatmap. A 1D FFT is first applied along the ADC-sample dimension, followed by another along the chirp dimension. To enhance angular resolution, the virtual-antenna dimension, which originally encodes azimuth and elevation, is zero-padded from 12 to 64 and then transformed by a third 1D FFT. For a 1D FFT, the transform is

$$\mathbf{Y}(m) = \sum_{n=0}^{N-1} \mathbf{X}(n)\, \exp\!\big(-j\,\tfrac{2\pi}{N}\,nm\big), \qquad (1)$$

where N is the input length and n, m are the indices of input and output. The resulting tensor $\mathbf{Y} \in \mathbb{C}^{H\times D\times W} = \mathbb{C}^{64\times8\times256}$ organizes the data along **angle** (H), **doppler** (D), and **range** (W).

Figure 4 illustrates the summary of the radar preprocessing pipeline. The 3D heatmap generation simplifies the process and is far more efficient, cutting memory usage by 11× and latency by 8.6× compared to the conventional 4D approach.

3.2. Cross-View Fusion Mamba Encoder (CV-Mamba)

The preprocessed radar data from Sec. 3.1 are stacked across T frames and split into real and imaginary components to form a two-channel tensor of shape $C\times T\times H\times D\times W$ with $C = 2$. Following HuPR [13], we process the horizontal (h) and vertical (v) views in two parallel branches. Each branch starts with an MNet [28] block that merges the doppler dimension, then passes through three residual 3D convolutions and two down-sampling layers, reducing the angle H and range W resolutions by $4\times$. This produces feature maps $\mathbf{F}_h, \mathbf{F}_v \in \mathbb{R}^{C_f\times T\times \frac{H}{4}\times\frac{W}{4}}$.

Cross-View Fusion. We first add separate learnable positional embeddings $\mathbf{P}_h$ and $\mathbf{P}_v$ to the horizontal and vertical feature maps, encoding angle, and range dimension. The two views are then concatenated to form the encoder input $\mathbf{F} = [\mathbf{F}_h; \mathbf{F}_v] \in \mathbb{R}^{C_f\times T\times \frac{H}{4}\times\frac{W}{4}\times2}$.

Scanning Order. We convert $\mathbf{F}$ into a 1D sequence by a zigzag scan across *range* → *angle* → *view* ($h \rightarrow v$) → *frame* shown in Figure 3. This sequence is processed by two independent SSM branches in the forward and backward directions to provide bidirectional context.

