FROM HIDDEN TO HANDY: MAKING BIAS EXPLICIT AND ACTIONABLE IN MMWAVE POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Practical mmWave sensing systems operate with sparse returns, specular reflections, and occlusions, pushing pose estimators to lean on statistical priors rather than sensor evidence. This systems-level reality yields biased joint predictions that hurt downstream tasks (e.g., gesture/activity recognition). We present MMJOINTS, which augments a pre-trained, black-box mmWave 3D pose estimator with per-joint descriptors that characterize the degree of local radar support and the reliability of each predicted joint location. Instead of hiding bias, MMJOINTS makes it explicit and usable—improving interpretability of pose and accuracy of downstream models. Evaluated on over 115,000 frames across 13 pose-estimation settings, MMJOINTS consistently estimates informative ξ and κ descriptors across diverse models and sensing conditions, improves joint accuracy by up to 12.5%, and boosts activity recognition accuracy by up to 16%.

1 INTRODUCTION

mmWave-based 3D pose estimation over-relies on priors. Pose estimation, which estimates 3D joint coordinates, underpins downstream tasks such as gesture recognition (Liu et al., 2022), gait analysis (Alanazi et al., 2022), patient monitoring (Sengupta et al., 2020), and posture tracking (Kong et al., 2022), where accuracy and reliability are critical to prevent error propagation. Yet mmWave-based models (An & Ogras, 2021; Chen et al., 2022; Xue et al., 2021) show a distinct bias: when radar reflections are sparse, they lean more on learned human-body priors than on sensor evidence. Unlike vision systems with dense spatial detail, mmWave returns are often incomplete or noisy (Singh et al., 2019; Jiang et al., 2018; Xie et al., 2022; Li et al., 2022), encouraging models to fit statistical priors rather than interpret the signal; in extreme cases they down-weight the radar because priors yield higher average pose accuracy.

Example. If lower-body reflections are sporadically missing due to radar placement or occlusions, the model often ignores them and defaults to memorized lower-body coordinates (An & Ogras, 2021; Chen et al., 2022), yielding decent accuracy but lacking sensor-based verification. This can produce reasonable estimates, but it risks applications requiring precise tracking of task-sensitive joints (Zhang et al., 2023), where misplaced confidence in statistical priors can mislead downstream models (Liu et al., 2020; Xu et al., 2022b; Ramasubramanian, 2017; Xu et al., 2022a).

Priors are vital for learning, but over-reliance skews results. While model bias is a well-known phenomenon with various mitigation strategies (Xue et al., 2021; 2023; Chen et al., 2022; An & Ogras, 2021), it remains an inherent—and often necessary—component of mmWave sensing. mmWave radars capture inconsistent, unreliable body reflections (An et al., 2022; Wang et al., 2023); no placement ensures adequate returns from all joints, even with software/hardware augmentation (Zheng et al., 2024; Xue et al., 2022). A model trained solely on radar, without priors, struggles to learn human pose because signals are sparse across joints. Thus, *prior bias is crucial for learning meaningful representations, but over-reliance is harmful: models may ignore weak or inconsistent radar evidence, yielding inaccurate poses that propagate errors to downstream applications.*

A new proposal—let's make bias explicit and actionable. Instead of attempting to surgically mitigate bias—which is both computationally and theoretically challenging—we propose a fundamentally different approach: *making bias explicit in the predicted pose*. Specifically, we introduce two per-joint descriptors: a sensing score (ξ) measuring local radar support and a reliability score (κ) representing an error-derived reliability target. This explicit representation exposes mismatches between available sensor evidence and prediction reliability, making them actionable to downstream

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tasks. Rather than suppressing bias, we characterize these mismatches and use them to produce more interpretable mmWave pose representations and more reliable downstream predictions.

Estimating ξ and κ requires an inverse process. The ground-truth-defined ξ and κ cannot be obtained directly during inference because the true joint locations are unavailable. We therefore introduce a *surrogate pose*—a refined pose that best agrees with the observed data—estimated using only the input and the model's output. We learn an inverse mapping ***pose*** $\rightarrow$ ***expected mmWave signal distribution*** and compare predicted and observed signals; their discrepancy drives refinement from the model's estimate to obtain the surrogate. As a proxy for missing ground truth, this surrogate enables estimation of the sensing and reliability scores, providing an output-level characterization of radar support and prediction reliability for downstream use.

MMJOINTS: a bias-aware pose-augmentation module for robust downstream use. MMJOINTS augments a pre-trained pose estimator's output with per-joint ξ and κ, producing a 5D tuple (x, y, z, ξ, κ). This augmentation improves accuracy in downstream tasks (e.g., gesture/activity recognition) compared to using raw poses alone. MMJOINTS treats the estimator as a black box—no access to internals—and works with locally hosted or API-served models. Its overhead is $\sim$63 ms per inference, suitable for edge deployment. Note that MMJOINTS does not replace or fine-tune the base pose estimator; it is a separately trained post-hoc augmentation module. When pose refinement is desired, a lightweight downstream model can use the estimated descriptors to refine the base output.

Comprehensive evaluation of MMJOINTS across models, datasets, and real-world scenarios. We evaluate MMJOINTS across four state-of-the-art pose estimators plus several variants, using two public datasets and one self-collected dataset (over 115,000 signal frames, 13 pose-estimation settings). We further test in real-world scenarios with seven pose estimators, 10 activities, and four environments, targeting two downstream tasks: improving pose estimation and activity recognition. MMJOINTS consistently achieves accurate ξ and κ estimation across models and sensing conditions, improves joint positions by up to 12.5%, and boosts activity recognition accuracy by up to 16%.

2 SENSING, RELIABILITY, AND MODEL BIAS

We characterize each predicted joint using two scores—sensing ξ and reliability κ—that provide an output-level view of the available radar evidence and prediction quality, without requiring access to the estimator's internal decision process.

Sensing Score (ξ). The sensing score quantifies how strongly mmWave returns concentrate around a joint. We first compute a signal-strength proxy $\psi_j = \sum_{n=1}^{N} I_n / d_{j,n}^2$, where the point cloud contains N points, I_n is the return intensity of point n, and $d_{j,n}$ is its distance to joint j. We define the sensing score as $\xi_j = \sigma(\psi_j - \bar{\psi})$, where $\sigma(\cdot)$ maps the score to $(0, 1)$, and $\bar{\psi}$ is the training-set mean used as a configuration reference point for determining whether the joint is sensed.

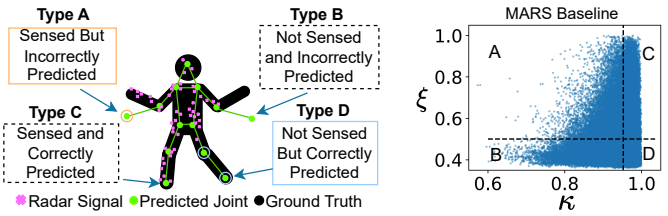


Figure 1: Joint types and their distribution for the MARS model.

Model Input	Accuracy
Joint	75.26%
Joint + Descriptor	86.41%

(a) Accuracy gain.

Joint Type	Correct		Incorrect	
	Not Sensed	Sensed	Not Sensed	Sensed
#Joint	1,078	31	1,048	47

(b) Cases where descriptors helped.

Table 1: Effect of pose characterization on activity recognition.

Reliability Score (κ). The reliability score quantifies confidence in a joint estimate by converting its error to a bounded, body-scale normalized value. Let D_j be the distance between predicted and ground-truth joint j, and define $f(D_j) = 1 - \sigma\big(D_j - \frac{1}{2}\bar{D}_0\big)$, where $\bar{D}_0$ is the training-set average torso length (as in PCK@0.5 scaling) and $\sigma(\cdot)$ bounds the range. We define the reliability score as $\kappa_j = f(D_j)/f(0)$, so $\kappa_j \in (0, 1]$ with $\kappa_j = 1$ when $D_j = 0$, ensuring interpretability and cross-subject comparability.

Bias in Existing Models. We examine MARS (An & Ogras, 2021), mmBody (Chen et al., 2022), and mmMesh (Xue et al., 2021) across their public datasets. Ideally, correct predictions should coincide

with strong sensing and reliability. Instead, Figure 1 shows that 36.84% of joints are correct despite low ξ, while 13.22% are incorrect despite high ξ. This mismatch persists across models and datasets (Appendix B), showing that prediction correctness alone does not reveal whether a joint is supported by radar evidence.

Are These Descriptors Useful? We test this using oracle (ξ, κ), which are available only with ground truth. On MARS, an activity classifier using predicted 3D joints achieves 75.26% accuracy; adding per-joint oracle (ξ, κ) increases accuracy to 86.41%, as shown in Table 1(a). Table 1(b) further illustrates how the descriptors distinguish signal-supported from uncertain predictions. These gains motivate estimating (ξ, κ) without ground truth using the surrogate pose in Section 6.

