

参赛学生姓名： 山闻涛

中学： 深圳国际交流书院

省份： 广东省

国家/地区： 中国

指导老师姓名： Shahla Ahmadova

指导老师单位： 深圳国际交流书院

论文题目： SRA-AF: Symmetric Reliability-Aware Adaptive Fusion for 3D Human Pose Estimation with mmWave Radar and Sparse IMUs

SRA-AF: Symmetric Reliability-Aware Adaptive Fusion for 3D Human Pose Estimation with mmWave Radar and Sparse IMUs

Wentao Shan

Abstract

Three-dimensional human pose estimation plays an important role in applications such as rehabilitation assessment, smart homes, and safety monitoring. Millimeter-wave (mmWave) radar and sparse inertial measurement units (IMUs) provide complementary cues for 3D human pose estimation, but the relative reliability of the two modalities varies across actions, subjects, and environments. Consequently, direct feature concatenation or fixed-weight fusion may lead to negative transfer. To address this issue, we propose the Symmetric Reliability-Aware Adaptive Fusion Transformer (SRA-AF), which estimates time-varying relative reliability weights from modality-specific radar and IMU features and their cross-modal interactions. These shared weights jointly guide bidirectional cross-modal rectification and final multimodal fusion. No additional supervision is required for reliability estimation, as the weights are learned end-to-end through the pose estimation loss. Five-fold cross-validation on mRI yielded MPJPEs of 94.29 mm and 73.01 mm under the P1 and P2 protocols, respectively. Under the cross-user protocol on OctoNet, SRA-AF achieved an MPJPE of 101.42 mm, reducing the error of the IMU-only model by 4.62 mm. Temporal-misalignment controls and ablations support the roles of synchronized cross-modal information, reliability estimation, and bidirectional rectification. Cross-scene results also show that the method remains sensitive to distribution shifts. Overall, the results support adaptive reliability-aware fusion and point to robust cross-scene fusion as a key remaining problem.

Keywords:

3D human pose estimation; Millimeter-wave radar; Inertial measurement units; Multimodal fusion; Reliability estimation; Transformer

Contents

1	Introduction	1
2	Related Work	3
2.1	MmWave Radar-Based Human Pose Estimation	3
2.2	Sparse IMU-Based Human Pose Estimation	3
2.3	Multimodal Human Pose Estimation and Adaptive Fusion	4
3	Method	5
3.1	Problem Formulation	5
3.2	Framework Overview	5
3.3	Modality Feature Encoding	6
3.4	Symmetric Relative Reliability Estimation	7
3.5	Reliability-Guided Bidirectional Cross-Modal Rectification	7
3.6	Adaptive Fusion and Pose Regression	8
3.7	Training Objective	9
4	Experimental Setup	9
4.1	Datasets and Task Settings	9
4.1.1	mRI Dataset	9
4.1.2	OctoNet Dataset	10
4.2	Evaluation Metrics	11
4.3	Implementation Details and Comparison Methods	11
5	Results	12
5.1	Cross-Subject Results on mRI	12
5.2	Cross-User and Cross-Scene Results on OctoNet	13
5.3	Ablation Study	14
5.3.1	Ablation Results on mRI	14
5.3.2	Ablation Results on OctoNet	14
5.4	Qualitative Pose Analysis	15
5.5	Action-wise Performance Analysis	16
5.6	Relative Reliability Analysis	17
6	Conclusion	19
	References	20
	Acknowledgments	23

1 Introduction

Three-dimensional human pose estimation recovers the 3D positions of human joints from sensor observations. It is widely used in intelligent perception [1], human-computer interaction [2], rehabilitation assessment [3], motion analysis [4], action understanding [5], and safety monitoring [6]. As deep learning and sensing hardware have improved, pose estimation has moved from handcrafted features and kinematic models toward end-to-end data-driven prediction. RGB, depth, and multi-view camera methods have made strong progress, especially when textures are visible, lighting is stable, and occlusion is limited [7], [8].

Visual sensors still face practical limitations [1], [9]. RGB and RGB-D cameras are affected by illumination changes, occlusion, restricted viewpoints, and background clutter. Their estimates can degrade in low light, smoke, crowded indoor scenes, or when body parts are hidden. Cameras also record appearance, faces, clothing, and indoor surroundings, which raises privacy concerns in long-term settings such as home monitoring, rehabilitation, and elder care. In addition, camera systems require suitable placement and field of view. Pose estimates may become unstable when a person leaves the view, turns away, or is partly blocked by furniture. These issues have encouraged the use of non-visual sensors in privacy-sensitive and visually difficult environments.

Millimeter-wave (mmWave) radar is an active wireless sensor and a useful alternative for non-visual pose estimation. By transmitting and receiving electromagnetic waves, it measures range, velocity, angle, point-cloud structure, and related motion information from human reflections. Unlike cameras, mmWave radar does not capture appearance images. It works in low light, is largely unaffected by illumination, and is better suited to privacy-sensitive smart-home, healthcare, and safety-monitoring settings. Recent studies have estimated pose from radar point clouds, range-Doppler maps, range-azimuth maps, and related representations, showing that radar signals contain meaningful information about human structure and motion [10], [11].

Radar-only 3D pose estimation remains difficult. Radar usually has lower spatial resolution than visual sensors, and human point clouds are often sparse, noisy, and unstable. Reflections from different body parts change with material, orientation, distance, and occlusion. Distal joints such as wrists and ankles are especially hard to observe consistently [12]. As a result, radar-only models may confuse left and right limbs or produce larger errors in fine-grained limb positions and dynamic motions.

Inertial measurement unit (IMU) provides a complementary signal by measuring local motion directly at their attachment site. The IMU typically measures acceleration, angular velocity, orientation, or related motion features, giving direct information about local body movement. Compared with radar, IMU has high sampling rates, responsive motion measurements, compact hardware, and low computational cost, which makes

them suitable for continuous sensing [3]. Many IMU-based pose methods place several IMU nodes on different body parts to recover whole-body pose. However, in daily use, multi-IMU setups add wearing, deployment, and calibration burden. A single IMU or a small set of IMUs is more practical, but it observes only a few body locations and cannot fully determine the 3D skeleton by itself.

Combining radar’s whole-body spatial cues with local motion measurements from sparse IMUs is a practical way to compensate for the weaknesses of each sensor. The two modalities are complementary, yet fusion does not automatically improve accuracy. Radar and IMU data differ in structure, noise, and physical meaning. If one modality suffers from missing reflections, scene changes, or temporal misalignment, direct concatenation can pass unreliable information into the shared representation. A fixed fusion ratio also cannot respond to changing input conditions. Radar describes the spatial distribution of the body, whereas IMUs record local motion at the attached locations. Their relative usefulness can therefore change over time, so the fusion module should estimate each modality’s contribution from the current input.

To address this challenge, we propose the Symmetric Reliability-Aware Adaptive Fusion Transformer (SRA-AF). The model encodes temporal radar and IMU features in separate branches, then estimates time-varying relative reliability weights from the unimodal representations, their shared responses, and their differences. These weights control both bidirectional cross-attention rectification and the final fusion ratio between the rectified modalities. Because they are learned end to end from pose prediction error, the weights should be interpreted as relative task-specific contributions rather than calibrated probabilities of sensor quality.

The main contributions of this study are as follows:

1. We formulate radar-IMU 3D human pose estimation as a symmetric reliability-aware fusion problem. The method does not assume that either modality is always dominant; instead, it models modality reliability across samples and time.
2. We develop SRA-AF, which combines dual-branch temporal encoding, interaction-based reliability estimation, reliability-controlled bidirectional cross-attention rectification, and adaptive fusion in one network.
3. We evaluate the method with cross-subject, cross-user, and cross-scene experiments on mRI and OctoNet. Temporal-misalignment controls, component ablations, and visual analyses are used to examine how the method works.

The rest of the paper is organized as follows. Section 2 reviews mmWave radar, sparse IMUs, and multimodal human pose estimation. Section 3 describes the proposed SRA-AF method. Section 4 presents the datasets and experimental settings. Section 5 reports the main results, ablations, and qualitative findings. Section 6 concludes the paper.