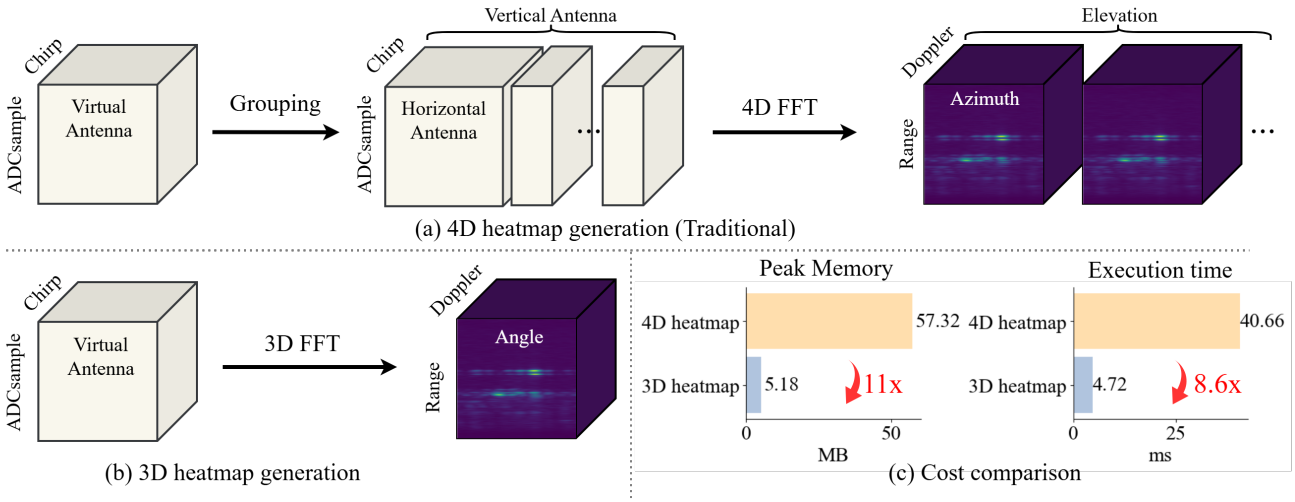


Figure 4. Comparison of heatmap generation. (a) The traditional 4D approach [25] applies separate FFTs for range, doppler, azimuth, and elevation after antenna grouping. (b) Our 3D pipeline performs a unified spatial FFT without grouping, yielding a compact representation. (c) Cost comparison between 4D and 3D heatmaps, showing 11× reduction in memory and 8.6× reduction in latency.

Vision Mamba Encoder. [40] The token sequence is processed by a stack of Vision Mamba layers. Each layer integrates a gating mechanism, two directional SSMs, and residual connections to efficiently capture long-range dependencies. Each SSM updates the hidden state h_{t+1} as:

$$\mathbf{h}_{t+1} = A\,\mathbf{h}_t + B\,u_t, \qquad \mathbf{y_t} = C\,\mathbf{h}_t + D\,u_t, \tag{2}$$

where u_t, y_t denote the input token and the output token at time step t, and A, B, C, D are layer-specific learnable parameters. Stacking L_e such Vision Mamba layers enables the model to extract global context across long sequences while preserving linear complexity. Through this design, our encoder captures features across multiple frames and multiple views, enabling efficient global context modeling.

3.3. Multi-Pose Spatio-Temporal-Cross Attention Decoder (STCA)

To predict human poses from the encoded radar representations, we propose a Multi-Pose Spatio-Temporal Cross-Attention (STCA) decoder that leverages learnable keypoint queries to jointly model spatial structure, temporal dynamics, and encoder–decoder interactions.

The decoder operates on a fixed set of $J \times T$ learnable keypoint queries $\{\mathbf{q}_{f,j}\}$ where $f \in \{0, \ldots, T-1\}$ indexes frames and $j \in \{0, \ldots, J-1\}$ enumerates joints. Each decoder layer includes *Spatio-Temporal Attention*, *Cross-Attention*, and a position-wise *MLP*. This layer is stacked L_d times, allowing queries to iteratively refine their representations.

Keypoint Query Embedding. Inspired by the object query strategy of DETR [1], each keypoint query represents a particular joint in a specific frame. All $J \times T$ keypoint queries are learnable and serve as the decoding tokens for pose prediction.

Spatio-Temporal Attention Module. Each decoder layer begins with two self-attention blocks that operate directly on the keypoint queries. Spatial Attention is applied first, operating within each frame to aggregate the J joints and capture inter-joint relationships. Temporal Attention attends to the same joint across all T frames, thereby enforcing motion consistency.

$$\mathbf{q}'_{f,\cdot} = \mathrm{SA}(\mathbf{q}_{f,\cdot}) = \mathrm{softmax}\big(Q_f K_f^\top / \sqrt{d}\big) V_f, \tag{3}$$

$$\mathbf{q}''_{\cdot,j} = \mathrm{TA}(\mathbf{q}'_{\cdot,j}) = \mathrm{softmax}\big(Q_j K_j^\top / \sqrt{d}\big) V_j. \tag{4}$$

Cross-Attention to Encoder Features. The updated keypoint queries attend to encoder features $\mathbf{F}'$ through standard cross-attention. This mechanism enables the model to utilize contextual information from all frames, improving its ability to estimate missing keypoints.

$$\widetilde{\mathbf{q}}_{f,j} = \mathrm{CrossAttn}\big(\mathbf{q}''_{f,j}, \mathbf{F}'\big). \tag{5}$$

Finally, each refined query $\widetilde{\mathbf{q}}_{f,j}$ is passed through a prediction head to produce 2D keypoint coordinates, generating a sequence of T pose estimates.