3 Overview of mmJoints

mmJoints augments a pre-trained pose estimator's output joints (x_i, y_i, z_i) with two additional per-joint descriptors—a sensing score (ξ_i) and a reliability score (κ_i)—forming a 5D representation $(x_i, y_i, z_i, \xi_i, \kappa_i)$. Because ground truth is unavailable at inference, mmJoints computes a surrogate pose in a learned latent space, seeded by the pre-trained model's output. It learns an inverse mapping from pose to expected sensor signal, compares predicted and observed signals, and updates the latent pose. Operating entirely in latent space improves robustness to noise and uncertainty while enabling principled estimation of sensing and reliability scores for each joint.

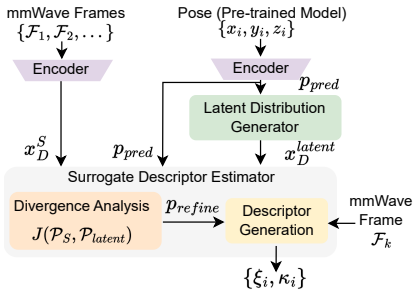


Figure 2: Overview of mmJoints.

Figure 2 illustrates mmJoints's processing pipeline: from mmWave frames $\{\mathcal{F}_t\}$ and a pre-trained pose estimator, *Signal and Pose Encoders* aggregate a fixed window into a distributional signal feature x_D^S and encode the predicted pose into p_{pred}, respectively. *Latent Distribution Generator* learns an inverse map to synthesize the expected latent signal x_D^{latent} and *Surrogate Descriptor Estimator* contrasts it with x_D^S via a divergence $J(\cdot, \cdot)$, yielding a refined surrogate pose p_{refine}. Finally, *Descriptor Generator* conditions on $(p_{\text{pred}}, p_{\text{refine}}, \mathcal{F}_k)$ to produce per-joint sensing and reliability scores (ξ, κ), where k denotes the time stamp corresponding to the predicted pose.

4 Signal and Pose Encoding

4.1 Pose Encoding

We represent diverse human poses as linear combinations of a small set of basis poses, enabling efficient interpolation, generalization, and manipulation. Pose modeling consists of three steps:

Step 1: We cluster training poses using a Gaussian Mixture Model (GMM), selecting the number of clusters C ($C \geq D_p$) by the Bayesian Information Criterion (BIC) (Vrieze, 2012). Each cluster represents a pose category. We rank clusters by $d_i|\{\text{pose}_i\}|$, where d_i is the Euclidean distance between the ith cluster centroid and the mean pose, and select the top-D_p clusters for basis construction.

Step 2: We encode raw poses while preserving continuity and smoothness using a VAE (Kingma & Welling, 2013; Higgins et al., 2017). Its latent features follow a Gaussian distribution with mean μ_p and standard deviation σ_p. We optimize $\mathcal{L}_{\text{P}_{\text{step2}}} = \mathcal{L}_{\text{REC}} + \lambda_{\text{step2}}\mathcal{L}_{\text{G}}$, where $\mathcal{L}_{\text{REC}}$ is the MSE reconstruction loss and $\mathcal{L}_{\text{G}}$ is a Gaussian-prior regularization term for the latent representation.

Step 3: We fine-tune the D_p pose categories to be mutually orthogonal in latent space using Orthogonal Projection Loss (OPL) (Ranasinghe et al., 2021): $\mathcal{L}_{\text{P}_{\text{step3}}} = \mathcal{L}_{\text{OPL-d}} + \lambda_{\text{step3}}\mathcal{L}_{\text{OPL-s}}$, promoting inter-class orthogonality while preserving latent clustering.

4.2 Signal Encoding

mmWave pose-estimation inputs are typically multidimensional, unordered, and sparse point clouds (Chen et al., 2022; Xue et al., 2021; An & Ogras, 2021), making stable reconstruction

difficult for standard encoder-decoder models (Kingma & Welling, 2013). We therefore learn a structured feature space that maps signals from similar poses closer together and those from different poses farther apart.

We use a Point-4D-Transformer-based feature extractor (Fan et al., 2021), designed for point cloud inputs. It generates the distribution feature x_D^S, which consists of the mean μ_S and standard deviation σ_S, forming a Gaussian distribution to represent consecutive K signal instances. The objective is $\mathcal{L}_{\mathrm{S}} = \lambda_0 \mathcal{L}_{\mathrm{triplet}} + \lambda_1 \mathcal{L}_{\mathrm{CE}} + \lambda_2 \mathcal{L}_{\mathrm{G}}$, where $\mathcal{L}_{triplet}$ (Balntas et al., 2016) captures instance-wise correlation, $\mathcal{L}_{\mathrm{CE}}$ preserves pose-cluster structure, and $\mathcal{L}_{\mathrm{G}}$ provides Gaussian-prior regularization of the latent signal representation. λ_0, λ_1 and λ_2 indicate the contribution of these three loss terms. Detailed definitions of all training objectives used throughout Sections 4–6 are provided in Appendix G.

4.3 Signal Distribution Modeling

Recall that in our inverse modeling approach, we map poses to signal distributions. The justification for this approach is provided in Appendix C. To represent the distribution, we extract feature representations of signal instances and model them using a Gaussian Mixture Model. Signal distribution modeling consists of two key steps:

Step 1: For each pose instance $pose_i$, we select the top M closest poses, up to a maximum of M_{max}, by measuring their distances in the latent feature space using the l_2-norm between their respective μ_p. These distances are limited to $||\sigma_{p_i}/2||$. By collecting the corresponding M signal inputs and their associated latent features, we obtain M normal distribution components: $\{\mathcal{N}(\mu_{S_i}, \mathrm{diag}(\sigma_{S_i}^2))\}_{i=1}^{M}$

Step 2: To formalize the distribution representation for effective training, rather than using a nonstandard set form, we construct a GMM with G components, using samples randomly drawn from the distribution components $\mathcal{N}(\mu_{S_i}, diag(\sigma_{S_i})^2)$. G is selected based on the BIC indicator applied to a randomly sub-sampled dataset, with an upper bound to prevent overfitting. This results in a closed-form distribution: $\mathcal{P}(\mathbf{x}) = \sum_{i=1}^{G} \phi_i \mathcal{N}(\mathbf{x} \mid \mu_i, \Sigma_i),\ s.t. \sum_{i=1}^{G} \phi_i = 1$ with respect to the latent signal feature $\mathbf{x}$. Similar to the proposition in (Kingma & Welling, 2013), we use diagonal covariance for each component to maintain an efficient modeling strategy.

5 Latent Distribution Generation

5.1 Why Signal Distribution

Covariance analysis reveals that pose changes correlate weakly with changes in a single radar instance (Δx) but markedly more strongly with changes in the signal's *distribution* (Δx_D). This motivates our inverse formulation: learn pose $\rightarrow$ signal distribution and exploit a compact latent proxy during inference. Because a standalone radar cannot directly sense distributions at runtime, MMJOINTS estimates a latent signal-distribution descriptor from short windows to regularize pose estimation. See Appendix C for the full method—how we define $\Delta x, \Delta p$, compute covariances, reduce features via PCA, fit GMMs for x_D, use symmetric KL for Δx_D, and compare mmWave vs. vision baselines.

5.2 Latent Signal Distribution Construction

We construct an analytically tractable, closed-form signal distribution for an arbitrary input pose. We first define an orthogonal pose basis, then learn to map a pose expressed in this basis to its expected signal distribution, using the empirical signal distributions of the basis poses as references for convergence. Figure 3 illustrates the three steps.

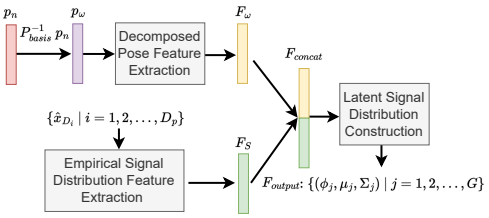


Figure 3: Latent signal distribution construction.

Step 1: From the D_p pose clusters in Section 4.1, we construct $\{(F_{basis_i}, \hat{x}_{D_i}) \mid i = 1, \ldots, D_p\}$, where F_{basis_i} is the average of $\{\mu_p\}$ in the ith cluster and the $\{F_{basis_i}\}$ are mutually orthogonal. The

corresponding empirical signal distribution is $\hat{x}_{D_i} = \{(\phi^i_{basis_j}, \mu^i_{basis_j}, \Sigma^i_{basis_j}) \mid j = 1, \ldots, G\}$. The resulting pose basis is $P_{basis} = [F_{basis_1}, \ldots, F_{basis_{D_p}}] \in \mathbb{R}^{D_p \times D_p}$, which is invertible. For an arbitrary pose $pose_n$, we decompose its latent feature p_n in this basis as $p_\omega = P^{-1}_{basis} p_n$, from which the fine-tuned pose feature F_ω is extracted.

Step 2: We encode the empirical distributions $\{\hat{x}_{D_i}\}$ with a SetTransformer (Lee et al., 2019), using input of size $D_p \times D_S \times G \times D_C$, where D_S is the signal latent dimension and D_C is the dimension of concatenated $(\phi_{basis_j}, \mu_{basis_j}, \mathrm{diag}(\Sigma_{basis_j}))$. The extracted empirical feature F_S is concatenated with F_ω and fed to the latent signal distribution generator. Its output is the flattened feature $F_{output} = \{(\phi_j, \mu_j, \sigma_j^{\circ 2}) \mid j = 1, \ldots, G\}$, where $(\cdot)^{\circ 2}$ denotes the Hadamard power. This empirical representation remains available during inference.

