M4Human: A Large-Scale Multimodal mmWave Radar Benchmark for Human Mesh Reconstruction

Supplementary Material

Overview. This supplementary material is organized as follows. (i) We first describe the overall dataset structure and a series of data compression and acceleration techniques for efficient use of our large-scale `M4Human` (cf. Sec. 2). (ii) We then present the action design in our dataset and its intended use cases (cf. Sec. 3). (iii) We provide full implementation details for RT-Mesh, all competing methods, the multi-modal fusion setup, and the HAR benchmarks (cf. Sec. 4). (iv) Finally, we offer a more comprehensive assessment of radar-based HMR across different actions and sensing ranges, together with additional visualizations comparing modalities, results under the more challenging S2 and S3 splits (cf. Sec. 5). We also include a demo.mp4 for video visualization.

1. Ethics Statement

The `M4Human` data has been de-identified by a facial mask. The subject recruitment is voluntary, and the involved subject has been informed that the de-identified data will be made publicly available for research purposes. As far as we know, this research does not endanger any person directly. Nevertheless, it is acknowledged that pose estimation and activity recognition research can potentially be used with malicious intent, such as user behavior monitoring.

2. Dataset Structure

Overall File Structure of Raw Data. We provide two ways to download `M4Human`: (1) a full archive of the raw data and (2) modality-specific preprocessed archives. The full archive preserves the original directory hierarchy so that users can easily inspect the recovered dataset and cross-check samples with the provided annotations (c.f. Fig. 1). For efficient training, we additionally release preprocessed packages in `.lmdb` format that is directly compatible with our dataloaders and avoids repeated on-the-fly preprocessing. (c.f. Fig. 2)

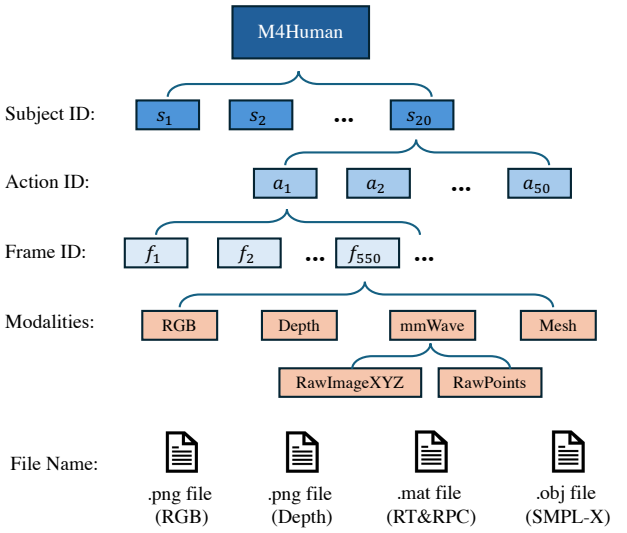
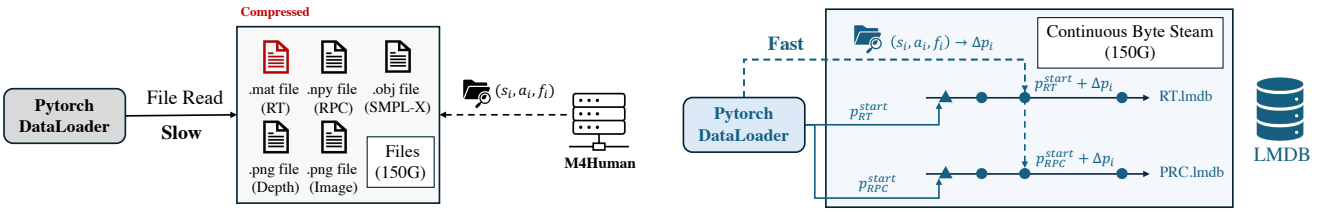


Figure 1. The expanded directory of raw `M4Human` dataset.

LMDB-based Data-Loading Acceleration. Raw radar tensors are highly I/O-intensive and can severely slow down dataset usage. To improve data-loading efficiency and overall usability, we design RT-LMDB, an RT-aware data-loading system built on the Lightning Memory-Mapped Database (LMDB) architecture [3, 8]. LMDB is an embedded transactional key–value store that keeps arbitrary key–value pairs as byte arrays in a contiguous memory-mapped region on disk, enabling fast random access. Similar LMDB-based layouts have been widely adopted in large-scale datasets (e.g., ImageNet [4, 13]) to accelerate data loading.

Our pipeline first applies lossless sparsity-aware compression to each RT. Instead of storing the full dense voxel grid, we keep only non-zero entries and their voxel indices $(id_x, id_y, id_z, \text{intensity})$, so that the original tensor can be exactly reconstructed by scattering intensities back to the corresponding voxels. This compression reduces the storage requirement from approximately 20 TB to about 150 GB.

