MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Heatmap-to-SMPL Multi-View Radar Transformer for Multi-Person 3D Pose Estimation

Kato, Sorachi; Wang, Pu; Fujihashi, Takuya; Markham, Andrew

TR2026-040 March 28, 2026

Abstract

Radar-based 3D human pose estimation can be achieved using either sparse radar point clouds or, more recently, high-resolution multi-view radar heatmaps. Point-cloud approaches typically lever- age strong body-shape priors, e.g., Skinned Multi-Person Linear Model (SMPL), but depend on point-based backbones and potentially temporal aggregation to compensate for weak features; heatmap approaches preserve richer, reflectivity-level radar features, yet usually regress only 3D keypoints, ignoring body-shape priors. In this paper, by retaining heatmap fidelity and simultaneously exploiting shape priors, we propose RHAMP: a Radar HeAtmap- to-SMPL Pose transformer for 3D human pose estimation. Specifi- cally, each radar view is encoded by the backbone network, and a set of person queries cross-attends to the multi-view radar features to produce per-instance SMPL parameters in a single end-to-end stage. Experiments on the public HIBER dataset confirm the effectiveness of the proposed approach over a list of baselines.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

HEATMAP-TO-SMPL MULTI-VIEW RADAR TRANSFORMER FOR MULTI-PERSON 3D POSE ESTIMATION

Sorachi Kato[1,2] *Pu (Perry) Wang*[1,3], *Takuya Fujihashi*[2], *Andrew Markham*[3]

[1]Mitsubishi Electric Research Laboratories (MERL), [2]University of Osaka, [3]University of Oxford

ABSTRACT

Radar-based 3D human pose estimation can be achieved using either sparse radar point clouds or, more recently, high-resolution multi-view radar heatmaps. Point-cloud approaches typically leverage strong body-shape priors, e.g., Skinned Multi-Person Linear Model (SMPL), but depend on point-based backbones and potentially temporal aggregation to compensate for weak features; heatmap approaches preserve richer, reflectivity-level radar features, yet usually regress only 3D keypoints, ignoring body-shape priors. In this paper, by retaining heatmap fidelity and simultaneously exploiting shape priors, we propose **RHAMP**: a **R**adar **HeA**tmap-to-**SMP**L **P**ose transformer for 3D human pose estimation. Specifically, each radar view is encoded by the backbone network, and a set of person queries cross-attends to the multi-view radar features to produce per-instance SMPL parameters in a single end-to-end stage. Experiments on the public HIBER dataset confirm the effectiveness of the proposed approach over a list of baselines.

***Index Terms*—** Pose estimation, radar perception, SMPL.

1. INTRODUCTION

Complementary to vision solutions, emerging radar technology has emerged as a particularly promising technology for indoor perception tasks, offering robustness under occlusion and low-light conditions, while providing high sensitivity and a broad sensing range that enables the capture of subtle body movements [1–21]. Existing studies on radar-based 3D pose estimation are mainly based on sparse point clouds (PCs) and radar heatmaps; see Fig. 1 (a) and (b).

For PC-based approaches, mmMesh [5] and its extension M^{4}esh [10] reconstruct 3D human meshes from sparse radar points by aggregating features around anchor points using LSTM networks. To mitigate weak features, both methods leverage the Skinned Multi-Person Linear (SMPL) model [22] as a strong body-shape prior to enforce anatomically plausible pose prediction. Building on this paradigm, mmBaT [13] introduces a multi-task framework that predicts body translations and reconstructs meshes in a coarse-to-fine manner via a skeleton-aware estimator, enhancing robustness to noisy inputs. mmDEAR [18] instead addresses point cloud sparsity with a two-stage pipeline: radar points are first densified through image-guided enhancement during training, followed by refinement via 2D–3D feature fusion. Despite these advances, e.g., a diffusion process [16], to mitigate sparsity, radar points remain inherently limited in geometric detail and fail to capture subtle reflections.

Radar heatmaps, on the other hand, provide dense and structured observations that preserve both spatial continuity and motion cues. Heatmap-based approaches include RF-Pose3D [3], which pioneered the inferring of 3D skeletons from RF signals by traditional convolution backbone networks. RPM 2.0 [14] advanced this direction by introducing multi-view fusion and spatiotemporal attention for multiperson estimation. More recently, QRFPose [15] reformulated the task as a query-based framework with deformable attention. Beyond these, subsequent works have further explored richer radar representations, multi-target scenarios, and advanced deep learning strategies [8, 12, 19, 20]. However, these methods directly regress keypoint coordinates from radar representations, with the suppression of implausible body configurations (e.g., unrealistic joint orientations) relying solely on the model's implicit capacity to learn anatomical priors.