By iteratively refining keypoint queries through spatio-temporal self-attention and cross-attention with encoder features, the STCA decoder produces temporally consistent

2D pose sequences, effectively capturing both per-frame joint relations and motion continuity.

3.4. Training Objectives

We employ two losses. First, the Object Keypoint Similarity L_{oks} penalizes mismatches between predicted and ground-truth locations. Second, the velocity loss L_{vel} encourages temporal smoothness by minimizing the error between predicted and ground-truth joint velocities [34]. We define the predicted velocity of joint j at frame f as the difference between its positions at consecutive frames: $\widehat{\mathbf{v}}_{f,j} = \widehat{\mathbf{p}}_{f+1,j} - \widehat{\mathbf{p}}_{f,j}$ and $\mathbf{v}_{f,j}$ is computed from ground-truth keypoint positions. The velocity loss is computed as:

$$L_{\text{vel}} = \frac{1}{(T-1)\,J} \sum_{f=1}^{T-1} \sum_{j=1}^{J} \left\| \widehat{\mathbf{v}}_{f,j} - \mathbf{v}_{f,j} \right\|_2^2. \tag{6}$$

The overall training objective is $L = L_{\text{oks}} + \lambda_{\text{vel}}\, L_{\text{vel}}$, where λ_{vel} balances between pose accuracy and temporal consistency. At inference time, only the central frame prediction within each window is retained, while ensuring that the output reflects the most accurate estimate.

4. Experimental Results

4.1. Implementation Details

Our main method follows the default configuration described below. The model takes input from two radar sensors, each capturing a sequence of $T = 9$ frames. It outputs 9 consecutive pose predictions, from which only the center pose is used during inference. The model is trained using the Adam optimizer with a learning rate of 0.00005, a batch size of 8, and a weight decay of 0.0001. The velocity supervision term in the loss function is weighted by $\lambda_{\text{vel}} = 0.05$. All experiments are conducted on a single NVIDIA Tesla V100 GPU.

4.2. Datasets and Evaluation Metrics

We evaluate our method on two benchmark mmWave radar-based 2D HPE datasets.

TransHuPR Dataset [12] comprises 440 sequences, totaling over 7 hours of video recorded from 22 subjects. It consists exclusively of fast and dynamic actions, presenting a considerable challenge due to the high motion complexity and diversity. We follow the original data split protocol: 352 sequences are used for training, 44 for validation, and 44 for testing.

HuPR Dataset [13] contains 235 sequences, amounting to approximately 4 hours of video from 6 subjects. In contrast to TransHuPR, the majority of actions in this dataset are relatively static. We adopt the same split protocol: 193 sequences for training, 21 for validation, and 21 for testing.

For evaluation, we use Average Precision (AP) based on Object Keypoint Similarity (OKS) [17]. The overall AP is computed by averaging OKS over 10 thresholds uniformly spaced from 0.50 to 0.95. We also report AP^{50} and AP^{75}, which correspond to OKS thresholds of 0.50 and 0.75, representing loose and strict matching criteria, respectively.

4.3. Performance Comparison

Table 2 presents a comparison between our method and existing radar-based HPE approaches on the TransHuPR dataset. Our method consistently outperforms all baselines across all AP metrics, including both overall and joint-wise AP, while also maintaining a relatively compact model size. In particular, compared to the baseline *TransHuPR* [12], our approach achieves a substantial improvement of 11.0 AP. Notably, for the most challenging joint-wrist, which is both fast-moving and often affected by specular reflection, our model achieves an AP of 46.9, demonstrating its robustness in inferring joints with high uncertainty. Our model consistently improves accuracy across all keypoints, indicating reliable performance over the entire body structure. Figure 5 presents qualitative results on the TransHuPR dataset. Additional examples are in the supplementary material.

Table 3 shows a similar trend on the HuPR dataset. Our method achieves up to 84.0 AP, indicating high prediction accuracy on relatively static actions. Compared to the baseline *HuPR* [13] which requires 68.6 GMACs and 35.5 M parameters, our method achieves higher accuracy with only 34.4 GMACs and 4.0 M parameters, highlighting the computational efficiency of our design. All baseline methods adopt a single-frame output strategy, whereas our milliMamba employs a multi-frame output strategy.