However, directly reading compressed files from the file system remains time-consuming, while caching all RTs in RAM is prohibitively expensive and does not scale to multi-GPU training (e.g., 4 GPUs would roughly require $4\times$ more RAM). Therefore, we convert all RT files into a single LMDB database, `RT.lmdb`, where each entry stores the compressed bytes of one frame. Each additional modality is stored in its own LMDB (e.g., `RPC.lmdb`, `mesh.lmdb`). In `M4Human`, every sample is uniquely identified by the triplet (s_i, a_i, f_i), denoting subject ID, action ID, and frame ID, which we use as the LMDB key. As illustrated in Fig. 2, the dataloader first queries LMDB to obtain the start pointer p^{start} of the corresponding byte segment, and then reads the bytes directly from disk. Empirically, RT-LMDB achieves random-access speed comparable to keeping all data in RAM, while only storing lightweight pointers in memory, thereby substantially accelerating RT loading for large-scale training. Under a single-GPU, single-process



(a) File-based DataLoader (b) LMDB-based DataLoader

Figure 2. (a) A conventional file-system dataloader repeatedly performs file-name lookup and disk I/O on large `.mat` files (e.g., RT), which quickly becomes a bottleneck at scale. (b) Our LMDB-based system converts all data into a single contiguous byte stream stored in a memory-mapped database, indexed by the key (s_i, a_i, f_i). At query time, the dataloader first retrieves a base pointer p^{start} and an offset Δp_i associated with the key, and then directly reads the required bytes from disk at $p^{\text{start}} + \Delta p_i$, avoiding expensive directory traversal and repeated file open/close operations.

setting, the total data-loading time for `M4Human` is reduced from > 18 hrs/epoch to < 1 hr/epoch.

Modality Dimensionality. For storage and I/O efficiency, RGB images are stored with shape $3 \times 640 \times 480$ and depth images with shape $1 \times 640 \times 480$. Radar point clouds (RPC) are stored as tensors of shape $N \times 4$, where N is the number of detected points (typically ranging from 400–600) and each point is represented by $(x, y, z, intensity)$ coordinate and energy response. The high-resolution raw radar tensor (RT) has shape $121 \times 111 \times 31$, corresponding to the spatial resolutions along the x, y, and z axes, respectively. The value within each voxel corresponds to the intensity.

3. Detailed Action Set

As summarized in Table 1, we define 50 actions in three groups: 30 *in-place* daily/rehabilitation exercises, 5 *sit-in-place* chair-based exercises, and 15 *non-in-place* dynamic activities. Across all protocols, the subject–sensor distance ranges from 0.5,m to 4,m. Actions are performed with different facing directions, and subjects are allowed to move within an area of roughly $6, \text{m}^2$. This setup provides diverse motion patterns, enabling a robust evaluation of model performance.

`M4Human` covers common daily activities and a complete exercise routine, including warm-up, strength training, and post-exercise stretching at different intensity levels. (i) In-place actions. These actions are performed while the subject remains approximately at a fixed location. They are designed to span the main stages of an exercise session. Gentle chest expansions (horizontal/vertical) and left/right trunk twists serve as warm-up and mobility exercises. Squats, various front and side lunges, and jumping or high-knee runs target lower-limb strength and balance. Upper/lower limb extension movements are used as stretching exercises after higher-intensity actions. (ii) Sit-in-place and rehabilitation-oriented actions. Several rehabilitation-oriented patterns are included, such as sideways walking to train lateral stability. Sit-in-place actions are executed on a chair and focus on lower-limb extensions and unilateral arm curls with light loads (e.g., an empty water bottle), mimicking common geriatric and post-surgery rehabilitation routines. (iii) Non-in-place actions. These actions require more complex whole-body coordination and involve substantial translation and direction changes. Examples include straight and curved walking at different speeds, lunge and front-kick walking, as well as sports-like activities such as three-step layups in basketball, badminton with lateral footwork and racket swings, table tennis and volleyball strokes, and free-pace boxing with bouncing footwork and rapid punching.

We expect that this carefully designed action set can support a wide range of real-world applications, including (but not limited to) the following:

Privacy-preserving healthcare applications. Privacy-preserving sensing is critical for elderly care and disease analysis, as monitored individuals are often unwilling to expose their daily activities, whereas there are strong ethical constraints on releasing patient videos captured by cameras for research purposes. For example, current motion analysis for Parkinson's disease commonly relies on RGB-D–based skeleton extraction as a compromise for privacy, but such modalities only provide sparse keypoints information [11]. Although depth cameras are more privacy-friendly than RGB by removing facial textures, they still reveal body shape, posture, and facial outlines. They are also affected by clothing bulk (e.g., whether a person is naked) and can potentially be used for re-rendering and re-identification. By contrast, RF sensors, such as mmWave radar, offer truly privacy-preserving sensing. They do not expose the above visual cues, yet still provide abundant motion information, making them a promising modality for data collection and human-motion analysis in healthcare scenarios. Our dataset is the first to demonstrate the potential of extracting 3D dense human meshes for more dynamic activities in an open room, thereby pushing RF sensing toward higher-fidelity perception, better generalizability, and more advanced algorithm

Table 1. Action set in M4Human.