Step 3: Because GMM components are unordered, training requires a permutation-invariant loss. Let $\Gamma_j = (\phi_j, \mu_j, \Sigma_j)$ and $T(\cdot)$ measure the discrepancy between two components. We formulate

$$\mathcal{L}_{\mathrm{LAP}} = \min_{\{a_{i,j}\}} \sum_{1 \le i,j \le G} T(\Gamma_i, \Gamma_j) a_{i,j}, \quad \text{s.t. } \sum_j a_{i,j} = 1,\ \sum_i a_{i,j} = 1,\ a_{i,j} \in \{0, 1\}.$$

This Linear Assignment Problem is solved with the Hungarian algorithm (Martello & Toth, 1987). We use $T(\Gamma_i, \Gamma_j) = \mathcal{L}_{1-\phi} + \mathcal{L}_{1-\mu} + \mathcal{L}_{1-\Sigma}$, i.e., the ℓ_1 losses on the mixture weights, means, and covariances. The final objective is $\mathcal{L}_{\mathrm{Distribution}} = \mathcal{L}_{\mathrm{LAP}} + \lambda_{\mathrm{Div}} \mathcal{L}_{\mathrm{Div}}$, where $\mathcal{L}_{\mathrm{Div}} = \hat{J}(P_S, P_G)$ is the approximate displacement defined in Section 6.

6 Surrogate Descriptor Estimation

6.1 Divergence Analysis

A closed-form metric is defined that quantifies the discrepancy between the expected signal distribution and the current signal input. This metric is used in later stages of surrogate pose construction. The two key steps in this analysis are shown below:

Step 1: To begin with, we consider an input signal feature x^S_D in a Gaussian distribution form with mean μ_S and standard deviation σ_S. The displacement between this representation and the latent signal distribution x^{latent}_D is measured using Jeffreys divergence: $J(\mathcal{P}_S, \mathcal{P}_G) = KL(\mathcal{P}_G || \mathcal{P}_S) + KL(\mathcal{P}_S || \mathcal{P}_G)$, where $KL(\cdot)$ is the KL-divergence (Kullback, 1997); $\mathcal{P}_S$ is the posterior probability from Gaussian distribution $\mathcal{N}(\mu_S,\ diag(\sigma_S)^2)$; $\mathcal{P}_G$ is the posterior probability from x^{latent}_D in GMM form: $\mathcal{P}_G(\mathbf{x}) = \sum_{i=1}^G \phi_i \mathcal{N}(\mathbf{x} \mid \mu_i, \Sigma_i)$ s.t. $\sum_{i=1}^G \phi_i = 1$ ($\phi_i \in [0, 1]$). Our goal is to leverage information from $J(\mathcal{P}_S, \mathcal{P}_G)$ to generate surrogate pose that better aligns with signal space.

Step 2: However, there is no analytically tractable Jeffreys / KL divergence when GMM is involved. Thus, we utilize the approximation of $J(\mathcal{P}_S, \mathcal{P}_G)$, which serves as an upper bound with a tractable closed-form expression. Specifically, we have

$$\begin{aligned} J(\mathcal{P}_S, \mathcal{P}_G) &= \int \mathcal{P}_G \ln\Big(\frac{\mathcal{P}_G}{\mathcal{P}_S}\Big)\, dx + \int \mathcal{P}_S \ln\Big(\frac{\mathcal{P}_S}{\mathcal{P}_G}\Big)\, dx \\ &= \int \Big(\sum_{i=1}^G \phi_i \mathcal{P}^i_c\Big) \ln\Big(\Big(\sum_{i=1}^G \phi_i \mathcal{P}^i_c\Big) \Big/ \mathcal{P}_S\Big)\, dx \\ &\quad + \int \mathcal{P}_S \ln\Big(\mathcal{P}_S \Big/ \Big(\sum_{i=1}^G \phi_i \mathcal{P}^i_c\Big)\Big)\, dx \le \sum_{i=1}^G \phi_i J(\mathcal{P}_S, \mathcal{P}^i_c) \end{aligned}$$

where $\mathcal{P}^i_c$ is the probability from i^{th} mixture component of GMM. The approximation term (weighted Jeffreys divergence) on the right-hand side of the inequality is analytically tractable and can be used effectively for surrogate pose construction.

6.2 Surrogate Pose and Descriptor Construction

By comparing the expected and current signal distributions, we refine the pre-trained pose into a surrogate pose and use it to estimate sensing (ξ) and reliability (κ) descriptors. The three major steps are:

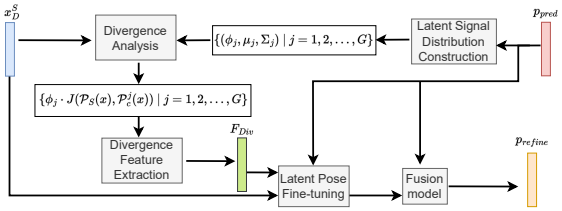


Figure 4: Surrogate pose construction module.

Step 1: As shown in Figure 4, we first construct the inputs for surrogate pose refinement by combining the approximated divergence term with features from the pre-trained model's output, while using μ_S from the signal feature x_D^S as an anchor for the distribution representation.

Step 2: The displacement $\{\phi_i J(P_S, P_c^i) \mid i = 1, \ldots, G\}$ is processed by a SetTransformer to extract the divergence feature F_{Div}, with a scaling coefficient applied beforehand to avoid overflow. The resulting intermediate fine-tuned pose feature is fused with the pre-trained model's output feature to generate the surrogate pose p_{refine}. We optimize $\mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{huber}-\text{final}} + \mathcal{R}$, where $\mathcal{L}_{\text{huber}-\text{final}}$ is the Huber loss (Huber, 1992) between the ground-truth latent pose and p_{refine}, and $\mathcal{R} = \lambda_{\text{inter}}\mathcal{L}_{\text{huber}-\text{inter}}$ regularizes the intermediate fine-tuned pose feature.

Step 3: We estimate the sensing (ξ) and reliability (κ) scores using the input signal data, pre-trained model's output, and the surrogate pose. We use a two-branch Hourglass-like architecture (Newell et al., 2016), but with linear layers instead, to estimate both the sensing score and reliability score, leveraging p_{refine}, p_{pred}, and feature of the current signal frame (extracted from a SetTransformer). The optimization process combines sensing and reliability scores, with the objective function defined as their summed MSE: $\mathcal{L}_{\text{desc}} = \mathcal{L}_{\text{MSE}-\xi} + \mathcal{L}_{\text{MSE}-\kappa}$. These estimated scores, along with the predicted pose from the pre-trained model, are fed into the downstream task to enhance application performance.

7 Evaluation

7.1 Experimental Setup

Datasets, Models and Baselines. We evaluate on three mmWave pose-estimation datasets: MARS (An & Ogras, 2021), mmBody (Chen et al., 2022), and a Self-Collected dataset. The Self-Collected dataset contains five subjects across four environments and two radar positions, captured using an AWR1843 mmWave radar. Ground-truth poses are obtained using stereo and depth cameras following (An et al., 2022); the setup is shown in Figure 5. The dataset contains nine standard activities—left/right/both lateral raises, left/right/both bicep curls, half squats, and left/right kicks—along with additional freestyle exercises. Radar signals are converted to point clouds containing 3D coordinates, Doppler velocity, and energy features following (Xue et al., 2021; Chen et al., 2022). For mmBody, we use the first 14 sequences for training, two for validation, and four for testing. The Self-Collected test set includes unseen environments and subjects to evaluate generalization. Dataset details are in Table 2.

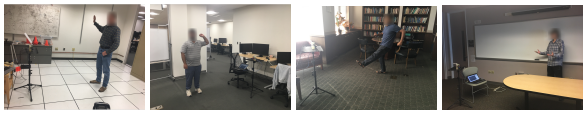


Figure 5: The Self-Collected dataset contains data from lab area, office, library, and classroom.

Datasets	Frames for Training	Frames for Validation	Frames for Testing	Joints	Radar
MARS	24,066	8,033	7,984	19	TI IWR1443
mmBody	27,821	4,125	7,946	22	Phoenix
Self-Collected	21,864	5,476	7,706	17	TI AWR1843

Table 2: Dataset details.

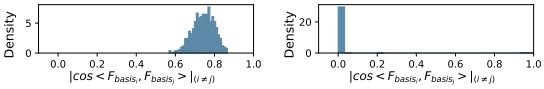
We use four state-of-the-art pose estimation models: MARS (An & Ogras, 2021), mmBody (Chen et al., 2022), mmMesh (Xue et al., 2021), and SynMotion (Zhang et al., 2022). The MARS baseline takes a single frame as input, the mmBody baseline uses five consecutive frames, and the mmMesh and SynMotion baselines use ten consecutive frames. For SynMotion, we use heatmap (Xue et al., 2021) as input, which is available in the Self-Collected dataset. We also compare mmJoints against three closely related uncertainty-estimation baselines for descriptor estimation: Monte Carlo Dropout (MCD) (Gal & Ghahramani, 2016), Deep Evidential Regression (DER) (Amini et al., 2020), and Learnable Uncertainty under Laplace Approximations (LULA) (Kristiadi et al., 2021). Each baseline receives the same radar observation window as mmJoints for the corresponding dataset. All methods are evaluated on the same evaluation samples.