While *TransHuPR* model exhibits lower MACs, our method offers significantly higher accuracy with only a moderate increase in compute. This demonstrates a favorable balance between efficiency and performance.

4.4. Ablation Studies

We conduct a series of ablation studies on the TransHuPR dataset to evaluate the impact of key design choices. Unless otherwise specified, all experiments are performed using a simplified variant of our model that employs only a single vertical radar. In this setting, the second CNN branch is removed, while the rest of the architecture remains unchanged. In addition, the number of default input frames $T = 9$.

Input Representation. As shown in Table 4, we compare different radar pre-processing strategies. The *density map*, a 2D projection derived from the 3D point cloud representation [12], yields the lowest performance, underscoring the limitations of compressing 3D point clouds into 2D representations. Heatmap-based representation performs better, and the *3D FFT*-based heatmap, which omits elevation padding and one FFT dimension used in 4D FFT,

Table 2. Comparison of model performance and complexity across methods on the TransHuPR dataset [12]. The complexity excludes radar signal preprocessing.

Method	Complexity			Joint-wise AP								Overall AP		
	MACs	Params	Mem	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	AP	AP^{50}	AP^{75}
mmPose [23]	**85.5 M**	15.0 M	**67.2 MB**	51.2	58.2	46.8	32.6	17.7	79.5	68.3	57.3	48.4	88.4	47.4
HuPR [13]	68.6 G	35.5 M	339.7 MB	57.1	65.3	54.6	35.2	20.6	80.8	69.8	60.9	51.5	89.5	53.7
TransHuPR [12]	5.8 G	5.3 M	230.8 MB	68.4	74.3	65.4	54.9	36.5	88.3	81.5	74.3	67.5	96.9	76.7
Ours	34.4 G	**4.0 M**	224.1 MB	**83.5**	**87.4**	**81.7**	**69.3**	**46.9**	**93.2**	**86.7**	**80.6**	**78.5**	**98.7**	**89.3**

Table 3. Comparison of model performance and complexity across methods on the HuPR dataset [13]. The complexity excludes radar signal preprocessing.

Method	Complexity			Joint-wise AP								Overall AP		
	MACs	Params	Mem	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	AP	AP^{50}	AP^{75}
mmPose [23]	**85.5 M**	15.0 M	**67.2 MB**	56.1	60.9	40.6	24.9	14.2	63.2	58.6	56.1	41.4	79.4	38.3
HuPR [13]	68.6 G	35.5 M	339.7 MB	77.5	81.9	70.3	45.5	22.3	88.1	82.2	73.1	63.4	97.0	74.0
TransHuPR [12]	5.8 G	5.3 M	230.8 MB	77.1	78.6	63.2	55.6	44.9	84.5	83.6	80.0	69.4	95.1	79.9
Ours	34.4 G	**4.0 M**	224.1 MB	**90.0**	**91.8**	**83.2**	**75.2**	**59.5**	**94.3**	**93.6**	**89.3**	**84.0**	**98.5**	**94.9**

Table 4. Ablation study on the impact of different input representations.

Input Representation	AP	AP^{50}	AP^{75}
density map	58.5	92.5	62.7
4D FFT	72.0	97.3	81.8
3D FFT	**74.5**	**98.5**	**84.7**

Table 5. Ablation study on the multi-pose output mechanism.

Prediction Strategy	AP	AP^{50}	AP^{75}
Many-to-one	70.4	97.0	81.0
Many-to-many	**74.5**	**98.5**	**84.7**

achieves comparable accuracy. Moreover, as illustrated in Figure 4(c), 3D FFT incurs significantly lower preprocessing cost than 4D FFT, thereby validating our design choice.