Action ID	Action Name	Protocol	Action ID	Action Name	Protocol
1	Chest expanding horizontally	In-Place Daily	26	Right limbs extension	In-Place Rehab
2	Chest expanding vertically	In-Place Daily	27	Jumping up	In-Place Daily
3	Left side twist	In-Place Daily	28	Tai Chi	In-Place Rehab
4	Right side twist	In-Place Daily	29	High knees	In-Place Rehab
5	Raising left arm	In-Place Daily	30	Neck rotations while standing with both legs	In-Place Rehab
6	Raising right arm	In-Place Daily	31	Slowly stand up and sit down from a chair	Sit-In-Place Daily
7	Waving left arm	In-Place Daily	32	Sit and left leg kick/extension	Sit-In-Place Rehab
8	Waving right arm	In-Place Daily	33	Sit and right leg kick/extension	Sit-In-Place Rehab
9	Picking up things	In-Place Daily	34	Sit and raise left dumbbell (arm curls)	Sit-In-Place Rehab
10	Throwing toward left side	In-Place Daily	35	Sit and raise right dumbbell (arm curls)	Sit-In-Place Rehab
11	Throwing toward right side	In-Place Daily	36	Walk in straight line (fast)	Non-In-Place Daily
12	Kicking toward left direction using right leg	In-Place Daily	37	Walk in straight line (slow)	Non-In-Place Daily
13	Kicking toward right direction using left leg	In-Place Daily	38	Walk in curves (fast)	Non-In-Place Daily
14	Bowing forward	In-Place Daily	39	Walk in curves (slow)	Non-In-Place Daily
15	Stretching and relaxing in free form	In-Place Daily	40	Non-in-place Lunge	Non-In-Place Sports
16	Mark time	In-Place Rehab	41	Front kick walk	Non-In-Place Sports
17	Left upper limb extension	In-Place Rehab	42	Sideways walking	Non-In-Place Daily
18	Right upper limb extension	In-Place Rehab	43	Badminton	Non-In-Place Sports
19	Left front lunge	In-Place Rehab	44	Table tennis (ping pong)	Non-In-Place Sports
20	Right front lunge	In-Place Rehab	45	Baseball	Non-In-Place Sports
21	Both upper limbs extension	In-Place Rehab	46	Volleyball	Non-In-Place Sports
22	Squat	In-Place Rehab	47	Free-place Boxing	Non-In-Place Sports
23	Left side lunge	In-Place Rehab	48	Walk in straight line & pick up item	Non-In-Place Daily
24	Right side lunge	In-Place Rehab	49	Walk in curve & pick up item	Non-In-Place Daily
25	Left limbs extension	In-Place Rehab	50	Basketball	Non-In-Place Sports

design.

Smart Home and Human Computer Interaction. The proposed dataset includes a rich action set covering both in-place and non–in-place daily movements, such as in-place waving, hand raising, throwing, kicking, and non-in-place walking, at different paces/speeds and sitting-standing. These actions naturally support standard action recognition and whole-body gestures, enabling RF-based smart-home applications such as device on/off control, adaptive energy management, and presence/behavior-aware automation. Moreover, the action/gesture samples provide a basis for privacy-preserving HCI, where users can interact with TVs, AR/VR displays, or service robots through RF-sensed body movements without exposing their visual appearance.

VR Rendering and Fitness Gaming. Thanks to the high-precision marker-based motion capture system and synchronized RGB-D streams, our dataset provides rich annotations for a wide range of 3D rendering and reconstruction tasks. The annotations include 3D SMPL-X human meshes, 3D dense poses, and RGB-textured human meshes, which together enable learning-based 3D human reconstruction from RGB-D, from privacy-preserving mmWave radar, and from multi-modal fusion. Beyond static reconstruction, the diverse action set (covering in-place and non–in-place daily activities, fitness/rehab exercises, and free-space sports-like motions) facilitates the learning of human motion priors, such as coordinated upper–lower limb movements. Consequently, the dataset can support downstream tasks including motion generation, stochastic motion prediction, avatar driving from RF signals, and controllable motion reenactment. This is particularly valuable for VR/AR content creation and fitness gaming, where plausible, temporally coherent human motion is required, as well as for privacy-preserving VR interaction without exposing the user's visual appearance.