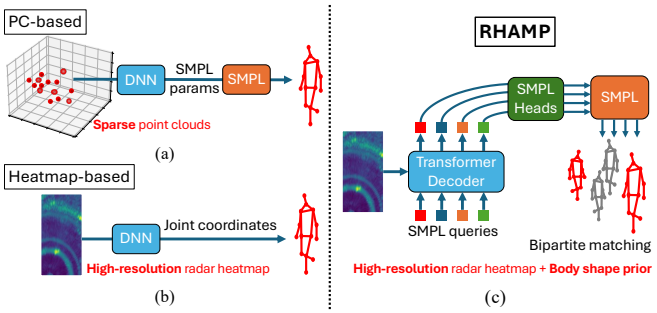

Fig. 1. RHAMP (c) integrates the strengths of existing radar-based 3D pose estimation pipelines (a) and (b) by retaining high-resolution radar heatmaps to preserve strong backbone features and incorporating human body-shape priors (e.g., SMPL) within a query-based detection framework for effective multi-person handling.

In this paper, we propose **RHAMP**: a **R**adar **HeA**tmap-to-**SMP**L **P**ose transformer for 3D human pose estimation; see Fig. 1 (c). RHAMP retains the fidelity of high-resolution radar heatmaps to preserve strong backbone features, while simultaneously exploiting human body-shape priors (e.g., SMPL) within a query-based detection framework to effectively handle multi-person scenarios. Specifically, each radar view is encoded by a backbone network, and a set of person queries cross-attends to the multi-view radar features to regress per-instance SMPL parameters in a single end-to-end stage. This design couples the geometric richness of radar heatmaps with the structural regularization of parametric human body-shape SMPL priors, thereby improving robustness under occlusion and challenging conditions. Extensive experiments on the public HIBER dataset demonstrate that RHAMP consistently outperforms strong baselines, validating the effectiveness of integrating high-resolution radar representations with human body-shape priors.

SK's work was initiated during his internship at MERL, while PW's work was done in part during his visit to the University of Oxford as an Academic Visitor.

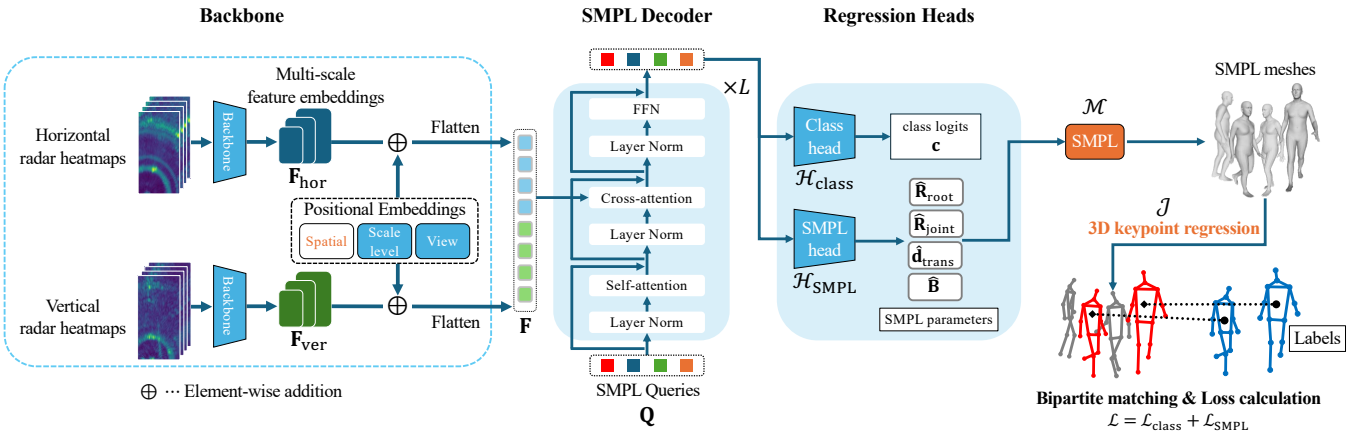

Fig. 2. The architecture of RHAMP. Blue-shaded modules denote learnable components, while orange components (e.g., spatial positional embedding, SMPL model, and 3D keypoint regression) are fixed.