Datasets	Random Basis Reconstruction Error ($\times 10^3$)	mmJoints Basis Reconstruction Error ($\times 10^3$)	CCS-based Signal Margin Difference	mmJoints Signal Margin Difference
MARS	9.35	**0.19**	7.75	**36.19**
mmBody	16.90	**2.66**	13.17	**22.77**
Self-Collected	5.03	**0.11**	7.50	**33.36**

Table 3: Pose and signal feature representation evaluation.

Configurations	MSE-mean ($\mu \times 10^{-1}$)	MSE-covariance ($\Sigma \times 10^{-3}$)	MSE-weights ($\phi \times 10^{-2}$)
KL Divergence-Based	14.50	5.59	2.03
mmJoints	**4.68**	**4.86**	**1.22**

Table 4: Evaluation of latent signal distribution construction.

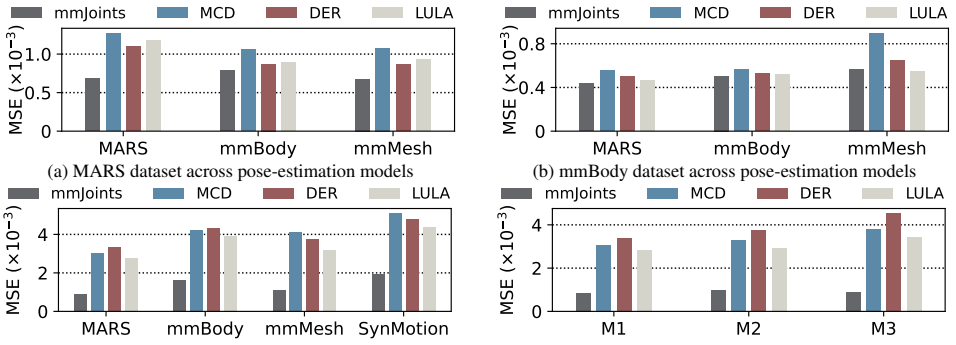


(a) Random basis (b) Our approach

Figure 6: Pose basis orthogonality in MARS dataset.

Dataset	Model	Pre-trained MSE ($\times 10^{-2}$)	Refined mmJoints MSE ($\times 10^{-2}$)
MARS	mmBody	3.81	**2.50**
mmBody	MARS	22.48	**15.38**
Self-Collected	mmMesh	6.63	**6.02**

Table 5: Evaluation of surrogate pose construction.



(a) MARS dataset across pose-estimation models (b) mmBody dataset across pose-estimation models

(c) Self-Collected dataset across pose-estimation models (d) Self-Collected dataset with stress-test model variants

Figure 7: Descriptor estimation accuracy across datasets. Each subfigure fixes one dataset and compares mmJoints with MCD, DER, and LULA baselines across different pose-estimation models.

Configurations and Metrics. We use $D_p = 32$, $D_S = 64$, $\bar{D}_0 = 20$, and $D_C = 129$, where the sizes of μ_S and σ_S are both 64. K is determined based on the sampling rate of the radar configuration and corresponds to half-second intervals of the signal. The Adam optimization algorithm is used for training. The evaluation metrics are defined as: Mean Per Joint Position Error (MPJPE) (An et al., 2022), Mean Absolute Error (MAE) (An & Ogras, 2021), Mean Squared Error (MSE) (Bishop & Nasrabadi, 2006), Correct Joint Percentage (PCK@0.5) (Badiola-Bengoa & Mendez-Zorrilla, 2021), Signal Degradation Scores (Ning et al., 2018; Rusu et al., 2008), Basis Reconstruction Error (representation capability of P_{basis}), and Margin Difference (representation capability of signal feature μ_S, higher is better). These metrics are detailed in Appendix H.

7.2 Component Analysis

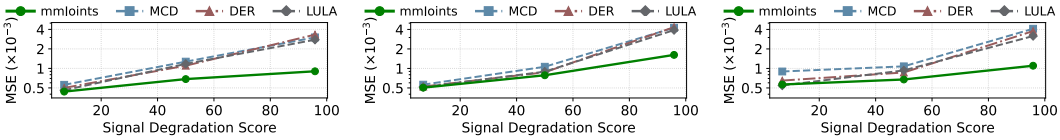
Signal and Pose Encoding. We evaluate the representation quality of the pose and signal encoders. As shown in Table 3, the mmJoints pose basis achieves a 49.07× lower reconstruction error than a random basis on MARS. Figure 6 further shows strong orthogonality, with pairwise basis-vector cosine values concentrated near zero. For signal encoding, we compare against the encoder of (An & Ogras, 2021) using only coarse cluster-wise structure (CCS); our method yields a 4.67× larger margin difference. Similar trends across datasets demonstrate the effectiveness of both encoders.

Latent Signal Distribution Construction. We compare our latent distribution construction with a baseline using the same architecture but directly optimizing the KL divergence between the current signal feature and target distribution from the pose feature, without modeling component-wise internal structure. Using SynMotion as the pre-trained pose estimator on the Self-Collected dataset, Table 4 shows improvements of 67.72%, 13.12%, and 39.85% in estimating the GMM mean, covariance, and weights, respectively. Further ablations are provided in Section 7.4.

Surrogate Pose Construction. We compare our surrogate pose construction with directly encoding the pre-trained model output as the surrogate pose. As shown in Table 5, on MARS with mmBody as the pre-trained model, our method reduces the surrogate-pose error by 34.30%. Similar trends hold across other configurations. Its impact is further analyzed in Section 7.4.

7.3 System Evaluation

mmJoints on State-of-the-Art Models. Figure 7 (a)-(c) compares descriptor-estimation MSE against MCD, DER, and LULA across datasets and pose-estimation models. mmJoints consistently



(a) MARS model (b) mmBody model (c) mmMesh model

Figure 8: Descriptor estimation robustness under increasing Signal Degradation Score, comparing MMJOINTS with MCD, DER, and LULA baselines across pose-estimation models.

Model Input	MSE ($\times 10^{-4}$)
p_{pred}	28.89
p_{pred} + signal	9.09
p_{pred} + signal + p_{refine}	**6.84**

(a) Using p_{refine} and signal features improves descriptor estimation.

Configurations	Pre-trained	Pre-trained + Divergence	Pre-trained + Divergence + Signal Feature
MSE ($\times 10^{-2}$)	9.66	4.41	**2.67**

(b) Component analysis showing how distributional training improves surrogate pose construction.

Table 6: Ablation evaluation of component impact.

achieves low estimation error, with particularly clear advantages on the Self-Collected dataset, reducing MSE by 55.49%–67.40% relative to the strongest uncertainty baseline across MARS, mmBody, mmMesh, and SynMotion. On mmBody, where all methods already achieve low estimation error, MMJOINTS maintains consistently strong performance across pose-estimation models.

MMJOINTS under Degraded Base-Model Quality. We further stress-test MMJOINTS using three deliberately constrained MARS-based models: M1 is trained only for upper-body pose estimation, M2 only for lower-body pose estimation, and M3 uses randomly initialized, untrained parameters. Across all three stress-test configurations, as shown in Figure 7 (d), MMJOINTS achieves the lowest descriptor-estimation error, reducing MSE by 66.65%–73.49% relative to the strongest uncertainty baseline. These results demonstrate that MMJOINTS remains effective across a wide range of base-estimator quality, including severely degraded predictions.

Robustness to Input Degradation. Figure 8 evaluates descriptor estimation across datasets with increasing Signal Degradation Scores, with each subfigure using the same pose-estimation model. As radar inputs become sparse and noisy due to variations in device configuration, placement, sensing geometry, and surrounding environments, the advantage of MMJOINTS over uncertainty baselines becomes increasingly pronounced. At the lowest degradation level, MMJOINTS performs comparably to the strongest baseline, while reducing MSE by up to 37.89% at the intermediate level and by 58.55%–67.40% at the highest level. This widening improvement demonstrates that MMJOINTS remains particularly effective under increasingly challenging sensing conditions.

7.4 ABLATION STUDY

Ablation on Surrogate Descriptor Estimation. We evaluate the impact of the surrogate pose p_{refine} and current signal features on descriptor estimation using the MARS model and dataset. We compare three settings: ① using only p_{pred}; ② using p_{pred} together with signal features; and ③ the proposed approach using p_{pred}, p_{refine}, and signal features. As shown in Table 6 (a), adding signal features reduces the error by 68.53% over ①, while further incorporating p_{refine} yields an additional 24.76% MSE reduction over ②.

Ablation on Surrogate Pose Construction. We evaluate the impact of latent signal distribution and signal features on surrogate pose construction. As shown in Table 6 (b), incorporating latent signal distribution information reduces MSE by 54.35% compared with directly using the pre-trained pose features, while further adding signal features yields an additional 39.46% reduction.