2 Related Work

2.1 MmWave Radar-Based Human Pose Estimation

MmWave radar provides spatial and motion observations without recording identifiable human appearance. It is therefore suitable for privacy-sensitive and low-light environments. Early work mainly tested whether human pose could be recovered from radar observations. Mm-Pose used a convolutional neural network to regress skeletal joints from commercial mmWave radar signals [13]. MARS estimated 3D joint positions from temporal point clouds for assisted rehabilitation [11]. These studies showed that radar reflections contain information related to human structure and motion.

Later work has mainly followed two directions: designing more informative radar representations and adding prior knowledge. MPose combined spatial feature extraction with domain-adversarial learning to improve generalization across environments and subjects [14]. HuPR provided synchronized visual and dual-radar data, establishing a public benchmark for radar-based human pose estimation [15]. MmPose-NLP represented radar point-cloud sequences as tokens and used a sequence model to predict joint positions [16]. MmPose-FK incorporated forward-kinematic constraints to reduce joint jitter and anatomically implausible predictions [17]. Despite this progress, radar-only pose estimation remains limited by sparse reflections, multipath interference, and uneven visibility across body parts. Temporal and kinematic priors can compensate for brief observation gaps, but they cannot fully recover inadequately observed local motion. Therefore, the introduction of a complementary wearable sensor is a common extension.

2.2 Sparse IMU-Based Human Pose Estimation

IMU directly measures local orientation and acceleration at the body parts where they are attached. Conventional inertial motion-capture systems use many sensors to constrain whole-body motion, but their wearing and calibration procedures are too cumbersome for many everyday scenarios. Sparse Inertial Poser (SIP) reduced this burden by reconstructing whole-body motion from six IMUs through multi-frame joint optimization [18]. Deep Inertial Poser (DIP) used synthetic inertial data to train a bidirectional recurrent neural network that maps sparse inertial measurements to whole-body pose [19]. These studies established sparse IMUs as an alternative to dense inertial motion-capture systems.

Later methods added structural, temporal, and physical constraints. TransPose decomposed the task into end-joint estimation, full-joint position completion, and inverse kinematics while also estimating global body translation [20]. Physical Inertial Poser

incorporated physics-based optimization to improve motion stability and physical plausibility [21]. Transformer Inertial Poser modeled temporal dependencies with a Transformer and used terrain and contact information to mitigate long-term drift [22]. Recent work has also explored more flexible sensor configurations. MobilePoser uses IMUs embedded in consumer devices [23], whereas DiffusionPoser supports arbitrary sparse sensor layouts through an autoregressive diffusion model [24]. Sparse IMUs still cannot directly observe the body’s overall structure in external space. Similar measurements at a few sensor locations may correspond to different whole-body poses, leaving the reconstruction problem underconstrained. Radar observations can supply the spatial information missing from sparse inertial inputs.

2.3 Multimodal Human Pose Estimation and Adaptive Fusion

Multimodal datasets have made sensor-fusion research possible for human pose estimation. mRI provides synchronized mmWave radar, RGB-D, IMU, and 3D pose data, while MM-Fi includes several non-invasive sensing modalities [9]. OctoNet further increases the diversity of participants, activities, scenes, and sensor configurations [25]. Methods such as ImmFusion, AdaptiveFusion, and Babel use attention-based token interactions to combine heterogeneous sensor inputs, mainly for radar-vision or general multimodal sensing [26], [27], [28].

Direct radar-IMU fusion for pose estimation remains less explored. mRI evaluated several modality combinations for rehabilitation movements [3], and mmEgoHand combined head-mounted radar with an IMU for egocentric hand-pose estimation [29]. Kim et al. [30] studied whole-body radar-IMU fusion at the pose-output level. Firstly, the 3D joint positions are predicted by two unimodal encoders respectively, and then the relationship between the two sets of pose outputs is modeled by bidirectional cross-attention. The auxiliary supervision is constructed by using the unimodal joint errors to learn the fusion weights that change with joints and time.

Multimodal pose-estimation methods commonly fuse sensor features through concatenation, attention, or gating. When one modality is noisy, temporally misaligned, or shifted out of distribution, direct fusion may inject harmful information into the shared representation. Fixed fusion ratios also cannot reflect changes in modality usefulness across input conditions. Existing approaches mostly use fusion weights for feature aggregation. Few methods use the same reliability estimate to control both bidirectional cross-modal rectification and final fusion. We address this gap by estimating task-specific relative reliability from radar features, IMU features, and their interactions, and then using that estimate in both stages.

3 Method

3.1 Problem Formulation

Given synchronized mmWave radar and IMU sequences of length T ,

$$X^r = \{x_t^r\}_{t=1}^T, \quad X^i = \{x_t^i\}_{t=1}^T \quad (3.1)$$

where x_t^r and x_t^i denote the radar and inertial observations at time step t , respectively. The model uses this temporal window to predict the 3D human pose at the final time step:

$$\hat{Y}_T^{\text{rel}} = f_\theta(X^r, X^i) = \{\hat{p}_{T,j}^{\text{rel}}\}_{j=1}^J, \quad \hat{p}_{T,j}^{\text{rel}} \in \mathbb{R}^3 \quad (3.2)$$

where J is the number of body joints and $\hat{p}_{T,j}^{\text{rel}}$ is the predicted coordinate of joint j .

To remove the effect of global translation, we use the root joint as the coordinate origin. All joints are expressed relative to this root:

$$p_{T,j}^{\text{rel}} = p_{T,j} - p_{T,\text{root}} \quad (3.3)$$

The prediction target is the joint configuration relative to the root rather than the body's absolute translation in global coordinates.

3.2 Framework Overview

The overall SRA-AF architecture is shown in Fig. 1. It comprises five stages: modality feature encoding, relative reliability estimation, bidirectional cross-modal rectification, adaptive fusion, and pose regression.

The radar and IMU inputs first pass through separate frame-level and temporal encoders. The resulting representations capture temporal context within each modality. The reliability estimator then analyzes the unimodal features and their interactions to assign relative reliability weights at each time step.

These weights determine how strongly one modality rectifies the other and how much each rectified modality contributes during final fusion. The model does not set radar or IMU as a fixed dominant source. Instead, it adjusts their roles according to the current input, and the fused temporal representation is used to regress the final-frame 3D pose.

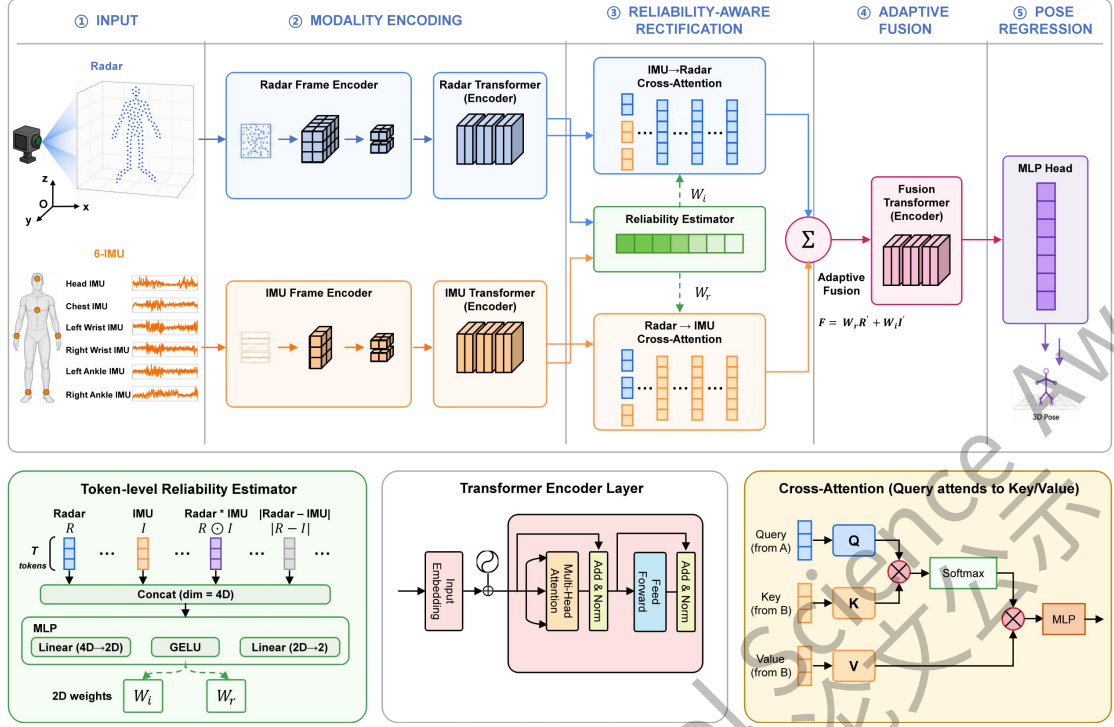


Fig. 1 Overview of the proposed SRA-AF architecture.