2. PROBLEM FORMULATION

We formulate radar-based 3D pose estimation as a problem of estimating human keypoints in the 3D real-world coordinate system from multi-view radar heatmaps, while simultaneously leveraging the SMPL model as a parametric human body prior. The SMPL model uses a compact set of parameters: root orientation parameters $\mathbf{R}_{\text{root}} \in \mathbb{R}^{1\times 6}$, joint rotation parameters $\mathbf{R}_{\text{joint}} \in \mathbb{R}^{K\times 6}$ body shape parameters $\mathbf{B} \in \mathbb{R}^{1\times 10}$ and a global translation $\mathbf{d}_{\text{trans}} \in \mathbb{R}^{1\times 3}$. Through a differentiable SMPL model $\mathcal{M}$, it generates a realistic human body mesh $\mathbf{V}$ with V vertices, from which 3D keypoint coordinates $\mathbf{P}$ can be obtained via a linear regressor $\mathcal{J}$,

$$\begin{aligned} \mathbf{V} &= \mathcal{M}(\mathbf{R}_{\text{root}}, \mathbf{R}_{\text{joint}}, \mathbf{d}_{\text{trans}}, \mathbf{B}) \in \mathbb{R}^{V\times 3}, \\ \mathbf{P} &= \mathcal{J}(\mathbf{V}) \in \mathbb{R}^{K\times 3}. \end{aligned} \tag{1}$$

Given horizontal (range-azimuth) $\mathbf{X}_{\text{hor}}(t) \in \mathbb{R}^{H\times W}$ and vertical (range-elevation) heatmaps $\mathbf{X}_{\text{ver}}(t) \in \mathbb{R}^{H\times W}$ at time t, we group T consecutive frames as $\mathbf{X}_{\text{hor/ver}} \in \mathbb{R}^{T\times H\times W}$. Existing radar heatmap based approaches directly regress 3D keypoint using multi-view radar views as:

$$\{\hat{\mathbf{P}}_i\}_{i=1}^{N_q} = f(\mathbf{X}_{\text{hor}}, \mathbf{X}_{\text{ver}}; \boldsymbol{\theta}), \tag{2}$$

where f represents the 3D pose estimation pipeline with weights $\boldsymbol{\theta}$ and N_q is the number of potential human objects. For the PC-based approach, we have

$$\{\hat{\mathbf{P}}_i\}_{i=1}^{N_q} = f\big(\text{CFAR}(\mathbf{X}_{\text{hor}}, \mathbf{X}_{\text{ver}}); \boldsymbol{\theta} \mid \mathcal{M}, \mathcal{J}\big), \tag{3}$$

where radar PCs are often obtained by applying the CFAR (constant false alarm rate) operation over the radar heatmaps, and $\mathcal{M}$ and $\mathcal{J}$ are defined above. The problem of interest here is to integrate the strengths of the two radar-based 3D pose pipelines as

$$\{\hat{\mathbf{P}}_i\}_{i=1}^{N_q} = f(\mathbf{X}_{\text{hor}}, \mathbf{X}_{\text{ver}}; \boldsymbol{\theta} \mid \mathcal{M}, \mathcal{J}). \tag{4}$$

3. HEATMAP-TO-SMPL 3D POSE ESTIMATION

Fig. 2 illustrates the architecture of the proposed multi-view radar transformer, which consists of the following key components.

3.1. Architecture

Backbone: Given $\mathbf{X}_{\text{hor}}$ and $\mathbf{X}_{\text{ver}}$, separate backbone networks generate multi-scale feature embeddings:

$$\mathbf{F}_{\text{hor/ver}} = \{\mathbf{F}^{(i)}_{\text{hor/ver}}\}_i = \texttt{backbone}(\mathbf{X}_{\text{hor/ver}}), \tag{5}$$

where the i-th scale feature embedding is denoted as $\mathbf{F}^{(i)}_{\text{hor/ver}} \in \mathbb{R}^{\frac{H}{s_i}\times\frac{W}{s_i}\times d}$, with $s_i \in \{8, 16, 32\}$ representing the stride and d the feature dimension. In our implementation, we employ ResNet-18 [23] as the backbone network.