8 APPLICATION EVALUATION

Pose Estimation. We apply MMJOINTS to improve pre-trained pose estimators by concatenating the estimated surrogate descriptors with their outputs and feeding them into a lightweight MLP-based downstream model, adjusting only its input layer for fairness. Across four state-of-the-art and three stress-test models on the Self-Collected dataset, MMJOINTS consistently reduces MPJPE and MAE while improving PCK@0.5 in both cross-environment and cross-subject settings, with larger gains for the stress-test models (up to $6.42\times$ reduction in MPJPE) and consistent improvements for

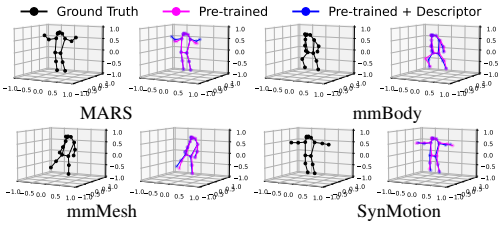


Figure 9: Improved pose estimation examples on the self-collected dataset across four baselines.

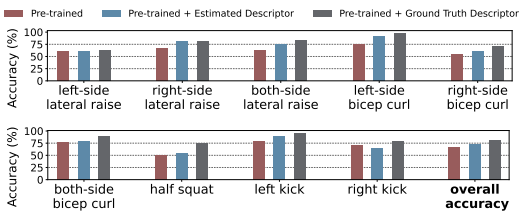


Figure 10: Accuracy across activities in cross-subject settings.

state-of-the-art models. Detailed quantitative results are provided in Appendix E. Figure 9 further shows representative examples across four state-of-the-art baselines, where upper- and lower-limb movements are refined to more precise positions after incorporating the descriptor information.

Activity Recognition. We evaluate MMJOINTS for activity recognition in health rehabilitation using an LSTM-based downstream classifier over five consecutive poses estimated by the pre-trained MARS model. As shown in Figure 10, incorporating the estimated descriptors improves overall cross-subject accuracy by 6.49 percentage points over predicted poses alone, with gains of up to 16.07 percentage points for individual activities. The largest improvements occur for high-variability activities such as both-side lateral raises, where sensing support and reliability provide complementary information beyond the predicted joint coordinates. Performance remains stable for lower-variability activities, resulting in a substantial net improvement across the full test set. Additional quantitative results are provided in Appendix F.

9 Related Work

mmWave-based Applications. mmWave-based sensing applications, such as gesture and activity recognition (Singh et al., 2019; Jiang et al., 2018; Xie et al., 2022; Li et al., 2022), human counting (Li & Hishiyama, 2023; Ren et al., 2023), tracking (Dai et al., 2023; Huang et al., 2021), detection (Cui & Dahnoun, 2021), biometric imaging (Meng et al., 2020; Yang et al., 2020), pose estimation (An & Ogras, 2021; Xue et al., 2021), and sensor fusion (Li et al., 2020; Gadre et al., 2022), have advanced rapidly. We aim to characterize pre-trained pose estimators by expanding the representation of the output joint explicitly, enhancing system reliability and trustworthiness.

mmWave-based Human Pose Estimation. Recent work in mmWave-based human pose estimation has utilized various learning-based methods to generate human pose skeletons or meshes from radar signals (Xue et al., 2021; An & Ogras, 2021; Chen et al., 2022; Zhang et al., 2022), focusing on learning the pattern of single (An & Ogras, 2021) or multiple (Xue et al., 2021; Chen et al., 2022; Zhang et al., 2022) signal frames. To explicitly characterize mmWave-based pose estimation models, we introduce descriptors that offer knowledge-level interpretability while also enhancing the precision of state-of-the-art methods and downstream tasks.

Explainable Representation in Pose Estimation. Several explainable AI methods are proposed for human pose estimation using different input modalities. For vision-based tasks, existing methods (Dittakavi et al., 2022; Yang et al., 2021) highlight joints that preserve specific dependencies using RGB image as input. For mmWave-based tasks, methods either leverage implicit information (e.g., body parts) to improve pose estimation (Li et al., 2023; Cao et al., 2022) or generate explicit body-part explanations for better downstream tasks (Wang et al., 2023). However, they lack explicit interpretability for arbitrary black-box models' outputs.

10 Conclusion

We propose MMJOINTS, a framework that exposes the hidden sensing support and prediction reliability of pre-trained, black-box mmWave pose estimators through per-joint descriptors. By augmenting each predicted joint with sensing and reliability information, MMJOINTS provides a richer, actionable representation without accessing or modifying the estimator, consistently outperforming uncertainty-estimation baselines under challenging sensing conditions and improving downstream pose estimation and activity recognition across diverse datasets, models and configurations.

AI USE STATEMENT

Generative AI tools were used to assist with proofreading, language polishing, rephrasing, and improving the clarity and readability of the manuscript. They were also used for limited assistance with code editing and debugging. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was reviewed and tested for correctness. We independently verified the experimental results, analyses, and claims presented in this paper and take full responsibility for the final content of the work.

ETHICS STATEMENT

We follow the ICLR Code of Ethics. The collection of our dataset involving human participants was conducted under institutional IRB approval. Our study does not use private, sensitive, or personally identifiable information, and we do not identify any specific ethical concerns or potential harmful societal impacts arising from this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the MMJOINTS implementation and scripts/configurations used for the reported experiments in the supplementary materials, together with hyperparameter settings and execution instructions.

REFERENCES

Mubarak A Alanazi, Abdullah K Alhazmi, Osama Alsattam, Kara Gnau, Meghan Brown, Shannon Thiel, Kurt Jackson, and Vamsy P Chodavarapu. Towards a low-cost solution for gait analysis using millimeter wave sensor and machine learning. *Sensors*, 22(15):5470, 2022.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

Sizhe An and Umit Y Ogras. Mars: mmwave-based assistive rehabilitation system for smart healthcare. *ACM Transactions on Embedded Computing Systems (TECS)*, 20(5s):1–22, 2021.

Sizhe An and Umit Y Ogras. Fast and scalable human pose estimation using mmwave point cloud. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, pp. 889–894, 2022.

Sizhe An, Yin Li, and Umit Ogras. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in Neural Information Processing Systems*, 35: 27414–27426, 2022.

Aritz Badiola-Bengoa and Amaia Mendez-Zorrilla. A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise. *Sensors*, 21(18):5996, 2021.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, pp. 3, 2016.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

Zhongping Cao, Wen Ding, Rihui Chen, Jianxiong Zhang, Xuemei Guo, and Guoli Wang. A joint global–local network for human pose estimation with millimeter wave radar. *IEEE Internet of Things Journal*, 10(1):434–446, 2022.

Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3501–3510, 2022.

Han Cui and Naim Dahnoun. High precision human detection and tracking using millimeter-wave radars. *IEEE Aerospace and Electronic Systems Magazine*, 36(1):22–32, 2021.

James R Dabrowski and Ethan V Munson. Is 100 milliseconds too fast? In *CHI'01 extended abstracts on Human factors in computing systems*, pp. 317–318, 2001.

Yimin Dai, Xian Shuai, Rui Tan, and Guoliang Xing. Interpersonal distance tracking with mmwave radar and imus. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*, pp. 123–135, 2023.

Bhat Dittakavi, Divyagna Bavikadi, Sai Vikas Desai, Soumi Chakraborty, Nishant Reddy, Vineeth N Balasubramanian, Bharathi Callepalli, and Ayon Sharma. Pose tutor: an explainable system for pose correction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3540–3549, 2022.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Hehe Fan, Yi Yang, and Mohan Kankanhalli. 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*, pp. 14204–14213, 2021.

Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9919–9928, 2021.

Akshay Gadre, Deepak Vasisht, Nikunj Raghuvanshi, Bodhi Priyantha, Manikanta Kotaru, Swarun Kumar, and Ranveer Chandra. Milton: Sensing product integrity without opening the box using non-invasive acoustic vibrometry. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 390–402. IEEE, 2022.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Ivan Grishchenko, Valentin Bazarevsky, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, Richard Yee, Karthik Raveendran, Matsvei Zhdanovich, Matthias Grundmann, and Cristian Sminchisescu. Blazepose ghum holistic: Real-time 3d human landmarks and pose estimation. *arXiv preprint arXiv:2206.11678*, 2022.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Xu Huang, Hasnain Cheena, Abin Thomas, and Joseph KP Tsoi. Indoor detection and tracking of people using mmwave sensor. *Journal of Sensors*, 2021(1):6657709, 2021.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. Towards environment independent device free human activity recognition. In *Proceedings of the 24th annual international conference on mobile computing and networking*, pp. 289–304, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. m3track: mmwave-based multi-user 3d posture tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pp. 491–503, 2022.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Learnable uncertainty under laplace approximations. In *Uncertainty in Artificial Intelligence*, pp. 344–353. PMLR, 2021.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Shenglei Li and Reiko Hishiyama. Counting and tracking people to avoid from crowded in a restaurant using mmwave radar. *IEICE TRANSACTIONS on Information and Systems*, 106(6):1142–1154, 2023.

Wenwei Li, Ruofeng Liu, Shuai Wang, Dongjiang Cao, and Wenchao Jiang. Egocentric human pose estimation using head-mounted mmwave radar. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, pp. 431–444, 2023.