3.3 Modality Feature Encoding

MmWave radar and IMU data differ substantially in structure, numerical distribution, and physical meaning. Therefore, we use two independent branches to extract unimodal features and avoid mixing heterogeneous observations at an early encoding stage.

For the radar input x_t^r at frame t , the radar encoder first applies a shared feature mapping to each valid radar unit. Masked max pooling and masked mean pooling then aggregate the local responses. Meanwhile, a global mapping of the entire radar frame preserves frame-level spatial information. The three feature components are concatenated and projected to form the radar-frame representation:

$$e_t^r = \phi_f^r ([\text{MaxPool}(\phi_p^r(x_t^r)), \text{MeanPool}(\phi_p^r(x_t^r)), \phi_g^r(x_t^r)]) \quad (3.4)$$

where ϕ_p^r denotes the shared local feature mapping, ϕ_g^r denotes the frame-level global mapping, and ϕ_f^r combines the radar features. For padded radar inputs, pooling is performed only over valid units.

For the inertial input x_t^i at frame t , the inertial encoder maps observations from multiple attachment locations to a unified frame-level representation:

$$e_t^i = \phi^i(x_t^i) \quad (3.5)$$

The two frame-level features are then projected to the same dimension and aug-

mented with temporal positional encodings:

$$u_t^r = W^r e_t^r + \text{PE}(t), \quad u_t^i = W^i e_t^i + \text{PE}(t) \quad (3.6)$$

Two independent temporal Transformers model motion changes within the temporal window:

$$R = \text{Trm}_r(U^r), \quad I = \text{Trm}_i(U^i) \quad (3.7)$$

where $R = \{r_t\}_{t=1}^T$ and $I = \{i_t\}_{t=1}^T$ denote the temporally contextualized radar and inertial representations, respectively. The two temporal branches have identical architectures but independent parameters, preserving modality-specific information while mapping both inputs into a common feature space.

3.4 Symmetric Relative Reliability Estimation

Direct concatenation and fixed weighting treat modality contributions as constant across samples and time. They cannot adapt when input validity changes. SRA-AF instead jointly analyzes r_t and i_t at each time step and constructs a cross-modal interaction descriptor:

$$z_t = [r_t, i_t, r_t \odot i_t, |r_t - i_t|] \quad (3.8)$$

where r_t and i_t retain modality-specific information. The element-wise product $r_t \odot i_t$ captures their shared response, whereas the absolute difference $|r_t - i_t|$ describes their discrepancy.

The interaction descriptor is passed through the reliability estimator. A softmax operation then produces relative reliability weights for the radar and inertial modalities:

$$[w_t^r, w_t^i] = \text{Softmax}(\phi_{\text{rel}}(z_t)), \quad w_t^r + w_t^i = 1 \quad (3.9)$$

The weights vary across samples and time. A larger w_t^r indicates a greater relative contribution from radar to the current prediction, whereas a larger w_t^i indicates greater reliance on inertial information. Because these weights are learned end-to-end through the pose objective, we define them as task-specific relative reliability rather than absolute probabilities of sensor quality.

3.5 Reliability-Guided Bidirectional Cross-Modal Rectification

Weighted fusion changes each modality's contribution to the fused representation, but it does not let one modality recover information that is missing from the other be-

fore fusion. Therefore, SRA-AF applies multi-head cross-attention in both directions between the radar and IMU representations.

Using the radar representation as the query and the inertial representation as the key and value, the model extracts inertial information that complements the radar modality:

$$C_{i \rightarrow r} = \text{MHA}(Q = R, K = I, V = I) \quad (3.10)$$

The relative reliability of the inertial modality controls the strength of this rectification:

$$R' = \text{LN}(R + w^i \odot C_{i \rightarrow r}) \quad (3.11)$$

Conversely, the inertial representation serves as the query, while the radar representation serves as the key and value. This operation extracts radar information that complements the inertial modality:

$$C_{r \rightarrow i} = \text{MHA}(Q = I, K = R, V = R) \quad (3.12)$$

The radar reliability weight then controls the update to the inertial representation:

$$I' = \text{LN}(I + w^r \odot C_{r \rightarrow i}) \quad (3.13)$$

where $w^r = \{w_t^r\}_{t=1}^T$ and $w^i = \{w_t^i\}_{t=1}^T$ collect the reliability weights over time and are broadcast along the feature dimension. Each update uses the weight of the modality providing the rectification information. Therefore, reliable inertial information can more strongly rectify the radar representation, and reliable radar information can more strongly rectify the inertial representation. This design exchanges complementary information while reducing interference from the relatively less reliable modality.

3.6 Adaptive Fusion and Pose Regression

After bidirectional rectification, the same relative reliability weights fuse the two modality representations:

$$F = w^r \odot R' + w^i \odot I' \quad (3.14)$$

Bidirectional cross-modal rectification exchanges complementary information before fusion. Relative-reliability weighting determines the contributions of the rectified modalities to the shared representation. The two operations share the same weights, ensuring consistent control over information injection and final contribution.

The fused sequence F is then passed through a fusion Transformer, which further models its temporal dependencies:

$$H = \text{Trm}_f(F) \quad (3.15)$$

To preserve both the local state at the target time and motion across the temporal window, the model concatenates the final-step representation with temporal mean and max responses:

$$h = [H_T, \text{MeanPool}(H), \text{MaxPool}(H)]. \quad (3.16)$$

Finally, the pose regressor maps the aggregated representation to the 3D joint coordinates at the final time step:

$$\hat{Y}_T^{\text{rel}} = \phi_{\text{pose}}(h) \in \mathbb{R}^{J \times 3}. \quad (3.17)$$

3.7 Training Objective

SRA-AF is trained end-to-end using an L1 loss on the root-relative 3D coordinates. For a training batch containing B samples, the pose regression loss is

$$\mathcal{L}_{\text{pose}} = \frac{1}{3BJ} \sum_{b=1}^B \sum_{j=1}^J \|\hat{p}_{b,T,j}^{\text{rel}} - p_{b,T,j}^{\text{rel}}\|_1 \quad (3.18)$$

We do not require additional modality-reliability labels. Instead, the pose prediction error jointly optimizes the reliability estimator and the other network components. This allows the estimator to learn task-specific relative modality contributions. The complete model is trained using only the pose regression loss:

$$\mathcal{L} = \mathcal{L}_{\text{pose}} \quad (3.19)$$

4 Experimental Setup

4.1 Datasets and Task Settings

4.1.1 mRI Dataset

mRI is a dataset that synchronously records mmWave radar, RGB-D, IMU, and 3D human pose data [3]. We use its mmWave radar features, six IMU nodes, and 3D pose labels. Each radar frame has dimensions of $14 \times 14 \times 5$, and each IMU frame has dimensions of 6×12 . The pose labels contain 17 body joints. The model takes a continuous sequence of length $T = 16$ as input and predicts the 3D pose at the final time step. Every window remains within a single subject and action segment.

We use two action protocols. Protocol 1 (P1) contains 10 standardized actions, free-form actions, and walking, giving 12 action classes in total. Protocol 2 (P2) contains only the 10 standardized actions.