Positional Embedding: In addition to the feature embeddings $\mathbf{F}^{(i)}_{\text{hor/ver}}$, we incorporate three types of positional embeddings: 1) *spatial positional embedding*, where a 2D sinusoidal embedding $\mathbf{E}^{(i)}_{\text{pos}} \in \mathbb{R}^{\frac{H}{s_i}\times\frac{W}{s_i}\times d}$ is applied to the 2D range–angle (azimuth/elevation) coordinates of the two radar views; 2) *scale-level positional embedding*, where a learnable embedding $\mathbf{E}^{(i)}_{\text{lvl}} \in \mathbb{R}^{1\times 1\times d}$ is used at the i-th feature resolution scale; and 3) *view positional embedding*, where learnable embeddings $\mathbf{E}_{\text{hor/ver}} \in \mathbb{R}^{1\times 1\times d}$ are introduced to differentiate between the two radar views. The overall embeddings are given as

$$\tilde{\mathbf{F}}^{(i)}_{\text{hor/ver}} = \mathbf{F}^{(i)}_{\text{hor/ver}} + \mathbf{E}^{(i)}_{\text{pos}} + \tilde{\mathbf{E}}^{(i)}_{\text{lvl}} + \tilde{\mathbf{E}}_{\text{hor/ver}}, \tag{6}$$

where $\tilde{\mathbf{E}}^{(i)}_{\text{lvl}}$ and $\tilde{\mathbf{E}}_{\text{hor/ver}}$ denote the embeddings broadcast across all spatial range-angle coordinates. The embedded features are then flattened into a token sequence $\mathbf{F} \in \mathbb{R}^{(\sum_i 2HW/s_i^2)\times d}$. This unified representation allows subsequent modules to jointly attend to cues from all radar features from each view and resolution scales.

SMPL Decoder: We design the decoder in a query-based manner, following the DETR paradigm [24] but tailored to SMPL parameter regression. A set of N_q learnable SMPL queries $\mathbf{Q}^{(0)} \in \mathbb{R}^{N_q\times d}$ is introduced, each intended to represent a candidate human instance. The radar feature tokens $\mathbf{F}$ serve as key-value pairs, while queries act as dynamic slots that extract instance-specific information. Each decode layer $\mathcal{D}^{(l)}$ updates the queries through self-attention among queries, cross-attention to the radar features, and feed-forward networks (FFNs):

$$\mathbf{Q}^{(l)} = \mathcal{D}^{(l)}\left(\mathbf{Q}^{(l-1)}, \mathbf{F}\right), \quad l = 1, \cdots, L. \tag{7}$$

Regression Heads: After the L iterations of query refinement in the SMPL decoder, each query embedding $\mathbf{Q}^{(L)}$ is passed through

Table 1. MPJPE performance (↓) comparison between RHAMP and baselines (unit: mm).

	Methods	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Overall**	h	v	d
WALK	mmMesh†	175.7	106.3	154.2	154.6	211.2	135.9	141.6	203.6	163.2	96.7	43.0	97.0
	PETR‡	94.3	61.5	76.2	92.0	133.9	66.0	86.1	125.6	93.9	52.8	30.4	53.6
	QRFPose‡	88.4	59.7	71.4	86.1	125.5	63.9	83.0	121.8	89.4	54.9	48.8	**26.0**
	RHAMP‡	**46.2**	**42.2**	**51.6**	**62.8**	**84.9**	**40.0**	**60.4**	**82.8**	**61.0**	**33.6**	**21.3**	33.5
MULTI	mmMesh†	-	-	-	-	-	-	-	-	-	-	-	-
	PETR‡	79.0	52.8	64.1	79.4	124.3	57.5	88.0	146.5	89.3	49.7	28.6	51.2
	QRFPose‡	84.3	62.2	72.1	86.8	125.6	63.6	80.2	123.0	89.2	52.8	49.6	**27.6**
	RHAMP‡	**52.3**	**48.8**	**56.4**	**68.4**	**91.9**	**48.0**	**67.8**	**91.3**	**67.8**	**37.4**	**22.8**	38.2

†: PC-based method(s), ‡: Heatmap-based method(s).