Multi-Pose Output Mechanism. Table 5 compares our model, *Many-to-many*, with a simplified variant, *Many-to-one*, where the decoder is replaced by a vanilla Transformer [26] that receives only center-pose keypoint queries and predicts a single pose. Our Multi-Pose STCA Decoder achieves a 4.1 AP improvement in overall accuracy. Although our method predicts only the center pose at inference time, the prediction is guided by joint features from different time steps, enabling the model to infer missing or weakly reflected joints using rich spatial-temporal context. More details in supplementary materials.

Effect of Input Sequence Length. Table 6 presents the impact of varying the number of input frames T on pose estimation accuracy. The results indicate that increasing T consistently improves performance, particularly for joints affected by rapid motion or frequently missing due to specular reflection, such as the wrist and elbow. Considering the trade-off between accuracy and computational cost, we adopt $T = 9$ as the default setting.

Number of Radars. Table 7 explores different radar configurations. Surprisingly, even using a single vertical radar achieves competitive performance, showcasing the practicality of single-radar setups for real-world deployment. While single-radar systems are simpler and more cost-effective, the dual-radar configuration offers additional gains by compensating for the limited elevation resolution inherent to mmWave radar sensors.

Transformer vs Mamba. Table 8 compares the encoder performance of Transformer and Mamba. Due to the higher memory demands of the Transformer, we were only able to conduct experiments with $T = 3$ frames to avoid out-of-memory issues. The results show that Mamba achieves 1.5 AP higher than the Transformer, indicating comparable accuracy. More importantly, given limited memory resources, Mamba offers a practical solution that scales effectively to longer input sequences. The Transformer requires 14.9 GMACs at $T = 3$, computationally comparable to Mamba at $T = 7$ and $T = 9$ where our model achieves much higher AP, highlighting Mamba's superior scalability at similar computational cost.

5. Conclusion

We propose *milliMamba*, a dual-radar architecture for human pose estimation that tackles sparse reflections and high-dimensional radar signals. By combining efficient input processing with multi-frame sequence modeling, milliMamba captures rich spatio-temporal context for ro-

Table 6. Impact of input sequence length (T) on pose estimation performance. We investigate the effect of varying T to understand how temporal context contributes to accuracy.

	Complexity			Joint-wise AP								Overall AP		
T	MACs	Params	Mem	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	AP	AP^{50}	AP^{75}
3	5.6 G	3.2 M	44.7 MB	73.4	78.6	69.5	53.9	32.6	88.0	79.6	71.9	66.9	95.7	75.0
5	9.4 G	3.2 M	62.3 MB	78.3	81.9	73.3	58.3	35.8	89.5	82.1	75.0	70.3	97.1	79.5
7	13.3 G	3.2 M	86.6 MB	79.6	84.0	75.4	61.5	39.3	91.0	84.0	77.7	72.9	97.3	83.0
9	17.3 G	3.2 M	121.2 MB	80.6	84.6	77.7	63.7	41.0	91.5	84.4	78.6	74.5	98.5	84.7
11	21.4 G	3.2 M	163.5 MB	79.6	83.9	78.2	64.9	42.0	91.9	85.1	78.8	75.2	98.4	86.2
13	25.6 G	3.2 M	212.1 MB	81.2	85.0	78.8	65.6	44.2	91.8	85.5	79.1	75.8	98.5	85.9
15	29.9 G	3.2 M	269.3 MB	81.7	85.9	80.4	66.8	45.1	92.6	86.4	80.8	77.1	98.7	87.9



Figure 5. Qualitative results on TransHuPR [12] dataset.

Table 7. Effect of radar configuration on pose estimation performance. We compare three setups: horizontal-only (Hori), vertical-only (Vert), and dual-radar (Hori+Vert).

Radar Used	**Complexity**			**Overall AP**		
	MACs	Params	Mem	AP	AP^{50}	AP^{75}
Hori	17.3 G	3.2 M	121.2 MB	67.3	95.8	75.0
Vert	17.3 G	3.2 M	121.2 MB	74.5	98.5	84.7
Hori+Vert	34.4 G	4.0 M	224.1 MB	78.5	98.7	89.3

Table 8. Comparison of Transformer and Mamba encoders with 3-frame radar inputs. Transformer runs out-of-memory on our hardware when trained with longer sequences.