Both protocols use five-fold cross-validation at the subject level. The 20 subjects are divided into five test folds, each containing four test subjects. In each fold, the remaining subjects are further divided into 14 training subjects and two validation subjects. The training, validation, and test subjects do not overlap. We report the mean and standard deviation across the five test folds.

4.1.2 OctoNet Dataset

OctoNet is a multimodal dataset containing multiple users, scenes, activities, and sensing devices [25]. We use observations from five frequency-modulated continuous-wave (FMCW) radar nodes, six fixed IMU locations, and 3D pose labels. Each radar frame is represented by 150×5 point-cloud features, and each IMU frame contains 6×13 inertial features. The IMU data are standardized using training-set statistics. An inverse hyperbolic sine transformation reduces the influence of extreme values.

The model takes a continuous sequence of length $T = 9$ as input and predicts a 20-joint 3D pose at the final time step. Every window is constructed within a single activity segment.

We use the predefined OctoNet training, validation, and test splits. Results are reported under the cross-user (CU) and cross-scene (CS) settings. CU evaluates generalization to unseen users, whereas CS evaluates generalization to unseen scenes. CU includes 62 standardized activities, gym activities, and freestyle activities, giving 64 classes in total. After filtering for multi-node data completeness, CS contains 58 activity classes. Because OctoNet provides fixed data splits, we report results directly on the corresponding test sets without five-fold cross-validation.

Table 1 Experimental settings for mRI and OctoNet.

Item	mRI	OctoNet
Radar Input	$14 \times 14 \times 5$	150×5
IMU Input	6×12	6×13
Window	$T = 16$	$T = 9$
Pose	17 joints	20 joints
Main split	P1/P2 cross-subject five-fold	CU / CS

4.2 Evaluation Metrics

We evaluate pose prediction using the mean per-joint position error (MPJPE) and Procrustes-aligned mean per-joint position error (PA-MPJPE). Both metrics are measured in millimeters, and lower values indicate more accurate predictions.

MPJPE is the mean Euclidean distance between the predicted and ground-truth joint coordinates:

$$\text{MPJPE} = \frac{1}{BJ} \sum_{b=1}^B \sum_{j=1}^J \|\hat{p}_{b,j} - p_{b,j}\|_2 \quad (4.1)$$

where B is the number of test samples, and J is the number of joints. The coordinates $\hat{p}_{b,j}$ and $p_{b,j}$ denote the prediction and ground truth for joint j in sample b , respectively.

PA-MPJPE first aligns the predicted and ground-truth poses of each sample using a similarity transformation:

$$(s^*, Q^*, t^*) = \arg \min_{s, Q, t} \|s\hat{Y}Q + \mathbf{1}t^\top - Y\|_F^2, \quad Q \in SO(3) \quad (4.2)$$

where s , Q , and t denote scale, rotation, and translation, respectively. Reflections are excluded during alignment, after which MPJPE is computed for the aligned pose. Compared with MPJPE, PA-MPJPE focuses more strongly on the relative structure of the predicted skeleton.

For mRI, we report the mean and standard deviation across the five test folds. For OctoNet, we report results directly on the fixed CU and CS test sets.

4.3 Implementation Details and Comparison Methods

We implemented the model in PyTorch. We used Adam for mRI and AdamW for OctoNet, with an initial learning rate of 1×10^{-4} for both datasets. The batch sizes were 128 and 64, respectively. The best model was selected by validation MPJPE. Within each dataset, all methods used the same data splits, training configurations, and pose regression loss. Table II summarizes the main training hyperparameters.

We compare SRA-AF with four baselines: Radar-only, IMU-only, Concat Fusion, and Misaligned IMU. Radar-only and IMU-only regress pose from radar and IMU observations, respectively. Concat Fusion concatenates the encoded modality features and projects them into a shared representation space. Misaligned IMU retains the radar input while shifting the mRI and OctoNet IMU sequences 30 and four frames into the past, respectively. This control tests whether the model uses temporally synchronized cross-modal information. In contrast, SRA-AF estimates relative modality weights and uses them to control bidirectional cross-modal rectification and final feature fusion.

Table II Main training hyperparameters.

Parameter	mRI	OctoNet
Batch size	128	64
Learning rate	1×10^{-4}	1×10^{-4}
Optimizer	Adam	AdamW
Epochs	50	100
Early stopping	10	20
Loss	L1	L1
Dropout	0.1	0.1

We further evaluate three ablated versions of SRA-AF. The w/o Reliability variant fixes both modality weights at 0.5. The w/o Bidirectional Rectification variant removes bidirectional cross-modal rectification and directly fuses the two unimodal representations. The w/o Fusion Transformer variant retains reliability estimation, bidirectional rectification, and weighted fusion but removes post-fusion temporal modeling.

5 Results

5.1 Cross-Subject Results on mRI

Table III reports the subject-level five-fold results on mRI under P1 and P2.

Table III Five-fold cross-subject results on mRI (mm)

Protocol	Method	MPJPE ↓	PA-MPJPE ↓
P1	Radar-only	111.09 ± 20.91	97.62 ± 32.32
	IMU-only	94.34 ± 18.39	84.82 ± 31.12
	Concat Fusion	95.99 ± 19.12	86.77 ± 31.09
	Misaligned IMU	105.33 ± 25.17	88.36 ± 35.31
	SRA-AF	94.29 ± 18.75	82.86 ± 31.29
P2	Radar-only	89.72 ± 5.08	66.93 ± 4.20
	IMU-only	75.64 ± 6.59	53.66 ± 8.45
	Concat Fusion	75.95 ± 6.48	54.75 ± 5.09
	Misaligned IMU	84.69 ± 10.58	57.83 ± 9.16
	SRA-AF	73.01 ± 5.30	50.22 ± 4.68

Under P1, IMU-only clearly outperformed Radar-only, indicating that six IMUs already provide strong motion constraints. SRA-AF achieved an MPJPE of 94.29 mm, nearly matching the 94.34 mm obtained by IMU-only, and reduced PA-MPJPE from 84.82 to 82.86 mm. Temporal misalignment increased the MPJPE of Misaligned IMU to 105.33 mm, showing that synchronized inertial information was important for effective

fusion. These results show that SRA-AF preserved overall joint localization accuracy while improving the aligned skeletal structure.

Under P2, SRA-AF achieved an MPJPE of 73.01 mm and a PA-MPJPE of 50.22 mm, the lowest values among all methods. Compared with the strongest baseline IMU-only, it reduced MPJPE by 2.63 mm and PA-MPJPE by 3.44 mm. Misaligned IMU produced an MPJPE of 84.69 mm. These results indicate that whole-body radar observations complemented local motion information from synchronized IMUs, whereas temporal misalignment weakened this complementarity.

Overall, SRA-AF matched the strong IMU-only baseline in joint localization under P1 while reducing skeletal-structure error. Under P2, it improved both MPJPE and PA-MPJPE.

5.2 Cross-User and Cross-Scene Results on OctoNet

Table IV reports results on the fixed CU and CS splits of OctoNet.

Table IV Pose estimation results under the OctoNet CU and CS settings (mm).

Method	CU MPJPE ↓	CU PA-MPJPE ↓	CS MPJPE ↓	CS PA-MPJPE ↓
Radar-only	155.40	92.42	187.70	109.98
IMU-only	106.04	73.27	184.72	104.35
Concat Fusion	114.12	79.60	179.65	106.67
Misaligned IMU	103.67	71.50	213.17	103.54
SRA-AF	101.42	70.67	182.65	103.50

Under CU, SRA-AF achieved an MPJPE of 101.42 mm and a PA-MPJPE of 70.67 mm, the lowest values among all methods. Compared with IMU-only, it reduced MPJPE and PA-MPJPE by 4.62 and 2.60 mm, respectively. Compared with Misaligned IMU, it reduced the two errors by 2.25 and 0.83 mm. These results indicate that reliability-guided cross-modal interaction used synchronized radar and IMU information to improve pose estimation for unseen users.