4. Benchmark Implement Details

4.1. Overall Training Specifications

Unless specified otherwise, all methods share the same HMR prediction head to regress SMPL-X parameters and follow a unified training setup. All models are implemented in PyTorch and trained for 100 epochs with a batch size of 64. We use the Adam optimizer [10] with a learning rate of 2×10^{-4}, momentum 0.9, and a learning-rate decay factor of 0.1 applied every 5 epochs. We further apply norm-based gradient clipping to stabilize training. All methods are optimized with a unified mesh reconstruction loss $\mathcal{L}_{\text{mesh}}$, with loss weights set to $\lambda_\alpha = 1$, $\lambda_\beta = 0.3$, $\lambda_\tau = 10$, $\lambda_\theta = 15$, and $\lambda_g = 0.5$. These hyperparameters are chosen empirically; more advanced tuning strategies (e.g., Bayesian optimization) may further improve performance and are left for future work. All experiments are conducted on a local Ubuntu 20.04 server equipped with 4 NVIDIA RTX 3090 GPUs, an Intel Xeon(R) Platinum 8474C processor (15 cores), and 128 GB RAM. Training RT-Mesh with 4 GPUs in parallel takes approximately one day.

4.2. Implementation of RT-Mesh

BEV 2D Localization. We reshape X_{RT} into a BEV tensor along the X–Y plane by concatenating the remaining axes into channels, i.e., $C_{\text{2D}} = Z \times T$. We adopt BEV since

Table 2. Radar-based HMR results on `M4Human` using state-of-the-art indoor human sensing HMR/HPE models. All methods are evaluated on the S1, S2, and S3 splits under four protocols: P1 In-Place (IP), P2 Sit-In-Place (SIP), P3 Non-In-Place (NIP), and ALL (all actions).

Modality	Method	Protocol	S1 (Random)				S2 (Cross-Sub)				S3 (Cross-Act)				Efficiency	
			MVE	MJE	MRE	TE	MVE	MJE	MRE	TE	MVE	MJE	MRE	TE		
RPC	mmMesh [16]	**IP**	105.9	88.0	11.0	57.2	149.4	132.6	16.1	96.2	146.0	122.9	16.1	78.5	Latency (ms):	3.53
		SIP	183.3	164.2	10.7	92.8	202.8	173.5	12.2	100.5	194.5	170.9	10.8	100.3	Param. (M):	41.45
		NIP	201.4	170.6	15.4	105.4	226.6	188.6	17.6	113.0	223.3	187.4	18.7	115.8	GFLOPs:	2.87
		ALL	132.7	112.1	11.8	70.4	170.1	147.9	16.0	100.0	173.8	146.9	16.4	91.8		
	P4Transformer [7]	**IP**	71.5	59.0	8.0	37.3	129.3	114.3	14.9	81.1	132.3	111.3	15.9	69.9	Latency (ms):	7.17
		SIP	115.2	109.3	9.5	69.1	142.8	134.6	11.9	90.0	132.0	124.0	10.3	76.9	Param. (M):	129.01
		NIP	139.7	121.7	13.5	77.9	180.2	158.8	17.0	103.1	184.4	157.9	18.3	96.1	GFLOPs:	11.76
		ALL	**89.5**	**76.6**	9.2	48.6	140.8	125.2	15.0	86.4	147.8	126.4	16.1	78.4		
RT	RT-Pose [9]	**IP**	80.0	66.5	8.7	43.0	135.9	119.7	15.3	88.6	133.7	112.0	15.7	71.6	Latency (ms):	39.58
		SIP	130.2	126.0	9.9	81.6	152.4	145.1	12.0	100.9	139.5	132.6	10.5	86.1	Param. (M):	6.05
		NIP	158.5	139.3	14.2	92.2	188.4	164.4	17.0	109.0	196.0	169.0	18.4	110.4	GFLOPs:	50.67
		ALL	100.7	87.0	9.9	56.7	148.1	131.2	15.3	93.9	152.8	131.0	16.0	84.6		
	RETR [18]	**IP**	73.2	58.8	7.5	34.8	159.4	138.9	16.6	95.4	143.0	118.3	16.5	70.9	Latency (ms:)	17.87
		SIP	133.6	126.2	10.0	80.4	169.4	156.7	12.1	105.7	153.6	143.6	10.8	93.1	Param. (M):	52.42
		NIP	162.6	140.4	14.2	90.1	206.1	178.5	17.4	114.1	207.0	177.1	18.8	112.3	GFLOPs:	3.01
		ALL	97.1	81.8	**9.1**	50.4	169.7	148.6	16.3	100.1	163.1	138.3	16.6	85.4		
	RT-Mesh (Ours)	**IP**	72.4	59.1	8.3	36.2	123.6	109.6	14.7	82.1	128.5	107.2	15.3	68.1	Latency (ms):	2.74
		SIP	118.1	112.2	9.7	68.4	138.5	130.8	12.0	88.4	126.5	120.0	10.4	73.9	Param. (M):	63.25
		NIP	142.0	123.1	13.8	77.1	173.6	151.8	16.9	98.8	178.2	152.6	18.0	92.8	GFLOPs:	2.60
		ALL	90.9	77.2	9.6	**47.6**	**135.1**	**120.2**	**14.9**	**86.1**	**143.1**	**122.0**	**15.6**	**76.0**		