Encoder	**Complexity**			**Overall AP**		
	MACs	Params	Mem	AP	AP^{50}	AP^{75}
Transformer	14.9 G	3.9 M	610.3 MB	65.4	95.5	73.5
Mamba	5.6 G	3.2 M	44.7 MB	66.9	95.7	75.0

bust pose prediction. Experiments on two benchmarks show state-of-the-art performance with competitive efficiency. Future work will explore multi-person and cross-environment scenarios while further reducing computational cost.

6. Acknowledgement

This work is supported by National Science and Technology Council (NSTC), Taiwan, under Grants 113-2634-F-A49-007-, 112-2221-E-A49-092-MY3, and 114-2221-E-A49-035-MY3. We thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 5

[2] Lin Chen and Guoli Wang. Cpformer: End-to-end multi-person human pose estimation from raw radar cubes with transformers. *IEEE Sensors Journal*, 2025. 2, 3

[3] Lin Chen, Xuemei Guo, and Guoli Wang. Mptformer: Towards robust arm gesture pose tracking using dual-view radar system. *IEEE Sensors Journal*, 23, 2023. 2

[4] Jaeho Choi, Soheil Hor, Shubo Yang, and Amin Arbabian. Mvdoppler-pose: Multi-modal multi-view mmwave sensing for long-distance self-occluded human walking pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27750–27759, 2025. 2

[5] Lukas Engel, Jonas Mueller, Eduardo Javier Feria Rendon, Eva Dorschky, Daniel Krauss, Ingrid Ullmann, Bjoern M Eskofier, and Martin Vossiek. Advanced millimeter wave radar-based human pose estimation enabled by a deep learning neural network trained with optical motion capture ground truth data. *IEEE Journal of Microwaves*, 2025. 2

[6] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 3

[7] Yuan-Hao Ho, Jen-Hao Cheng, Sheng Yao Kuan, Zhongyu Jiang, Wenhao Chai, Hsiang-Wei Huang, Chih-Lung Lin, and Jenq-Neng Hwang. Rt-pose: A 4d radar tensor-based 3d human pose estimation and localization benchmark. In *European Conference on Computer Vision*, pages 107–125. Springer, 2024. 2

[8] Teng Huang, Han Ding, Wenxin Sun, Cui Zhao, Ge Wang, Fei Wang, Kun Zhao, Zhi Wang, and Wei Xi. One snapshot is all you need: A generalized method for mmwave signal generation. In *IEEE INFOCOM 2025-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2025. 2

[9] Yingying Jiao, Zhigang Wang, Zhenguang Liu, Shaojing Fan, Sifan Wu, Zheqi Wu, and Zhuoyue Xu. Optimizing human pose estimation through focused human and joint regions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4102–4110, 2025. 2

[10] Can Jin, Xiangzhu Meng, Xuanheng Li, Jie Wang, Miao Pan, and Yuguang Fang. Rodar: Robust gesture recognition based on mmwave radar under human activity interference. *IEEE Transactions on Mobile Computing*, 23(12):11735–11749, 2024.

[11] Gon Woo Kim, Sang Won Lee, Ha Young Son, and Kae Won Choi. A study on 3d human pose estimation using through-wall ir-uwb radar and transformer. *IEEE Access*, 11:15082–15095, 2023. 2, 3

[12] Niraj Prakash Kini, Ruey-Horng Shiue, ryan chandra, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. Transhupr: Cross-view fusion transformer for human pose estimation using mmwave radar. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024. 1, 2, 3, 4, 6, 7, 8

[13] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5715–5724, 2023. 2, 3, 4, 6, 7

[14] Guangzheng Li, Ze Zhang, Hanmei Yang, Jin Pan, Dayin Chen, and Jin Zhang. Capturing human pose using mmwave radar. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2020. 2

[15] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pages 237–255. Springer, 2024. 2

[16] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 11313–11322, 2021. 2

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 6

[18] Jiajie Liu, Mengyuan Liu, Hong Liu, and Wenhao Li. Tcpformer: Learning temporal correlation with implicit pose proxy for 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5478–5486, 2025. 2