Yadong Li, Dongheng Zhang, Jinbo Chen, Jinwei Wan, Dong Zhang, Yang Hu, Qibin Sun, and Yan Chen. Towards domain-independent and real-time gesture recognition using mmwave signal. *IEEE Transactions on Mobile Computing*, 22(12):7355–7369, 2022.

Yang Li, Yutong Liu, Yanping Wang, Yun Lin, and Wenjie Shen. The millimeter-wave radar slam assisted by the rcs feature of the target and imu. *Sensors*, 20(18):5421, 2020.

Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kunpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 4(4):1–28, 2020.

Yilin Liu, Shijia Zhang, Mahanth Gowda, and Srihari Nelakuditi. Leveraging the properties of mmwave signals for 3d finger motion tracking for interactive iot applications. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(3):1–28, 2022.

Silvano Martello and Paolo Toth. Linear assignment problems. In *North-Holland Mathematics Studies*, volume 132, pp. 259–282. Elsevier, 1987.

Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pp. 506–516. IEEE, 2017.

Zhen Meng, Song Fu, Jie Yan, Hongyuan Liang, Anfu Zhou, Shilin Zhu, Huadong Ma, Jianhua Liu, and Ning Yang. Gait recognition for co-existing multiple people using millimeter wave sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 849–856, 2020.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pp. 483–499. Springer, 2016.

Xiaojuan Ning, Fan Li, Ge Tian, and Yinghui Wang. An efficient outlier removal method for scattered point cloud data. *PloS one*, 13(8):e0201280, 2018.

Karthik Ramasubramanian. mmwave radar for automotive and industrial applications. *Texas Instruments*, 2017.

Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Orthogonal projection loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12333–12343, 2021.

Liyuan Ren, Alexander G Yarovoy, and Francesco Fioranelli. Grouped people counting using mm-wave fmcw mimo radar. *IEEE Internet of Things Journal*, 10(22):20107–20119, 2023.

Mary Rosson and John M Carroll. *Usability engineering*, volume 13. Elsevier, 2001.

Radu Bogdan Rusu, Zoltan Csaba Marton, Nico Blodow, Mihai Dolha, and Michael Beetz. Towards 3d point cloud based object maps for household environments. *Robotics and Autonomous Systems*, 56(11):927–941, 2008.

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.

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*, pp. 51–56, 2019.

Texas Instruments. Dca1000evm. https://www.ti.com/tool/dca1000evm, 2026a. URL `https://www.ti.com/tool/DCA1000EVM`.

Texas Instruments. Ti awr1843. https://www.ti.com/product/awr1843, 2026b. URL `https://www.ti.com/product/AWR1843`.

Scott I Vrieze. Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological methods*, 17(2):228, 2012.

Shuai Wang, Dongjiang Cao, Ruofeng Liu, Wenchao Jiang, Tianshun Yao, and Chris Xiaoxuan Lu. Human parsing with joint learning for dynamic mmwave radar point cloud. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(1):1–22, 2023.

Yucheng Xie, Ruizhe Jiang, Xiaonan Guo, Yan Wang, Jerry Cheng, and Yingying Chen. Universal targeted attacks against mmwave-based human activity recognition system. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pp. 541–542, 2022.

Weiye Xu, Wenfan Song, Jianwei Liu, Yajie Liu, Xin Cui, Yuanqing Zheng, Jinsong Han, Xinhuai Wang, and Kui Ren. Mask does not matter: Anti-spoofing face authentication using mmwave without on-site registration. In *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking*, pp. 310–323, 2022a.

Xiangyu Xu, Jiadi Yu, Chengguang Ma, Yanzhi Ren, Hongbo Liu, Yanmin Zhu, Yi-Chao Chen, and Feilong Tang. mmecg: Monitoring human cardiac cycle in driving environments leveraging millimeter wave. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 90–99. IEEE, 2022b.

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*, pp. 269–282, 2021.

Hongfei Xue, Qiming Cao, Yan Ju, Haochen Hu, Haoyu Wang, Aidong Zhang, and Lu Su. M4esh: mmwave-based 3d human mesh construction for multiple subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pp. 391–406, 2022.

Hongfei Xue, Qiming Cao, Chenglin Miao, Yan Ju, Haochen Hu, Aidong Zhang, and Lu Su. Towards generalized mmwave-based human pose estimation through signal augmentation. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pp. 1–15, 2023.

Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11802–11812, 2021.

Xin Yang, Jian Liu, Yingying Chen, Xiaonan Guo, and Yucheng Xie. Mu-id: Multi-user identification through gaits using millimeter wave radios. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pp. 2589–2598. IEEE, 2020.

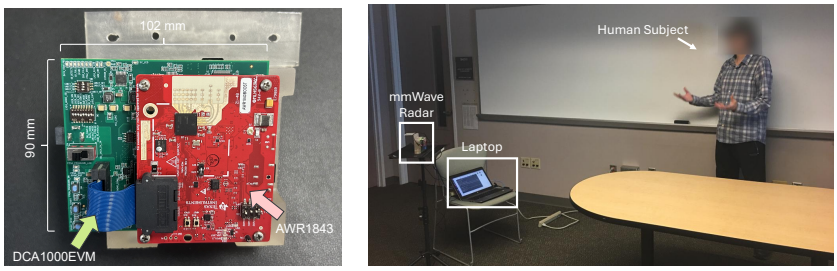
Jia Zhang, Rui Xi, Yuan He, Yimiao Sun, Xiuzhen Guo, Weiguo Wang, Xin Na, Yunhao Liu, Zhenguo Shi, and Tao Gu. A survey of mmwave-based human sensing: Technology, platforms and applications. *IEEE Communications Surveys & Tutorials*, 2023.

Xiaotong Zhang, Zhenjiang Li, and Jin Zhang. Synthesized millimeter-waves for human motion sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pp. 377–390, 2022.

Kai Zheng, Wuqiong Zhao, Timothy Woodford, Renjie Zhao, Xinyu Zhang, and Yingbo Hua. Enhancing mmwave radar sensing using a phased-mimo architecture. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*, pp. 56–69, 2024.

Model	Dataset	Correct		Incorrect	
		Not Sensed	Sensed	Not Sensed	Sensed
MARS (CNN-depth 6)	MARS	55,884	30,110	45,649	20,053
MARS (CNN-depth 4)	MARS	48,657	24,118	52,876	26,045
MARS (CNN-depth 6)	mmBody	114,867	17,338	34,145	8,462
mmBody (P4T-2 frames)	MARS	75,837	39,532	25,111	10,437
mmBody (P4T-5 frames)	MARS	86,588	43,177	12,805	6,010
mmBody (P4T-2 frames)	mmBody	117,449	16,505	31,482	9,288
mmBody (P4T-5 frames)	mmBody	118,169	17,046	30,523	8,722
mmMesh (3 frames)	MARS	76,249	35,761	25,295	14,391
mmMesh (10 frames)	MARS	90,552	43,317	11,000	6,827
mmMesh (3 frames)	mmBody	121,194	17,395	27,818	8,405
mmMesh (10 frames)	mmBody	119,432	16,818	29,580	8,982

Table 7: Distribution of joint types across models and datasets.



(a) mmWave radar hardware (b) MMJOINTS system setup

Figure 11: Radar device and MMJOINTS system setup.

A HARDWARE SETUP

The mmWave radar system for human pose estimation (Xue et al., 2021; 2023; An & Ogras, 2021) typically consists of a single-chip radar sensor and a data capture card. For our setup, we use a TI AWR1843 (Texas Instruments, 2026b) radar, featuring four receiving and three transmitting antennas, paired with a DCA1000EVM (Texas Instruments, 2026a) data capture card to collect our Self-Collected dataset. The radar emits millimeter-wave signals and processes reflections to detect object positions and motions. The DCA1000EVM enables raw signal acquisition from the sensor to a host device for processing and development. Figure 11 shows the radar hardware configuration, module dimensions, and MMJOINTS system setup.

B MODEL BIASES

We study bias in MARS (An & Ogras, 2021), mmBody (Chen et al., 2022), and mmMesh (Xue et al., 2021) across public datasets. Table 7 shows substantial mismatch between sensing and prediction correctness: on MARS, 66.76% of correct predictions are not sensed, while 32.66% of incorrect predictions are sensed, with similar trends on mmBody. This mismatch is consistent with reliance on learned priors and shows that prediction correctness alone does not reveal signal support.

C WHY SIGNAL DISTRIBUTION?

Signal distribution exhibits a stronger relationship to pose when examined via covariance analysis. This justifies our inverse modeling approach, where we map pose $\rightarrow$ signal distribution.