multi-layer perceptron (MLP) heads to regress the class logits and SMPL parameters. The class head $\mathcal{H}_{\texttt{class}}$ produces a scalar logit $\mathbf{c} \in \mathbb{R}$ per query, indicating whether the query corresponds to a valid person (1) or background (0). The SMPL head $\mathcal{H}_{\texttt{SMPL}}$ predicts a set of SMPL parameters, including root orientations $\hat{\mathbf{R}}_{\texttt{root}} \in \mathbb{R}^{N_q \times 6}$, joint rotations $\hat{\mathbf{R}}_{\texttt{joint}} \in \mathbb{R}^{N_q \times K \times 6}$ where K denotes the number of keypoints, global translations $\hat{\mathbf{d}}_{\texttt{trans}} \in \mathbb{R}^{N_q \times 3}$, and body shape coefficients $\hat{\mathbf{B}} \in \mathbb{R}^{N_q \times 10}$. To enhance training stability, root and joint rotations are represented using the continuous 6D formulation in [25]. All parameters are regressed jointly from queries as

$$\{\hat{\mathbf{R}}_{\texttt{root}}, \hat{\mathbf{R}}_{\texttt{joint}}, \hat{\mathbf{d}}_{\texttt{trans}}, \hat{\mathbf{B}}\} = \mathcal{H}_{\texttt{SMPL}}(\mathbf{Q}^{(L)}). \tag{8}$$

SMPL-guided Keypoint Estimation: From the regressed SMPL parameters, the SMPL model constructs posed meshes $\hat{\mathbf{V}}$ via a differentiable SMPL model $\mathcal{M}$ and a fixed keypoint regressor $\mathcal{J}$ maps mesh vertices to anatomical poses,

$$\{\hat{\mathbf{V}}_i\}_{i=1}^{N_q} = \mathcal{M}(\hat{\mathbf{R}}_{\texttt{root}}, \hat{\mathbf{R}}_{\texttt{joint}}, \hat{\mathbf{d}}_{\texttt{trans}}, \hat{\mathbf{B}}) \in \mathbb{R}^{N_q \times V \times 3}, \tag{9}$$

$$\{\hat{\mathbf{P}}_i\}_{i=1}^{N_q} = \mathcal{J}(\{\hat{\mathbf{V}}_i\}_{i=1}^{N_q}) \in \mathbb{R}^{N_q \times K \times 3}. \tag{10}$$

As the decoder outputs a set of N_q candidate predictions, we adopt bipartite matching with the Hungarian algorithm [24] to associate each prediction with ground-truth persons in a one-to-one manner.

3.2. Loss Function

The overall loss is a weighted combination of the class loss $\mathcal{L}_{\texttt{cls}}$ and the SMPL-related loss $\mathcal{L}_{\texttt{SMPL}}$. First, for bipartite matching, we match N_q predicted queries with N ground-truth labels. The matching cost for a ground-truth–prediction pair (i, j) is defined as

$$\mathcal{C}(i,j) = \lambda_{\texttt{cls}} \mathcal{C}_{\texttt{cls}}(\hat{\mathbf{c}}_j) + |\mathbf{V}_i - \hat{\mathbf{V}}_j| + |\mathbf{P}_i - \hat{\mathbf{P}}_j|, \tag{11}$$

where $\lambda_{\texttt{cls}}$ is a weight coefficient, $\mathcal{C}_{\texttt{cls}}$ denotes the Focal loss [26], and $\mathbf{V}_i$ and $\mathbf{P}_i$ are the ground-truth mesh vertices and 3D keypoints for person i, while $\hat{\mathbf{V}}_j, \hat{\mathbf{P}}_j$ denoting corresponding predictions. After determining the optimal assignment, the SMPL loss is computed on matched pairs as:

$$\mathcal{L}_{\texttt{SMPL}} = \sum_{i=1}^{N} (|\mathbf{V}_i - \hat{\mathbf{V}}_{\sigma(i)}| + |\mathbf{P}_i - \hat{\mathbf{P}}_{\sigma(i)}| + \lambda_\beta |\hat{\mathbf{B}}_{\sigma(i)}|), \tag{12}$$

where $\sigma(i)$ denotes the prediction matched to the ground truth i, $\hat{\mathbf{B}}_{\sigma(i)}$ are the predicted body shape coefficients, and λ_β is a loss weight. Finally, the total loss function is thus formulated as

$$\mathcal{L} = \lambda_{\texttt{cls}} \mathcal{L}_{\texttt{cls}} + \mathcal{L}_{\texttt{SMPL}}. \tag{13}$$