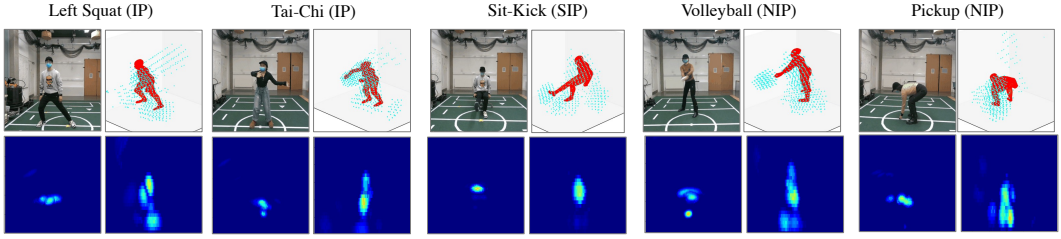


Figure 3. Sample visualizations of different modalities and ground-truth annotations. Our calibration system achieves good alignment between RGB-D, radar modalities, and the ground-truth meshes.

(X, Y) provides the highest spatial resolution for accurate human localization. The BEV tensor is encoded by 2D convolutions with downsampling ratio N_1, resulting in a feature map that is patchified into $\frac{X}{N_1} \times \frac{Y}{N_1}$ tokens. A lightweight 2D transformer performs patch-wise self-attention to amplify human responses. The output tokens are globally pooled and flattened into a unified BEV descriptor, from which we regress the human's BEV coordinates $(\hat{x}, \hat{y})$.

Local 3D Regression. Centered at $(\hat{x}, \hat{y})$, we crop a fixed-size 3D RoI $(\Delta X, \Delta Y, \Delta Z)=(24, 24, 31)$ from X_{RT}. This compact RoI enables fast processing and reduces clutter/multipath outside the human foreground. The cropped 3D tensor passes through an RoI-specific 3D transformer stack: a short 3D convolutional stem with downsampling ratio N_2 captures local 3D structure, the features are patchified, and a 3D transformer applies high-resolution self-attention. The output tokens are globally max-pooled and flattened to form the final 3D mesh feature. Finally, an MLP-based HMR head regresses SMPL-X parameters $(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\theta}, \hat{g})$ for reconstructing the SMPL-X mesh.

Training Objective. We jointly optimize BEV localization and 3D mesh regression:

$$\mathcal{L} \;=\; \lambda_{2D}\,\mathcal{L}_{2D} \;+\; \lambda_{\text{mesh}}\,\mathcal{L}_{\text{mesh}}\;, \tag{1}$$

where $\mathcal{L}_{2D} = \|\hat{x} - \tau[0]\|_2 + \|\hat{y} - \tau[1]\|_2$ supervises 2D BEV localization. The 3D mesh loss is the weighted sum over SMPL-X components:

$$\begin{aligned}\mathcal{L}_{\text{mesh}} = \lambda_\theta\,\mathcal{L}_{\text{rot}}(\hat{\theta}, \theta) \;+\; \lambda_\alpha\,\mathcal{L}_{\text{rot}}(\hat{\alpha}, \alpha) \;+\; \lambda_\beta\,\|\hat{\beta} - \beta\|_2^2 \\ +\; \lambda_\tau\,\|\hat{\tau} - \tau\|_1 \;+\; \lambda_g\,\text{BCE}(\hat{g}, g)\;,\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{rot}}$ denotes a geodesic rotation loss [19] and $\text{BCE}(\cdot)$ is the binary cross-entropy loss.