Under CS, Concat Fusion achieved the lowest MPJPE of 179.65 mm, followed by SRA-AF at 182.65 mm. SRA-AF still outperformed Radar-only, IMU-only, and Misaligned IMU. Its PA-MPJPE was numerically the lowest at 103.50 mm, nearly identical to the 103.54 mm of Misaligned IMU. However, Misaligned IMU had a much higher MPJPE of 213.17 mm, showing that similarity alignment can hide some global pose errors in PA-MPJPE. This should not be read as evidence that temporal misalignment is useful. For cross-scene evaluation, MPJPE gives the more direct assessment.

Overall, SRA-AF achieved the lowest MPJPE and PA-MPJPE under CU, showing better generalization to unseen participants. The CS results were less clear: SRA-AF

obtained the lowest PA-MPJPE among the evaluated methods, but its MPJPE was 3.00 mm higher than Concat Fusion, indicating remaining sensitivity to scene changes.

5.3 Ablation Study

5.3.1 Ablation Results on mRI

Table V reports ablations of the core SRA-AF components on mRI. We use subject-level five-fold aggregate results and MPJPE to measure changes in joint localization error.

Table V Ablation of core SRA-AF components on mRI (mm).

Variant	Reliability	Bi-Rectification	Fusion Transformer	P1 MPJPE↓	P2 MPJPE↓
Full SRA-AF	✓	✓	✓	94.29 ± 18.75	73.01 ± 5.30
w/o Reliability	✗	✓	✓	96.51 ± 20.18	76.83 ± 6.88
w/o Bi-Rectification	✓	✗	✓	98.99 ± 28.56	75.34 ± 7.68
w/o Fusion Transformer	✓	✓	✗	94.38 ± 19.62	73.31 ± 6.38

Replacing dynamic reliability weights with fixed equal weights increased MPJPE by 2.22 mm under P1 and 3.82 mm under P2. Therefore, input-dependent modality weighting was more effective than a fixed fusion ratio. Removing bidirectional cross-modal rectification increased MPJPE by 4.70 mm under P1 and 2.33 mm under P2, showing that explicit cross-modal information exchange was also important.

Removing the Fusion Transformer increased MPJPE by only 0.09 mm under P1 and 0.30 mm under P2. Reliability estimation and bidirectional rectification had substantially larger effects, so the main benefit of SRA-AF did not come simply from adding temporal modeling after fusion.

5.3.2 Ablation Results on OctoNet

The OctoNet ablations use the same fixed CU and CS splits as the main experiments. Table VI presents the results.

Under CU, removing reliability estimation, bidirectional cross-modal rectification, and the Fusion Transformer increased MPJPE by 4.36, 7.68, and 10.58 mm, respectively. Under CS, the corresponding increases were 0.95, 2.53, and 5.10 mm. All ablated variants underperformed the full model in both settings, indicating that each component contributed to the final performance.

Reliability estimation contributed more under CU than under CS. Dynamic modality weighting therefore helped with user differences, but was less effective under cross-scene shifts. Removing bidirectional rectification caused additional errors, showing that cross-modal information exchange provided a consistent complement. Unlike the mRI

Table VI Ablation of core SRA-AF components on OctoNet (mm).

Variant	Reliability	Bi-Rectification	Fusion Transformer	CU MPJPE↓	CS MPJPE↓
Full SRA-AF	✓	✓	✓	101.42	182.65
w/o Reliability	✗	✓	✓	105.78	183.60
w/o Bi-Rectification	✓	✗	✓	109.10	185.18
w/o Fusion Transformer	✓	✓	✗	112.00	187.75

results, removing the Fusion Transformer caused the largest degradation on OctoNet, suggesting that post-fusion temporal modeling is more important when activities are more diverse and complex.

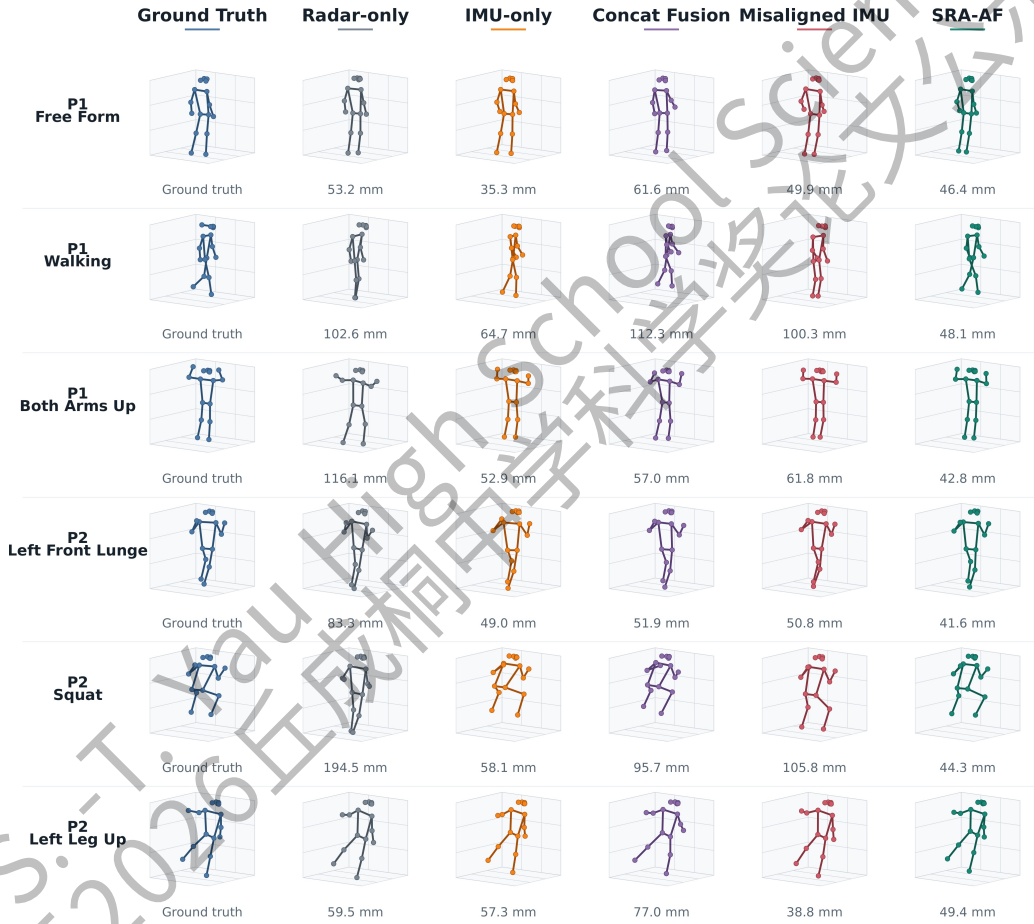


Fig. 2 Qualitative pose predictions on mRI using different methods.

The value below each predicted pose is the MPJPE for the displayed sample, measured in millimeters (mm).

5.4 Qualitative Pose Analysis

To compare the reconstructed skeletons visually, Fig. 2 shows 3D pose predictions from mRI. Across walking, arm raising, lunges, and squats, radar-only recovered the approximate skeleton but produced clear errors at several distal joints. IMU-only cap-

tured local motion well but provided limited constraints on the complete skeleton. In comparison, SRA-AF generally predicted trunk orientations and relative limb positions that were closer to the ground truth.

Fig. 3 presents 3D pose predictions from OctoNet. In CU examples, SRA-AF recovered the overall skeletons of unseen users accurately. The relative performance of fusion methods was more complex in the CS examples, consistent with the cross-scene results in Table IV.