4. PERFORMANCE EVALUATION

4.1. Dataset

We evaluate the performance on the HIBER dataset [6], which provides two high-resolution (horizontal and vertical) radar heatmaps with 3D keypoint labels. We exclude the MMVR [11] and HuPR [9] datasets as they contain only 2D keypoints and RT-Pose [12] as it includes a single high-resolution (horizontal) radar view. We use radar heatmaps from views 02 through 10 under the "WALK" (single-subject) and "MULTI" (two-subject) protocols for training, validation, and testing. To compute the SMPL-related loss $\mathcal{L}_{\texttt{SMPL}}$, we generate SMPL ground-truth parameters by solving an inverse kinematics (IK) problem that minimizes the error between the annotated keypoints $\mathbf{P}_i$ and corresponding SMPL joints $\hat{\mathbf{P}}_i$ over $\mathbf{R}_{\texttt{root}}$, $\mathbf{R}_{\texttt{joint}}$, $\mathbf{d}_{\texttt{trans}}$, and $\mathbf{B}$ using VPoser [27] with the L-BFGS optimizer.

4.2. Implementation and Metrics

We use $T = 4$ consecutive radar frames as input and set the number of SMPL decoder queries to $N_q = 10$ with a feature dimension of $d = 128$. Following the HIBER dataset, we use $K = 14$ keypoints for 3D human pose annotation. The loss weights are set to $\lambda_\beta = 0.5$ and $\lambda_{\texttt{cls}} = 0.5$. We train the models for 50 epochs using the AdamW optimizer with an initial learning rate of 2×10^{-4} and a batch size of 32. The learning rate follows a cosine decay schedule. The model with the lowest validation loss is selected, and early stopping with a patience of 5 epochs is applied to prevent overfitting.

For evaluation, we use the Mean Per Joint Position Error (MPJPE), measured in millimeters, which is defined as the average Euclidean distance between the predicted and ground-truth 3D keypoints. We also report axis-specific errors along the horizontal (h), vertical (v), and depth (d) dimensions to enable a more fine-grained analysis of prediction accuracy.

4.3. Baselines

We include the following radar-based 3D pose estimation baseline methods, prioritizing open-sourced pipelines:

PC-based: We adopt **mmMesh** [5] as a PC-based baseline. mmMesh is a recurrent neural network (RNN)-based method reconstructing human meshes by using multiple consecutive frames and fixed-size anchor points for spatial grouping. Since mmMesh is designed for single-person scenarios, we evaluate its performance with the "WALK" protocol. Radar points are generated from raw radar heatmaps by constructing a 3D voxel grid via a depth-wise outer product of multi-view heatmaps and selecting the top-K voxels.

Heatmap-based: We adopt **PETR** [28] and **QRFPose** [15] as radar heatmap-based baselines. Both are query-based methods that use deformable attention [29] to aggregate multi-view radar features

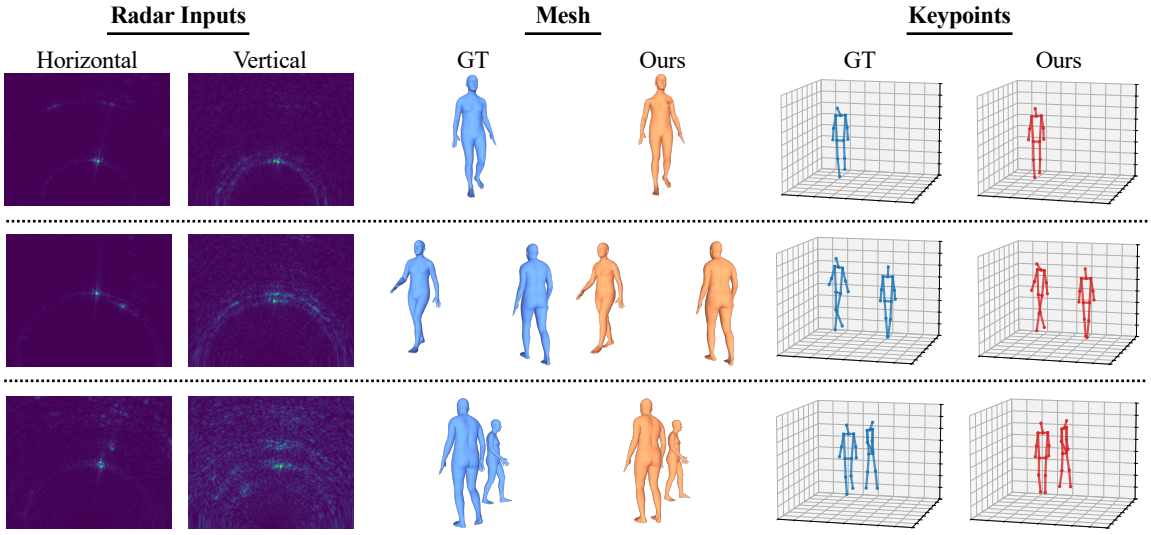

Fig. 3. Qualitative results of RHAMP. The predicted meshes and keypoints effectively capture the subtle body structure.