4.3. Implementations of Single-Modal Methods

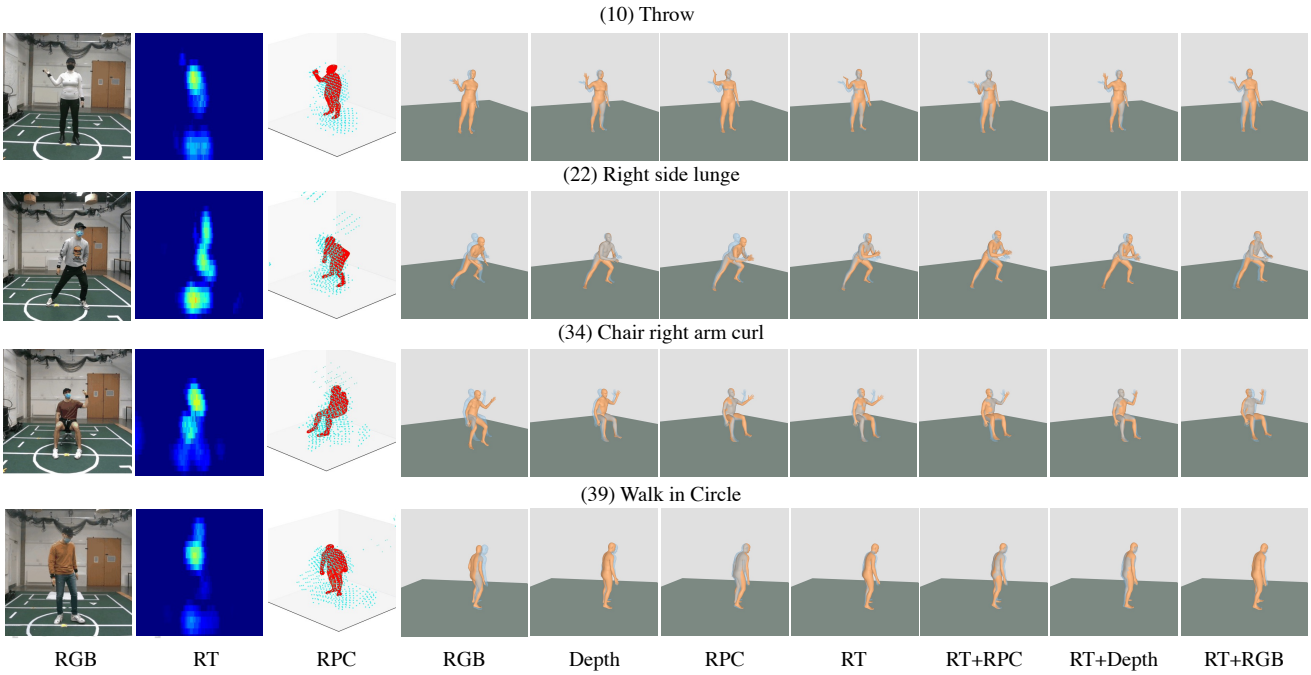


Figure 4. Visualization of single-modality predictions and multi-modal fusion. Predicted meshes are shown in orange and ground truth in blue; higher overlap indicates higher accuracy. The performance gap between line-of-sight (LoS) RGB-D and radio-frequency (RF) modalities is smaller than expected, thanks to the high-resolution radar in `M4Human`, making radar-based HMR feasible. Notably, RGB-only HMR struggles with accurate depth estimation, leading to larger reconstruction errors.

RGB. We use Detectron2 [15] to detect human-centered square bounding boxes, resized to 224×224. The cropped RGB patches are passed to a pretrained TokenHMR [6] model to predict SMPL-X parameters $(\hat{\alpha}_{\text{img}}, \hat{\beta}_{\text{img}}, \hat{\theta}_{\text{img}}, \hat{g}_{\text{img}})$, representing global orientation, body shape, pose, and gender, respectively, under the cropped-image coordinate system.

To localize the subject under the full-image coordinate, the model additionally predicts a normalized triplet $(\hat{s}, \hat{u}, \hat{v})$, where $\hat{s}$ is the weak-perspective scale (monotonically related to inverse depth) and $(\hat{u}, \hat{v}) \in [-1, 1]$ are normalized crop coordinates. To undo the crop to obtain full-image pixels, we leverage (i) the human detector's outputs: box size b and box center (c_x, c_y), and (ii) the camera intrinsics: focal length f (pixels) and principal point $(W/2, H/2)$. The human center pixels is computed as:

$$x_{\text{pix}} = c_x + \frac{b_s}{2}\,\hat{u},\; y_{\text{pix}} = c_y + \frac{b_s}{2}\,\hat{v},\; \hat{z}_{\text{img}} = \frac{2f}{b_s}, \quad (3)$$

where $b_s = b * \hat{s}$ is the effective box size and $\hat{z}_{\text{img}}$ is the estimated depth. We then apply pinhole back-projection with the weak-to-perspective depth relation to obtain full translation $\hat{\tau}_{img} = (\hat{x}_{\text{img}}, \hat{y}_{\text{img}}, \hat{z}_{\text{img}})$ under the world coordinates:

$$\hat{x}_{\text{img}} = \frac{x_{\text{pix}} - \frac{W}{2}}{f}\,\hat{z}_{\text{img}}, \qquad \hat{y}_{\text{img}} = \frac{y_{\text{pix}} - \frac{H}{2}}{f}\,\hat{z}_{\text{img}}. \quad (4)$$

For the best performance, we fine-tune the prediction heads for $(\hat{\alpha}_{\text{img}}, \hat{\beta}_{\text{img}}, \hat{\tau}_{\text{img}}, \hat{g}_{\text{img}})$ on our dataset to improve depth and localization accuracy. We keep the VQ-based pose tokenizer/decoder fixed, as it is pretrained on the large-scale AMASS dataset and provides strong prior performance. This fine-tuning improves the MVE from 190 mm to 97 mm on our benchmark, as reported in the main paper.

Depth. Following the same pinhole back-projection used in the RGB pipeline, depth frames (which provide measured rather than estimated depth) are back-projected with camera intrinsics K to obtain per-pixel dense 3D points in the camera/world frame, yielding a dense point cloud. We then adopt a widely adopted point-based encoder, P4Transformer [7], to process the depth-derived point cloud, which is adopted in prior multimodal benchmark mmBody [2] and LiDAR-based HMR benchmark (e.g., RELI11D). For a fair comparison with the RPC modality, we use the same HMR prediction head to regress SMPL-X parameters and optimize with the same mesh loss $\mathcal{L}_{\text{mesh}}$.