Signal Instance vs Pose. We systematically study the pose estimation problem via covariance analysis (Efron & Tibshirani, 1994) where we quantify the correlation strength of the changes to the signal, Δx and the corresponding changes to the ground truth pose, Δp. A larger absolute covariance between Δp and Δx indicates a stronger dependency between x and p, which is relatively easier to model. Figure 12 compares the distribution of $|Cov(\Delta p, \Delta x)|$ for four mmWave-based and two computer vision-based pose estimation baselines (An & Ogras, 2021; Chen et al., 2022; An et al., 2022; Mehta et al., 2017; Fieraru et al., 2021). The vertical lines represent the mean. For consistency across data domains, Δx is computed using the l_2-norm of input signal vectors before they are fed to

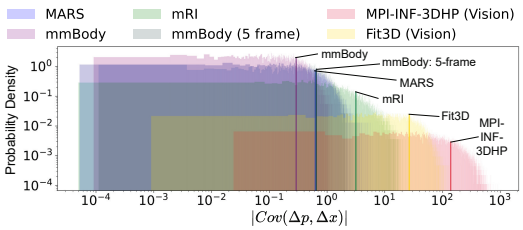


Figure 12: Covariance analysis: signal vs. pose.

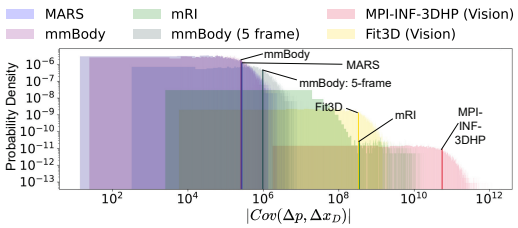


Figure 13: Covariance analysis: signal distribution vs. pose.

the pose estimator; Δp denotes the Euclidean distance between two poses. We observe that mmWave baselines are scattered around the left side of the plot, while the computer vision baselines coincide on the right.

This stark difference between vision and mmWave baselines can be attributed to the nature of signal reflections. mmWave, when reflected off the human body, are sparse, sporadic, and inconsistent. In contrast, light uniformly illuminates the body, resulting in dense, continuous, consistent reflections. As a result, mmWave radars face significantly greater challenges in achieving the same level of performance as cameras in accomplishing human sensing tasks.

Signal Distribution vs Pose. We can lower the modeling complexity by reformulating the problem. We replace signal instance, x with its distribution, x_D and study $|Cov(\Delta x_D, \Delta p)|$. We redo the covariance analysis for this new formulation to observe the dependency of changes in the ground truth pose, Δp and the corresponding signal distribution change, Δx_D. To obtain the distribution x_D from a collection of signal instances given a pose p, we first apply PCA to reduce the number of correlated features in the raw signal space and then fit this to a GMM. Symmetric KL-divergence (Jeffreys, 1946) is used to compute Δx_D.

Figure 13 shows that the mean $|Cov(\Delta x_D, \Delta p)|$ for both vision and mmWave baselines overlap at the right end with larger variance values. *This indicates that the signal distribution and pose maintain a stronger relationship compared to the association between pose and signal instances.* Despite their strong relationship, at runtime, a standalone radar system cannot sense the signal distribution directly. While some works attempt to approximate this by accumulating consecutive frames (An & Ogras, 2022; An et al., 2022), these frames are biased by their fixed location and do not produce high-quality signal distributions. To address this limitation, we introduce latent signal distribution in MMJOINTS, with the details of its construction provided in Section 5.2.

D TRAINING AND INFERENCE TIME

Training of MMJOINTS is performed offline and does not affect deployment latency. On the Self-Collected dataset (containing 35,046 signal instances), feature modeling takes 7.5 hours, while signal distribution construction, surrogate pose construction, and descriptor estimation require an additional 25 hours on a GTX 1080 GPU and an i7-7800X CPU.

At deployment, MMJOINTS incurs an average inference overhead of 62.90 ms per frame on a Jetson Orin Nano. Together with the base estimator execution time of 26.86 ms, the complete pipeline achieves an end-to-end latency of approximately 90 ms per frame. This latency supports real-time operation on edge hardware: it falls below the ≈100–200 ms near-instantaneous feedback regime identified in classic HCI guidance (Rosson & Carroll, 2001; Dabrowski & Munson, 2001) and is consistent with the 10–30 Hz operating range commonly used by pose-driven pipelines (Cao et al., 2019; Grishchenko et al., 2022) for applications such as indoor activity recognition and gesture-based control.

E APPLICATION—POSE ESTIMATION

Table 8 shows the results of all seven models on the Self-Collected dataset before and after applying MMJOINTS in cross-environment and cross-subject evaluations. Integrating the estimated surrogate descriptors generated by MMJOINTS consistently reduces joint localization error across all models and evaluation settings. For the stress-test models, MMJOINTS yields particularly large gains, reducing

Model	MPJPE (cm) before / after	MAE (cm) before / after	PCK@0.5 (%) before / after
MARS	9.36 / **9.06**	4.42 / **4.28**	72.24 / **73.98**
mmBody	9.23 / **8.89**	4.38 / **4.22**	72.92 / **75.78**
mmMesh	8.07 / **7.83**	3.81 / **3.70**	79.29 / **80.99**
SynMotion	11.69 / **11.21**	5.65 / **5.38**	55.81 / **59.79**
M_1	23.43 / **9.11**	10.55 / **4.34**	47.10 / **73.18**
M_2	41.71 / **8.84**	17.86 / **4.19**	27.49 / **75.33**
M_3	55.94 / **8.71**	24.14 / **4.19**	1.17 / **76.10**

(a) Evaluation in cross-environmental settings.

Model	MPJPE (cm) before / after	MAE (cm) before / after	PCK@0.5 (%) before / after
MARS	10.32 / **10.19**	4.86 / **4.83**	61.34 / **62.31**
mmBody	9.59 / **9.33**	4.59 / **4.48**	64.03 / **65.88**
mmMesh	8.59 / **8.50**	4.15 / **4.08**	70.68 / **70.97**
SynMotion	11.06 / **10.75**	5.19 / **5.03**	55.12 / **57.25**
M_1	23.43 / **10.13**	10.43 / **4.79**	39.73 / **62.44**
M_2	40.42 / **9.64**	17.15 / **4.58**	24.22 / **65.21**
M_3	53.90 / **9.69**	22.94 / **4.61**	1.47 / **65.18**

(b) Evaluation in cross-subject settings.

Table 8: MMJOINTS improves pose-estimation performance across different settings.

Configurations	Pre-trained	Pre-trained + Estimated Descriptor	Pre-trained + Ground Truth Descriptor
Accuracy (%)	55.83	61.00	72.86

(a) Overall accuracy comparison across different baselines.

Joint Type	Correct		Incorrect	
	Not Sensed	Sensed	Not Sensed	Sensed
#Joint	13,852	6,564	3,860	1,904

(b) Test cases where estimated descriptors correct mispredictions.

Table 9: Evaluation in cross-environmental settings.

MPJPE by up to $6.42\times$ and bringing their performance into the range of state-of-the-art models. At the same time, MMJOINTS consistently improves the four state-of-the-art pre-trained models across both evaluation settings.[1] A similar trend is observed in MAE and PCK@0.5.

We also present representative examples of pose instances involving upper and lower limb movements across four state-of-the-art baselines, as shown in Figure 9. After integrating descriptor information into this downstream task, we observe that the featured movements (e.g., arm movement in the MARS example and leg movement in the mmMesh example) are refined to more precise positions, resulting in more accurate pose estimation.

F Application—Activity Recognition

We further evaluate the accuracy improvements in unseen environmental settings with more complex indoor fixtures. As shown in Table 9 (a), the estimated descriptor improves activity recognition accuracy by 5.17 percentage points, while the ground truth descriptor achieves even higher accuracy of 72.86%, as expected. This result further demonstrates the effectiveness and generalizability of MMJOINTS in activity recognition.

We also investigate the distribution of joint types to understand the increase in accuracy. In Table 9 (b), we observe a considerable number of joints with limited signal support (not sensed) in both correctly and incorrectly predicted categories. In this evaluation, 67.85% of correctly predicted joints and 66.97% of incorrectly predicted joints are not sensed. The high proportion of unsensed joints in both categories highlights how the estimated descriptors identify latent patterns, supporting our findings in Section 2. We further examine the contribution of each descriptor in activity recognition. Adding ξ improves accuracy from 66.24% to 69.87% (+3.63 percentage points), while further incorporating κ increases it to 72.73% (+2.86 percentage points).

G Loss Function Details

We provide detailed definitions of the loss terms used in the main text. Let B denote the batch size, D_p and D_S the pose and signal latent dimensions, respectively, and J the number of body joints.

Gaussian-Prior Regularization. Both the pose and signal encoders parameterize their latent representations using a mean μ and standard deviation σ. We regularize these representations using

$$\mathcal{L}_{\mathrm{G}}(\mu, \sigma) = \frac{1}{B} \sum_{b=1}^{B} \sum_{d} \left[\sigma_{b,d}^2 + \mu_{b,d}^2 - \log \sigma_{b,d} - C \right], \tag{1}$$

where C is a normalization constant that sets the reference level of the regularization loss, avoiding an arbitrary offset in its magnitude; since it is independent of the learned parameters, it does not affect the gradients or the optimum.

[1] Because MPJPE averages error across all 17 joints, aggregate improvements can understate gains concentrated on individual joints. For example, for MARS, MMJOINTS achieves a 12.52% relative improvement on the left wrist while also reducing the overall MPJPE.