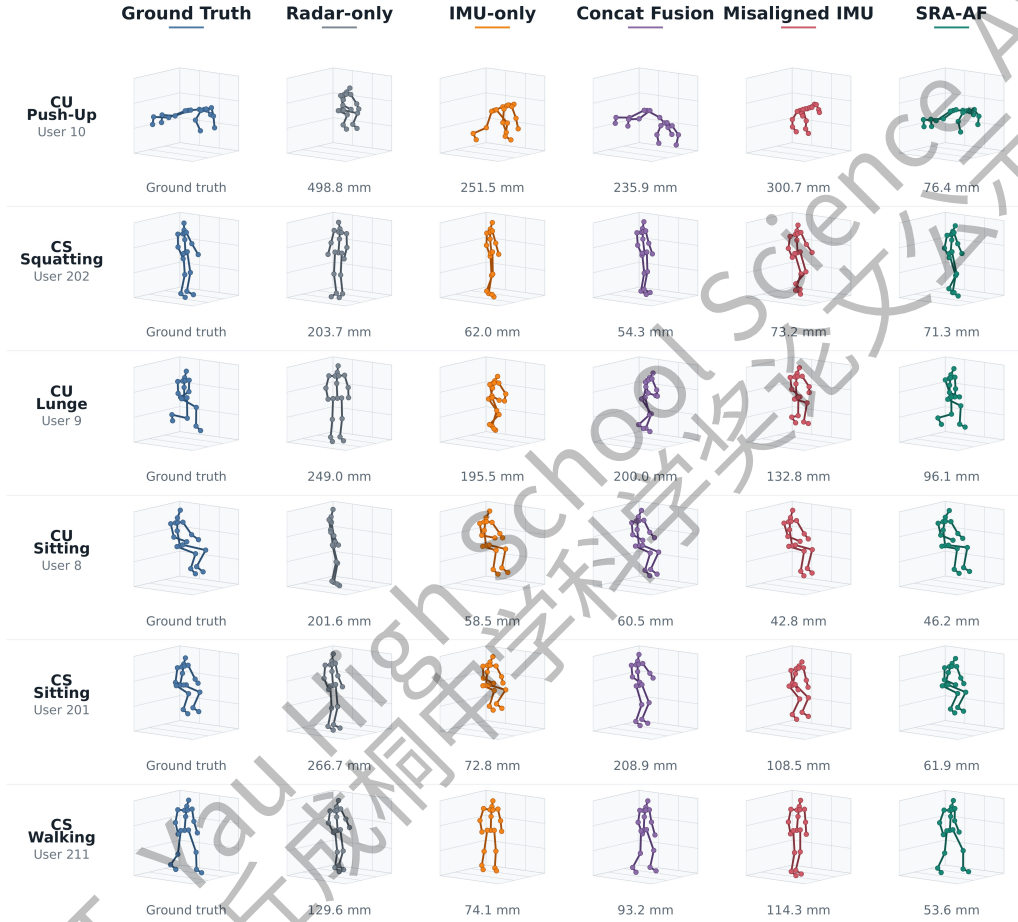


Fig. 3 Qualitative pose predictions on OctoNet using different methods.

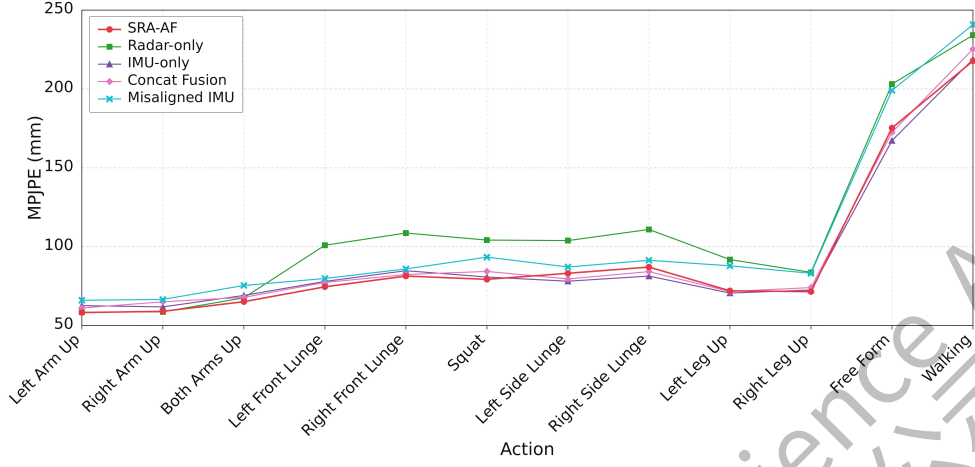
The value below each predicted pose is the MPJPE for the displayed sample, measured in millimeters (mm).

5.5 Action-wise Performance Analysis

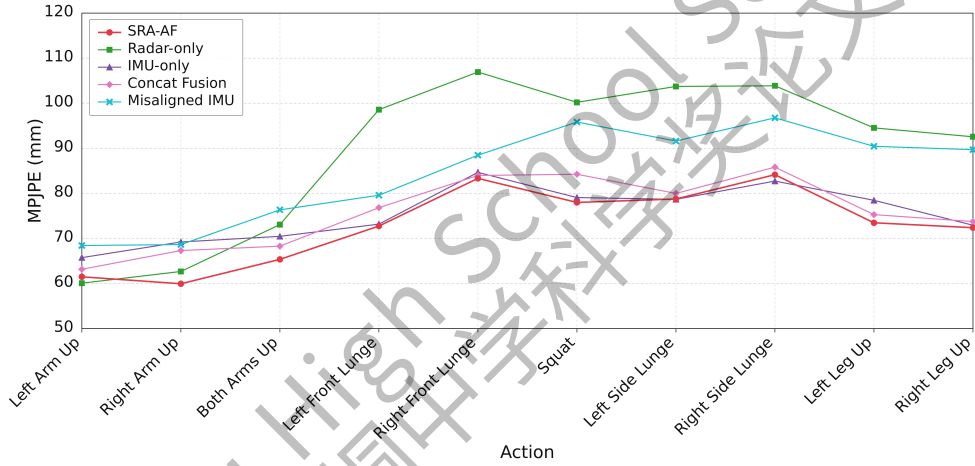
Aggregate errors can hide action-level differences, so we compare the action-level MPJPE of each method.

For mRI, Fig. 4 shows that SRA-AF achieved the lowest error in 7 of 12 P1 action classes, including arm raising, forward lunges, squats, single-leg raising, and walking. IMU-only matched or outperformed SRA-AF for some free-form actions, indicating that the relative value of each modality depends on the action. Under P2, SRA-AF achieved the lowest error in 7 of 10 action classes. The method was not superior for

every action, but its performance on most standardized actions was consistent with the aggregate advantage shown in Table III.



(a) mRI Protocol 1 Action-wise MPJPE.



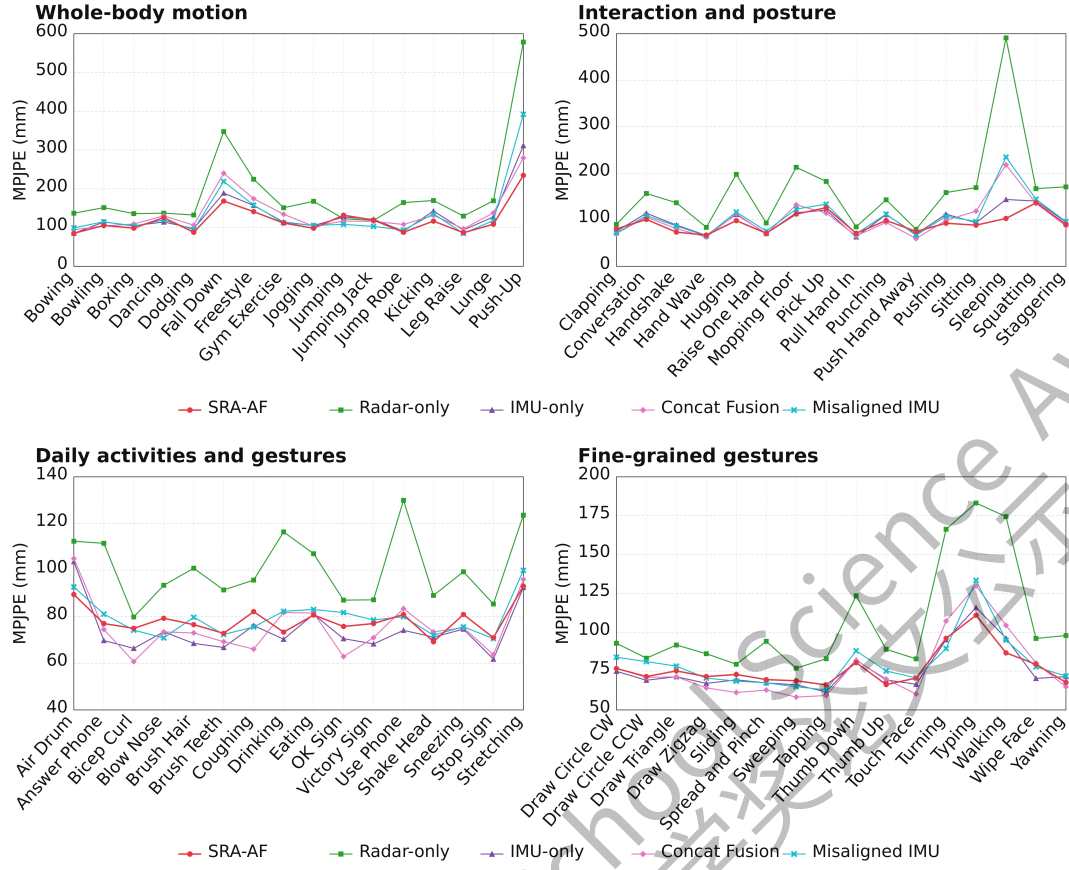
(b) mRI Protocol 2 Action-wise MPJPE.