[19] Yizhe Lv, Tingting Zhang, Yunpeng Song, Han Ding, Jinsong Han, and Fei Wang. Egohand: Ego-centric hand pose estimation and gesture recognition with head-mounted millimeter-wave radar and imus. *arXiv preprint arXiv:2501.13805*, 2025. 2

[20] Jonas Leo Mueller, Lukas Engel, Eva Dorschky, Daniel Krauss, Ingrid Ullmann, Martin Vossiek, and Bjoern M Eskofier. Radproposer: A framework for human pose estimation with uncertainty quantification from raw radar data. *arXiv preprint arXiv:2508.03578*, 2025. 2

[21] M Mahbubur Rahman, Ryoma Yataka, Sorachi Kato, Pu Wang, Peizhao Li, Adriano Cardace, and Petros Boufounos. Mmvr: Millimeter-wave multi-view radar dataset and benchmark for indoor perception. In *European Conference on Computer Vision*, pages 306–322. Springer, 2024.

[22] Arindam Sengupta and Siyang Cao. mmpose-nlp: A natural language processing approach to precise skeletal pose esti-

mation using mmwave radars. *IEEE Transactions on Neural Networks and Learning Systems*, page 267–281, 2022. 2, 3

[23] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, 20 (17):10032–10044, 2020. 3, 7

[24] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani 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. 2

[25] Texas Instruments. IWR1843BOOST Evaluation Module. https://www.ti.com/tool/IWR1843BOOST. 2, 5

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 7

[27] Haonan Wang, Jie Liu, Jie Tang, Gangshan Wu, Bo Xu, Yanbing Chou, and Yong Wang. Gtpt: Group-based token pruning transformer for efficient human pose estimation. In *European Conference on Computer Vision*, pages 213–230. Springer, 2024. 2

[28] Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: Radar object detection using cross-modal supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 504–513, 2021. 4

[29] Yingxiao Wu, Zhongmin Jiang, Haocheng Ni, Changlin Mao, Zhiyuan Zhou, Wenxiang Wang, and Jianping Han. mmhpe: Robust multi-scale 3d human pose estimation using a single mmwave radar. *IEEE Internet of Things Journal*, 2024. 2

[30] Chunyang Xie, Dongheng Zhang, Zhi Wu, Cong Yu, Yang Hu, and Yan Chen. Rpm 2.0: Rf-based pose machines for multi-person 3d pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):490–503, 2023.

[31] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2021.

[32] Zhihang Yan, Hongbo Ni, Xiaoguang Yu, and Weiqi Tan. Indoor 3d human pose estimation using single millimeter-wave radar and conditional generative adversarial network. *IEEE Sensors Journal*, 2024. 2

[33] Ryoma Yataka, Adriano Cardace, Perry Wang, Petros Boufounos, and Ryuhei Takahashi. Retr: Multi-view radar detection transformer for indoor perception. *Advances in Neural Information Processing Systems*, 37:19839–19869, 2024. 2

[34] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*, pages 625–642. Springer, 2022. 6

[35] Rui Zhang, Ruixu Geng, Yadong Li, Ruiyuan Song, Hanqin Gong, Dongheng Zhang, Yang Hu, and Yan Chen. Rf-mamba: Frequency-aware state space model for rf-based human-centric perception. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3

[36] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7356–7365, 2018. 2, 3

[37] Zhijie Zheng, Diankun Zhang, Xiao Liang, Xiaojun Liu, and Guangyou Fang. Radarformer: End-to-end human perception with through-wall radar and transformers. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2023. 3

[38] Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet of Things Journal*, 10(16):14128–14136, 2023. 2

[39] Bing Zhu, Zixin He, Weiyi Xiong, Guanhua Ding, Tao Huang, and Wei Xiang. Probradarm3f: mmwave radar-based human skeletal pose estimation with probability map guided multi-format feature fusion. *IEEE Transactions on Aerospace and Electronic Systems*, 2025. 2

[40] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3, 5