Pose Encoding. For the pose VAE, let p_b and $\hat{p}_b$ denote the input and reconstructed pose of sample b. The reconstruction loss is

$$\mathcal{L}_{\mathrm{REC}} = \frac{1}{B}\sum_{b=1}^{B} \|p_b - \hat{p}_b\|_2^2 . \tag{2}$$

The pose-encoding objective is therefore

$$\mathcal{L}_{\mathrm{P}_{\mathrm{step2}}} = \mathcal{L}_{\mathrm{REC}} + \lambda_{\mathrm{step2}}\mathcal{L}_{\mathrm{G}}(\mu_p, \sigma_p). \tag{3}$$

To construct an orthogonal pose basis, we further apply Orthogonal Projection Loss (OPL). Let $\mathcal{S}$ and $\mathcal{D}$ denote pairs of normalized latent features from the same and different pose clusters, respectively. We use

$$\begin{aligned} \mathcal{L}_{\mathrm{OPL-s}} &= 1 - \frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}} \langle f_i, f_j\rangle, \\ \mathcal{L}_{\mathrm{OPL-d}} &= \left| \frac{1}{|\mathcal{D}|}\sum_{(i,j)\in\mathcal{D}} \langle f_i, f_j\rangle \right|. \end{aligned} \tag{4}$$

The corresponding objective is

$$\mathcal{L}_{\mathrm{P}_{\mathrm{step3}}} = \mathcal{L}_{\mathrm{OPL-d}} + \lambda_{\mathrm{step3}}\mathcal{L}_{\mathrm{OPL-s}}, \tag{5}$$

encouraging inter-cluster orthogonality while preserving intra-cluster compactness.

Signal Encoding. For an anchor signal feature f_a, a positive feature f_p associated with a similar pose, and a negative feature f_n associated with a dissimilar pose, the triplet term is

$$\mathcal{L}_{\mathrm{triplet}} = \frac{1}{B}\sum_{b=1}^{B} \Big[d(f_a^{(b)}, f_p^{(b)}) - d(f_a^{(b)}, f_n^{(b)}) + m \Big]_+ , \tag{6}$$

where $d(\cdot,\cdot)$ denotes the latent-space distance, m is the margin, and $[x]_+ = \max(x, 0)$.

The cluster-level cross-entropy term is

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{B}\sum_{b=1}^{B} \log q_b(y_b), \tag{7}$$

where $q_b(y_b)$ is the predicted probability of the pose-cluster label y_b. The complete signal-encoding objective is

$$\mathcal{L}_S = \lambda_0\mathcal{L}_{\mathrm{triplet}} + \lambda_1\mathcal{L}_{\mathrm{CE}} + \lambda_2\mathcal{L}_{\mathrm{G}}(\mu_S, \sigma_S). \tag{8}$$

Latent Signal Distribution Construction. Let a predicted GMM component be $\hat{\Gamma}_i = (\hat{\phi}_i, \hat{\mu}_i, \hat{\Sigma}_i)$ and its target component be $\Gamma_j = (\phi_j, \mu_j, \Sigma_j)$. Their component-wise matching cost is

$$\begin{aligned} T(\hat{\Gamma}_i, \Gamma_j) &= \mathcal{L}_{l_1-\phi} + \mathcal{L}_{l_1-\mu} + \mathcal{L}_{l_1-\Sigma}, \\ \mathcal{L}_{l_1-\phi} &= |\hat{\phi}_i - \phi_j|, \\ \mathcal{L}_{l_1-\mu} &= \|\hat{\mu}_i - \mu_j\|_1, \\ \mathcal{L}_{l_1-\Sigma} &= \|\hat{\Sigma}_i - \Sigma_j\|_1. \end{aligned} \tag{9}$$

Because the ordering of GMM components is arbitrary, we formulate a permutation-invariant Linear Assignment Problem:

$$\mathcal{L}_{\mathrm{LAP}} = \min_{\{a_{ij}\}} \sum_{i=1}^{G}\sum_{j=1}^{G} T(\hat{\Gamma}_i, \Gamma_j) a_{ij}, \tag{10}$$

subject to $\sum_j a_{ij} = 1$, $\sum_i a_{ij} = 1$, and $a_{ij} \in \{0, 1\}$. The assignment is solved using the Hungarian algorithm.

We additionally constrain the generated distribution according to its discrepancy from the observed signal distribution. Let $P_S = \mathcal{N}(\mu_S, \Sigma_S)$ denote the observed signal feature and $P_c^i = \mathcal{N}(\mu_i, \Sigma_i)$ the i-th component of the generated GMM. The divergence regularizer is

$$\mathcal{L}_{\text{Div}} = \hat{J}(P_S, P_G) = \sum_{i=1}^{G} \phi_i \left[D_{\text{KL}}(P_S \| P_c^i) + D_{\text{KL}}(P_c^i \| P_S) \right], \tag{11}$$

corresponding to the weighted Jeffreys-divergence approximation described in Section 6. The final distribution-construction objective is

$$\mathcal{L}_{\text{Distribution}} = \mathcal{L}_{\text{LAP}} + \lambda_{\text{Div}} \mathcal{L}_{\text{Div}}. \tag{12}$$

Surrogate Pose Construction. Let p^{gt} denote the latent feature of the ground-truth pose, p^{inter} the intermediate fine-tuned pose feature, and p^{refine} the final surrogate pose feature. For residual e, the Huber penalty is

$$h_\delta(e) = \begin{cases} \frac{1}{2} e^2, & |e| \le \delta, \\ \delta \left(|e| - \frac{1}{2}\delta \right), & |e| > \delta, \end{cases} \tag{13}$$

where δ controls the transition between the quadratic and linear regions. We define

$$\begin{aligned} \mathcal{L}_{\text{huber-final}} &= \frac{1}{BD_p} \sum_{b,d} h_\delta \left(p_{b,d}^{refine} - p_{b,d}^{gt} \right), \\ \mathcal{L}_{\text{huber-inter}} &= \frac{1}{BD_p} \sum_{b,d} h_\delta \left(p_{b,d}^{inter} - p_{b,d}^{gt} \right). \end{aligned} \tag{14}$$

The surrogate-pose objective is

$$\mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{huber-final}} + \lambda_{\text{inter}} \mathcal{L}_{\text{huber-inter}}, \tag{15}$$

where the intermediate term regularizes the pose representation before feature fusion.

Surrogate Descriptor Generation. For predicted sensing and reliability scores $\hat{\xi}_{b,j}$ and $\hat{\kappa}_{b,j}$ and their corresponding training targets $\xi_{b,j}$ and $\kappa_{b,j}$, we use

$$\begin{aligned} \mathcal{L}_{\text{MSE}-\xi} &= \frac{1}{BJ} \sum_{b=1}^{B} \sum_{j=1}^{J} \left(\hat{\xi}_{b,j} - \xi_{b,j} \right)^2, \\ \mathcal{L}_{\text{MSE}-\kappa} &= \frac{1}{BJ} \sum_{b=1}^{B} \sum_{j=1}^{J} \left(\hat{\kappa}_{b,j} - \kappa_{b,j} \right)^2. \end{aligned} \tag{16}$$

The descriptor objective equally combines the two terms:

$$\mathcal{L}_{\text{desc}} = \mathcal{L}_{\text{MSE}-\xi} + \mathcal{L}_{\text{MSE}-\kappa}. \tag{17}$$

H Evaluation Metrics

The details of the evaluation metrics are provided below.

- *Mean Per Joint Position Error (MPJPE) (An et al., 2022):* The mean Euclidean distance between predicted and ground-truth joint positions to evaluate pose estimation accuracy.
- *Mean Absolute Error (MAE) (An & Ogras, 2021):* The average l_1-norm of errors across the x, y, and z axes for estimated joints.
- *Mean Squared Error (MSE) (Bishop & Nasrabadi, 2006):* The average squared difference between the estimated and ground-truth descriptor values; lower values indicate more accurate descriptor estimation.
- *Correct Joint Percentage (PCK@0.5) (Badiola-Bengoa & Mendez-Zorrilla, 2021):* The percentage of joints with prediction errors within 50% of torso diameter.
- *Basis Reconstruction Error:* The representation capability of P_{basis} for input feature F, defined as $||F - P_{basis} P_{basis}^T F||_2^2$.

- *Margin Difference:* The representation capability of signal feature μ_S is evaluated by comparing latent margins to positive signal μ_S^{pos} (similar pose) and negative signal μ_S^{neg} (dissimilar pose) representations, formulated as $||\mu_S - \mu_S^{neg}||_2^2 - ||\mu_S - \mu_S^{pos}||_2^2$; higher values indicate better representation.
- *Signal Degradation Score:* We characterize radar inputs using spatial sparsity and noise (Chen et al., 2022). Sparsity is the median mean kNN distance normalized by the frame's bounding-box diagonal (Ning et al., 2018), while noise is the Statistical Outlier Removal ratio, i.e., the fraction of points with mean kNN distance above $\mu + \alpha\sigma$ (Rusu et al., 2008). We normalize both dataset-level metrics to $[0, 100]$, and define $\mathrm{SDS} = (\widetilde{S} + \widetilde{N})/2$; higher values indicate greater input degradation.