Fig. 4 Action-wise MPJPE on mRI under P1 and P2.

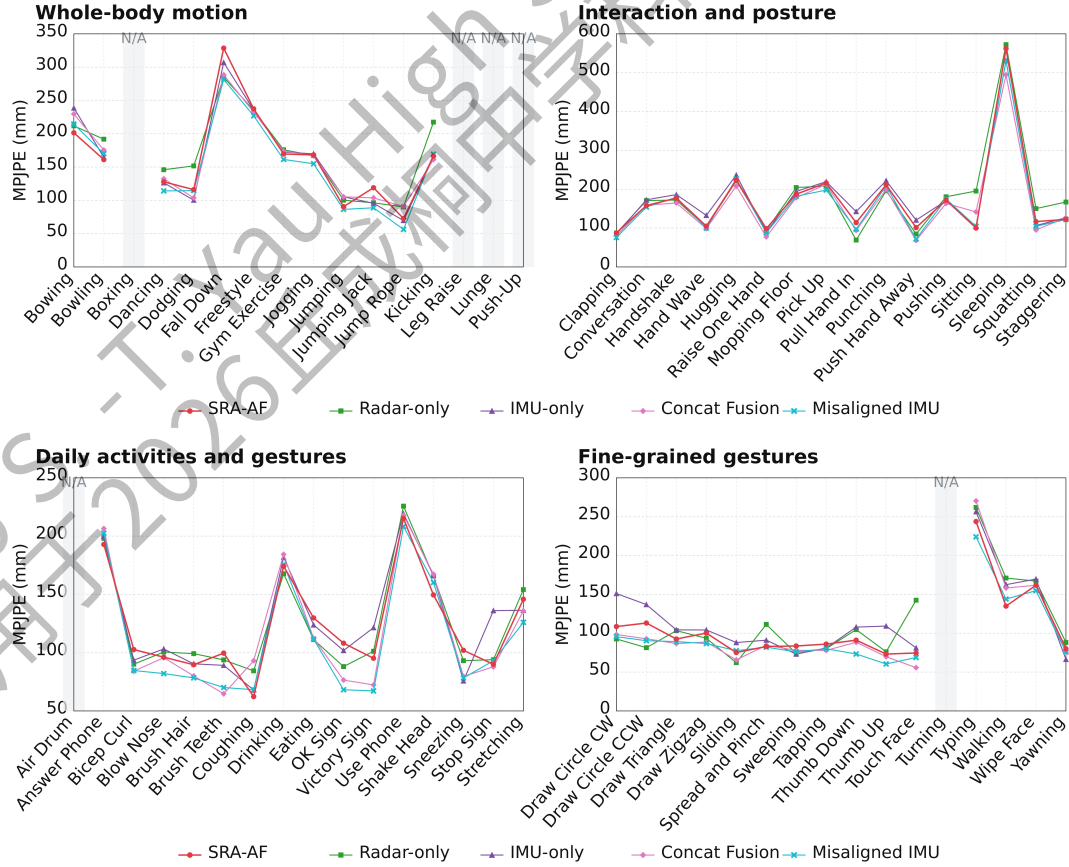
For OctoNet, the CU action-level results in Fig. 5a show that relative performance varied across activity types. SRA-AF achieved low errors for many activities but did not remain optimal for every class. Therefore, the benefit of dynamic fusion depends on the specific motion pattern. Fig. 5b shows greater variation among method rankings under CS than under CU. SRA-AF improved on unimodal results for some activities but did not consistently outperform Concat Fusion, which is consistent with the aggregate CS results in Table IV.

5.6 Relative Reliability Analysis

At each time step, SRA-AF outputs relative reliability weights for radar and IMU. These weights control both bidirectional cross-modal rectification and final fusion. Replacing the learned weights with fixed equal weights increased MPJPE by 2.22 mm un-



(a) Action-wise MPJPE under the OctoNet CU setting.



(b) Action-wise MPJPE under OctoNet CS setting.

Fig. 5 Comparison results on the OctoNet dataset.

der mRI P1 and 3.82 mm under P2. On OctoNet, the w/o Reliability variant increased MPJPE by 4.36 mm under CU and 0.95 mm under CS. Therefore, input-dependent dynamic weights outperformed a fixed fusion ratio on both datasets, although the size of the benefit depended on the evaluation setting.

The mean radar and IMU weights were approximately 0.41 and 0.59 under P1, and 0.27 and 0.73 under P2. The model assigned a higher average weight to IMU in both protocols, which is consistent with the overall strength of the IMU-only baseline. At the same time, whole-body spatial information from radar still complemented pose prediction through cross-modal rectification.

6 Conclusion

This study investigated 3D human pose estimation by fusing mmWave radar and sparse IMUs. Because the value of each modality changes with input conditions, simple fusion is not enough to allocate their contributions reliably. SRA-AF addresses this problem by extracting temporal radar and IMU features separately and estimating time-varying relative reliability from the unimodal features and their interactions. The same weights guide bidirectional cross-modal rectification and final fusion. Reliability estimation uses no additional manual labels and is learned end to end from pose prediction error.

On mRI, SRA-AF achieved an MPJPE of 94.29 mm under P1, nearly matching IMU-only while producing a lower PA-MPJPE. Under P2, it achieved an MPJPE of 73.01 mm and a PA-MPJPE of 50.22 mm, outperforming all comparison methods. Under the OctoNet CU setting, SRA-AF achieved an MPJPE of 101.42 mm and a PA-MPJPE of 70.67 mm. The temporal-misalignment controls showed that fusion gains depended on synchronized, valid cross-modal information. Component ablations further supported the contributions of relative reliability estimation and bidirectional cross-modal rectification.

Under the OctoNet CS setting, SRA-AF achieved the lowest numerical PA-MPJPE but did not outperform Concat Fusion in MPJPE. Therefore, the model remains sensitive to substantial scene-distribution shifts. Overall, the results show that task-driven relative reliability is more effective than fixed equal weighting, and that bidirectional interaction further improves complementary radar-IMU fusion. Maintaining this advantage across scenes remains an open challenge. Future work should also examine how the learned weights change with sensor quality and motion type.