and predict 3D keypoints. Since PETR is originally designed for RGB-based 2D keypoint estimation, we adapt its implementation to handle radar heatmaps and extend it to 3D keypoint estimation.

4.4. Comparison to Baseline Methods

Table 1 compares RHAMP with the selected baselines. RHAMP achieves the lowest MPJPE among all compared approaches by jointly exploiting radar heatmaps and the human body prior encoded in SMPL. While heatmap-based methods (PETR, QRFPose) already outperform the point cloud-based mmMesh, RHAMP yields further gains, reducing MPJPE by 31.8% on WALK and 24.0% on MULTI compared to the strongest baseline, QRFPose. Per-axis analysis reveals that the improvement is most pronounced along the vertical axis, where RHAMP consistently achieves the lowest error, markedly reducing MPJPE from $48.8\,\mathrm{mm}$ to $21.3\,\mathrm{mm}$ in WALK and from $49.6\,\mathrm{mm}$ to $22.8\,\mathrm{mm}$ in MULTI compared to QRFPose. We attribute this gain to the incorporation of SMPL, which encodes strong priors on the kinematic structure and vertical arrangement of human keypoints.

4.5. Effects of Single- and Dual-Views

We evaluate the effect of single-view and dual-view radar heatmap inputs in Table 2. Using only the vertical view substantially degrades performance, particularly along the horizontal (h) axis, as it lacks sufficient cues for infering the subject's lateral position. In contrast, the horizontal-only setting achieves performance comparable to the dual-view setting, indicating that SMPL prior enables coherent vertical reconstruction even in the absence of explicit vertical-view observations. Nevertheless, incorporating both views yields consistent improvement, with gains of 7.7% on WALK and 10.8% on MULTI over the horizontal-only setting.

4.6. Qualitative Results

Fig. 3 shows the qualitative results of RHAMP. The estimated meshes closely match the ground truth, yielding well-aligned keypoints that accurately capture body orientation, overall position, and peripheral joints such as elbows and knees. The top row shows accurate full-body reconstruction for an isolated subject, while the middle and bottom rows confirm that RHAMP recovers plausible poses even in close-proximity multi-person settings. These examples highlight the advantage of combining high-resolution radar heatmaps with SMPL priors to recover coherent full-body keypoints.

Table 2. Effect of Horizontal (Hor) and Vertical (Ver) radar views.

	Views	MPJPE	h	v	d
WALK	Hor	66.1	37.7	22.0	35.1
	Ver	92.2	56.7	24.0	49.3
	Hor + Ver	**61.0**	**33.6**	**21.3**	**33.5**
MULTI	Hor	76.0	41.0	29.7	40.5
	Ver	156.1	110.4	29.1	77.4
	Hor + Ver	**67.8**	**37.4**	**22.8**	**38.2**

4.7. Computational Complexity

The computational complexity of RHAMP is primarily determined by the backbone and the Transformer decoder. The ResNet-18 backbone consists of 2D convolutional layers, resulting in a complexity of $\mathcal{O}(HWd^2)$. The SMPL decoder has an approximate complexity of $\mathcal{O}(N_qHWd + N_qd^2 + N_q^2d)$, where the three terms correspond to cross-attention, linear projection layers, and self-attention, respectively. Consequently, the overall computational complexity is approximately $\mathcal{O}(HWd^2 + N_qHWd + N_qd^2 + N_q^2d)$.

5. CONCLUSION

We presented a transformer-based framework for 3D pose estimation using radar heatmaps and SMPL priors. Our query-based decoding process generates per-instance pose estimations end-to-end, combining the high fidelity of radar heatmaps with body shape constraints. Empirical evaluations on the HIBER benchmark show that our design outperforms existing point-cloud and heatmap-only approaches, validating the benefits of fine-grained radar features and parametric body models. As future work, it would be interesting to evaluate RHAMP beyond 3D keypoint accuracy by incorporating mesh-level metrics that assess surface geometry and structural detail, as well as to study its robustness to occlusions and cluttered environments using more diverse radar datasets such as RT-Pose.