RPC and RT. For RPC preprocessing, we follow prior designs [1, 12, 17] and adopt a sliding window that aggregates T=4 adjacent frames into a single input, which helps alleviate point-cloud sparsity and occasional body-part

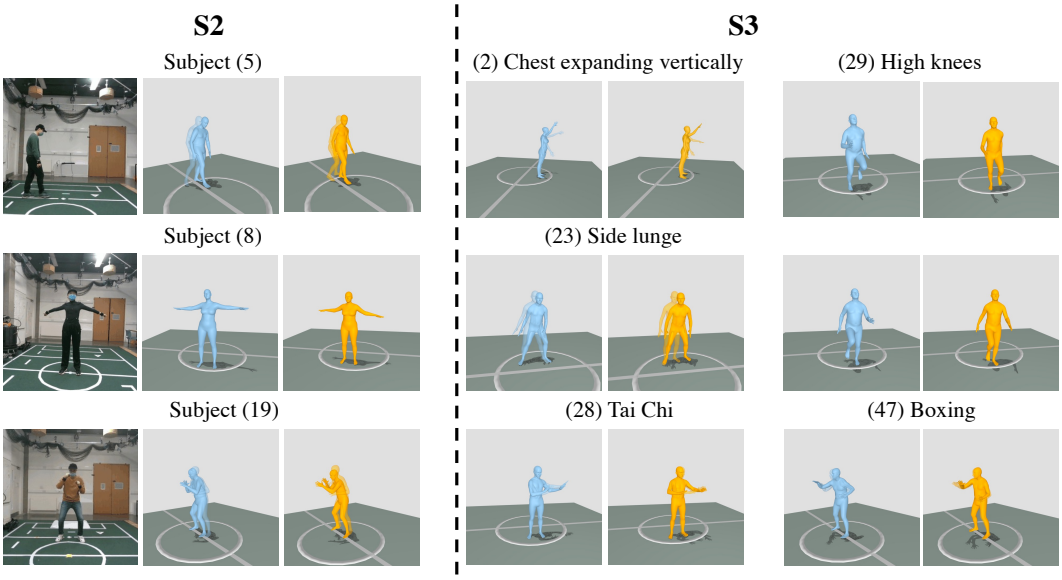


Figure 5. Visualization of RT-Mesh under S2 (cross-subject) and S3 (crpss-action) split. RT-Mesh demonstrates good generalization to unseen subjects and actions.

miss-detections in individual frames. Since the number of points varies across frames, we apply zero-padding to a fixed size of 1,000 points per aggregated frame, yielding a unified tensor shape for efficient PyTorch batching. We use P4Transformer [7] as the point-cloud baseline, as it achieves state-of-the-art performance on previous RPC-based HMR benchmarks (e.g., mmBody [2]) and is widely employed in point-based HMR (e.g., LiDAR). For RT preprocessing, we adopt the same T=4 sliding-window strategy to ensure a fair comparison, resulting in a 4D radar tensor. For RETR, we reshape the 4D tensor X_{RT} into two views, $X \times Y \times (Z \times T)$ for the horizontal (BEV) view and $Y \times Z \times (X \times T)$ for the vertical (XZ) view, and follow the official transformer encoder–decoder design for multi-view feature fusion. For RT-Pose, which relies on 3D convolutions and an HRNet-style head, we treat the temporal dimension T as the input feature (channel) dimension.

4.4. Implementation of Multi-modal Fusion

We adopt a simple feature-level fusion strategy to clearly assess the benefit of combining modalities. Let $f^{(m)} \in \mathbb{R}^{d_m}$ denote the final encoder-extracted global feature from modality m (e.g., RGB, depth, RPC, or RT), where $d_m = 1024$ is the feature dimension. Before being fed into the HMR prediction head, we concatenate the features from two modalities along the channel dimension:

$$f_{\text{fuse}} = \left[f^{(m_1)} \,\|\, f^{(m_2)} \right] \in \mathbb{R}^{2d_m}.$$

The fused feature f_{fuse} is then input to the same HMR prediction head architecture as in the single-modality case, with only the input dimension changed to $2d_m$ to match the fused feature size.

For fairness, we keep all modality backbones, the mesh loss $\mathcal{L}_{\text{mesh}}$, and the overall training setup identical to the unimodal experiments. This minimal design ensures that the performance gains arise from the inclusion of additional modalities rather than from altered training or loss configurations. We expect and encourage that more advanced multi-modal fusion mechanisms (e.g., attention-based fusion, gating, or cross-modal transformers) could further improve performance and leave such designs for future research.

4.5. Implementation of HAR

`M4Human` can also be used for skeleton-based human action recognition (HAR) across different modalities. Here, we construct a new HAR benchmark based on 3D skeletons extracted from our predicted SMPL-X meshes. Following the official splits used for HMR, we evaluate HAR under S1 (random split) and S2 (cross-subject split). For each sequence in S1 and S2, we first run HMR and then extract 3D joint trajectories from the predicted meshes. Each sequence is segmented into clips of 64 consecutive frames (approximately 5 seconds), yielding around *14,000* training clips, *1,000* validation clips, and *3,200* testing clips.