References

- [1] X. Chen and J. Yang, “X-Fi: A Modality-Invariant Foundation Model for Multimodal Human Sensing,” in *International Conference on Learning Representations (ICLR)*, 2025.
- [2] J. Shotton et al., “Real-time human pose recognition in parts from single depth images,” in *CVPR 2011*, IEEE, 2011, pp. 1297–1304.
- [3] S. An, Y. Li, and U. Ogras, “Mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors,” *Advances in neural information processing systems*, vol. 35, pp. 27 414–27 426, 2022.
- [4] S. D. Uhlich et al., “Opencap: Human movement dynamics from smartphone videos,” *PLoS computational biology*, vol. 19, no. 10, e1011462, 2023.
- [5] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “Ntu rgb+ d: A large scale dataset for 3d human activity analysis,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.
- [6] X. Wang, J. Ellul, and G. Azzopardi, “Elderly fall detection systems: A literature survey,” *Frontiers in Robotics and AI*, vol. 7, p. 71, 2020.
- [7] K. Isakov, E. Burkov, V. Lempitsky, and Y. Malkov, “Learnable triangulation of human pose,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 2019, pp. 7717–7726.
- [8] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, “3d human pose estimation in video with temporal convolutions and semi-supervised training,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2019, pp. 7745–7754.
- [9] J. Yang et al., “Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 756–18 768, 2023.
- [10] Z. Zeng, X. Liang, Y. Li, and X. Dang, “Vulnerable road user skeletal pose estimation using mmwave radars,” *Remote Sensing*, vol. 16, no. 4, p. 633, 2024.
- [11] S. An and U. Y. Ogras, “Mars: Mmwave-based assistive rehabilitation system for smart healthcare,” *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 20, no. 5s, pp. 1–22, 2021.
- [12] Y.-H. Ho et al., “Rt-pose: A 4d radar tensor-based 3d human pose estimation and localization benchmark,” in *European conference on computer vision*, Springer, 2024, pp. 107–125.

- [13] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "Mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE sensors journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.
- [14] C. Shi, L. Lu, J. Liu, Y. Wang, Y. Chen, and J. Yu, "Mpose: Environment-and subject-agnostic 3d skeleton posture reconstruction leveraging a single mmwave device," *Smart Health*, vol. 23, p. 100 228, 2022.
- [15] S.-P. Lee, N. P. Kini, W.-H. Peng, C.-W. Ma, and J.-N. Hwang, "Hupr: A benchmark for human pose estimation using millimeter wave radar," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2023, pp. 5704–5713.
- [16] A. Sengupta and S. Cao, "Mmpose-nlp: A natural language processing approach to precise skeletal pose estimation using mmwave radars," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8418–8429, 2023.
- [17] S. Hu, S. Cao, N. Toosizadeh, J. Barton, M. G. Hector, and M. J. Fain, "Mmpose-fk: A forward kinematics approach to dynamic skeletal pose estimation using mmwave radars," *IEEE sensors journal*, vol. 24, no. 5, pp. 6469–6481, 2024.
- [18] T. Von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse inertial poser: Automatic 3d human pose estimation from sparse imus," in *Computer graphics forum*, Wiley Online Library, vol. 36, 2017, pp. 349–360.
- [19] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time," *ACM Transactions On Graphics (TOG)*, vol. 37, no. 6, pp. 1–15, 2018.
- [20] X. Yi, Y. Zhou, and F. Xu, "Transpose: Real-time 3d human translation and pose estimation with six inertial sensors," *ACM Transactions On Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021.
- [21] X. Yi et al., "Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2022, pp. 13 157–13 168.
- [22] Y. Jiang, Y. Ye, D. Gopinath, J. Won, A. W. Winkler, and C. K. Liu, "Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–9.

- [23] V. Xu, C. Gao, H. Hoffmann, and K. Ahuja, “Mobileposer: Real-time full-body pose estimation and 3d human translation from imus in mobile consumer devices,” in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024, pp. 1–11.
- [24] T. Van Wouwe, S. Lee, A. Falisse, S. Delp, and C. K. Liu, “Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2024, pp. 2513–2523.
- [25] D. Yuan et al., “Octonet: A large-scale multi-modal dataset for human activity understanding grounded in motion-captured 3d pose labels,” *Advances in Neural Information Processing Systems*, vol. 38, 2025.
- [26] A. Chen et al., “Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions,” *arXiv preprint arXiv:2210.01346*, 2022.
- [27] A. Chen et al., “Adaptivefusion: Adaptive multi-modal multi-view fusion for 3d human body reconstruction,” *IEEE Transactions on Multimedia*, 2025.
- [28] S. Dai et al., “Babel: A scalable pre-trained model for multi-modal sensing via expandable modality alignment,” in *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, 2025, pp. 240–253.
- [29] Y. Lv et al., “Mmegohand: Egocentric hand pose estimation and gesture recognition with head-mounted millimeter-wave radar and imu,” *arXiv preprint arXiv:2501.13805*, 2025.
- [30] S. Kim, S. Oh, J. Kim, and S.-C. Kim, “Cross-attention fusion of mmwave radar and imu for robust 3d human pose estimation,” in *2026 IEEE Radar Conference (RadarConf26)*, IEEE, 2026.

Acknowledgments

My interest in this project began with a possible use of human pose estimation in medical rehabilitation. I wanted to explore whether recording a patient's body posture and movements could help clinicians follow recovery over time and identify movement difficulties that still need attention. This gave me a reason to look more closely at how body pose is measured. Cameras can be affected by lighting and occlusion and raise privacy concerns. Millimeter-wave radar does not capture appearance images, but its observations are sparse and noisy; a few inertial measurement units (IMUs) provide local motion information without fully describing the whole body. I therefore chose to study their fusion. The experiments in this paper use public datasets and do not constitute clinical validation.

I would like to thank Shahla Ahmadova, my computer science teacher at Shenzhen College of International Education (SCIE), who supervised this project without payment. With her help, I narrowed the research direction through an initial literature review. I then developed the overall framework and research plan and discussed them with her to confirm the scope. Her guidance also continued into the writing stage. After I completed the first draft, she read it and suggested changes to its organization, wording, and reasoning. I am grateful for the time she spent on these revisions. Her comments helped me see where my explanations were unclear and express the work more carefully.

As the only participating student, I carried out the technical research. The data came from the public mRI and OctoNet datasets, whose authors I thank for making this work possible. I downloaded and screened the data, converted formats, checked the dataset splits, and matched the radar and IMU observations with the corresponding pose labels. No new participants were recruited, and I collected no patient data.

Using published research and technical documentation as references, I wrote and debugged the code, trained the models, and adjusted their parameters. I designed the radar-only, IMU-only, and feature-concatenation comparisons and implemented SRA-AF, including its reliability estimator, bidirectional cross-modal rectification, and fusion Transformer. I ran five-fold cross-subject experiments on mRI and cross-user and cross-scene evaluations on OctoNet, together with ablation studies, temporal-misalignment controls, and action-wise analyses. I calculated and checked MPJPE and PA-MPJPE and organized the outputs for the paper.

Data preparation took more work than I had expected. OctoNet was too large to store and process conveniently on my personal computer. I moved data processing and model training to a server rented through AutoDL (www.autodl.com) and reorganized the files and workflow so that I could complete the experiments.

I originally designed the shifted-IMU experiment as a control: the model would still receive IMU input, but the temporal correspondence with the radar observations would be disrupted. This was intended to test whether the improvement came from useful, time-matched information, rather than simply from supplying additional input data. Unexpectedly, shifting the IMU input improved

performance in an early mRI experiment. I went back through the preprocessing code to check how the two streams had been paired. This involved checking their sampling frequencies, tracing the frame indices used to construct the input sequences, and examining the temporal windows to see whether the paired observations covered the same period. These checks revealed an alignment problem in the preprocessing. After correcting it, I repeated the synchronized and shifted-input comparisons. The artificially shifted inputs then produced higher MPJPE than the synchronized inputs. The control experiment therefore also helped me uncover a preprocessing error that had affected the interpretation of the earlier results.

I drafted every section of the report, from the introduction and related work to the methods, experimental setup, results, and conclusion. I also prepared the figures and tables and revised the manuscript in response to my supervisor's suggestions. Her role was to guide topic selection and review the writing; the implementation, experiments, numerical analysis, and manuscript drafting remained my responsibility. I also acknowledge the researchers whose published methods and technical resources I consulted throughout the project.

From July to September 2026, I frequently used OpenAI Codex (GPT-5.6) during code debugging and manuscript revision. For coding, I used it to examine errors and consider possible corrections. For writing, it helped with language polishing, translation, and document formatting, including revision of this acknowledgment. I carefully reviewed its suggestions and manually checked the text and code before adopting changes. I checked reported experimental results against the recorded outputs and calculations. The final choice of methods, interpretation of results, and responsibility for the submitted paper are mine.