6. REFERENCES

[1] Avik Santra et al., "Machine learning-powered radio frequency sensing: A review," *IEEE Sensors Journal*, vol. 25, no. 13, pp. 23164–23183, 2025.

[2] Mingmin Zhao et al., "Through-wall human pose estimation using radio signals," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 7356–7365.

[3] Mingmin Zhao et al., "RF-based 3D skeletons," in *Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)*, Aug. 2018, pp. 267–281.

[4] Peijun Zhao et al., "mID: Tracking and identifying people with millimeter wave radar," in *International Conference on Distributed Computing in Sensor Systems (DCOSS)*, May 2019, pp. 33–40.

[5] Hongfei Xue et al., "mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave," in *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, June 2021, pp. 269–282.

[6] Zhi Wu et al., "RFMask: A simple baseline for human silhouette segmentation with radio signals," *IEEE Transactions on Multimedia*, pp. 1–12, 2022.

[7] Peijun Zhao et al., "CubeLearn: End-to-end learning for human motion recognition from raw mmWave radar signals," *IEEE Internet of Things Journal*, vol. 10, no. 12, pp. 10236–10249, June 2023.

[8] Xie Qian et al., "mmPoint: Dense human point cloud generation from mmwave," in *The British Machine Vision Conference (BMVC)*, 2023.

[9] Shih-Po Lee et al., "HuPR: A benchmark for human pose estimation using millimeter wave radar," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 5715–5724.

[10] Hongfei Xue et al., "M^{4}esh: mmWave-based 3D human mesh construction for multiple subjects," in *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Nov. 2022, pp. 391–406.

[11] M. Mahbubur Rahman et al., "MMVR: Millimeter-wave multi-view radar dataset and benchmark for indoor perception," in *European Conference on Computer Vision (ECCV)*, 2024, p. 306–322.

[12] Yuan-Hao Ho et al., "RT-Pose: A 4D radar tensor-based 3D human pose estimation and localization benchmark," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 107–125.

[13] Jiarui Yang et al., "mmBaT: A multi-task framework for mmwave-based human body reconstruction and translation prediction," Apr. 2024, pp. 8446–8450.

[14] Chunyang Xie et al., "RPM 2.0: RF-based pose machines for multi-person 3D pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 490–503, Jan. 2024.

[15] Hong Wan et al., "QRFPose: Query-based 3D pose estimation using radio signals," in *International Conference on Wireless Communications and Signal Processing (WCSP)*, 2024, pp. 1271–1277.

[16] Junqiao Fan et al., "Diffusion model is a good pose estimator from 3D RF-Vision," in *European Conference of Computer Vision (ECCV)*, 2024, pp. 1–18.

[17] Ryoma Yataka et al., "RETR: Multi-view radar detection transformer for indoor perception," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 19839–19869, Nov. 2024.

[18] Jiarui Yang et al., "mmDEAR: mmWave point cloud density enhancement for accurate human body reconstruction," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2025, pp. 11227–11233.

[19] Sorachi Kato et al., "RAPTR: Radar-based 3D pose estimation using transformer," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[20] Bing Zhu et al., "ProbRadarM3F: mmWave radar-based human skeletal pose estimation with probability map guided multi-format feature fusion," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–11, 2025.

[21] Ryoma Yataka et al., "Indoor multi-view radar object detection via 3D bounding box diffusion," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[22] Matthew Loper et al., "SMPL: a skinned multi-person linear model," *ACM Transactions of Graphics*, vol. 34, no. 6, pp. 1–16, Nov. 2015.

[23] Kaiming He et al., "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[24] Nicolas Carion et al., "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[25] Yi Zhou et al., "On the continuity of rotation representations in neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 5738–5746.

[26] Tsung-Yi Lin et al., "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[27] Georgios Pavlakos et al., "Expressive body capture: 3D hands, face, and body from a single image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[28] Dahu Shi et al., "End-to-end multi-person pose estimation with transformers," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11059–11068, June 2022.

[29] Xizhou Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations (ICLR)*, 2021, pp. 1–16.