We implement several representative skeleton-based HAR methods: a simple 2D CNN [5] operating on the temporal–joint (T×J) dimension, AGCN [14] (CVPR'19), and BlockGCN [20] (CVPR'24). For AGCN and BlockGCN, we initialize from models pretrained on NTU-RGBD and fine-tune them on our dataset. We train the model for 1,000 epochs with a batch size of 32, using the Adam [10] opti-

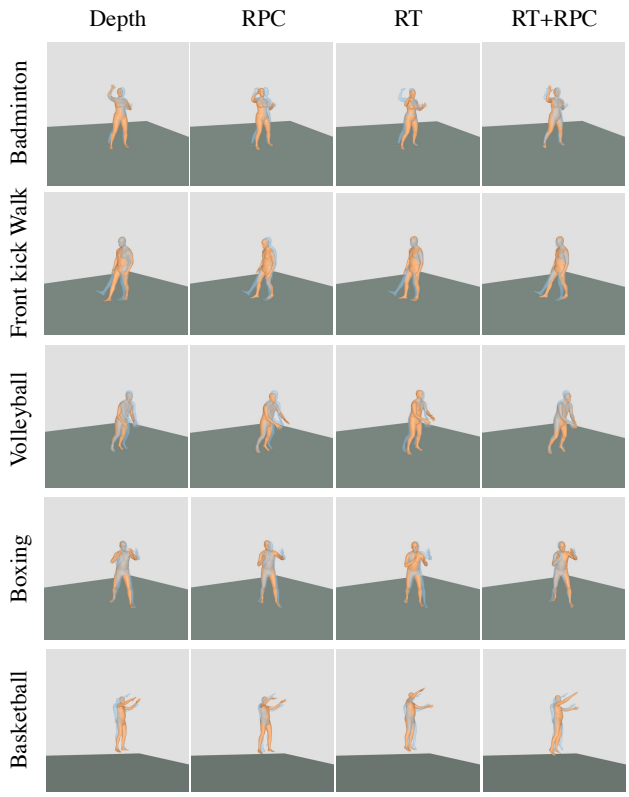


Figure 6. Visualization of RT–RPC fusion versus depth-based HMR under dynamic non-in-place actions. Predicted meshes are shown in orange and ground truth in blue; a higher overlap indicates higher accuracy. Fusing RT and RPC consistently improves radar-based HMR, narrowing the performance gap to the depth modality.

mizer with momentum 0.9, an initial learning rate of 0.01, and weight decay of 10^{-4}. We report Top-1 and Top-5 classification accuracy as our evaluation metrics. As shown in Fig. 9, the confusion matrices obtained using skeletons derived solely from the RT modality exhibit strong classification performance on both S1 and S2, indicating that our radar-predicted meshes are accurate enough to support challenging downstream HAR tasks.

5. More Qualitative Visualization

5.1. Qualitative Comparison Across Modalities

As illustrated in Fig. 4, we visualize HMR results for different single modalities (RGB, depth, RPC, RT) as well as their fused predictions. Predicted meshes are overlaid in orange and ground truth in blue. Due to the lack of explicit depth measurements, RGB-only HMR often exhibits noticeable depth offsets, which leads to biased global localization and reduced overlap with the ground-truth meshes. In contrast, depth-based HMR yields more accurate reconstructions, benefiting from high-precision per-pixel depth. RPC- and RT-based methods achieve comparable localization quality, even for challenging actions such as side lunges, demonstrating that high-resolution mmWave radar can support fine-grained body pose recovery. Moreover, multi-modal fusion generally outperforms single-modality models: in particular, RT+RPC fusion produces visibly more accurate meshes than either RT or RPC alone. Additional examples of dynamic non-in-place motions are shown in Fig. 6, where RT+RPC consistently surpasses single-modality radar and approaches the accuracy of depth-based HMR. These qualitative results indicate that radar is capable of high-fidelity human sensing. At the same time, as also seen in Fig. 4, radar representations do not reveal identifiable appearance or background details, making RF-based HMR a promising solution for privacy-sensitive applications.

5.2. HMR Visualization for Cross-Subject and Cross-Action

As illustrated in Fig. 5, we present qualitative results for the more challenging S2 (cross-subject) and S3 (cross-action) settings. In both cases, we observe that body motion and pose dynamics can be reasonably well recovered, whereas the estimated body shape parameters $\boldsymbol{\beta}$ are less accurate and sometimes inconsistent across frames. We hypothesize that this limitation stems from the lack of rich appearance cues and fine-grained geometric correspondences in radar data compared to RGB-D, which also contributes to its privacy-preserving nature.

For S3, our model shows promising generalization to unseen daily activities such as vertical chest expansion and high knees, even though these actions do not appear in the

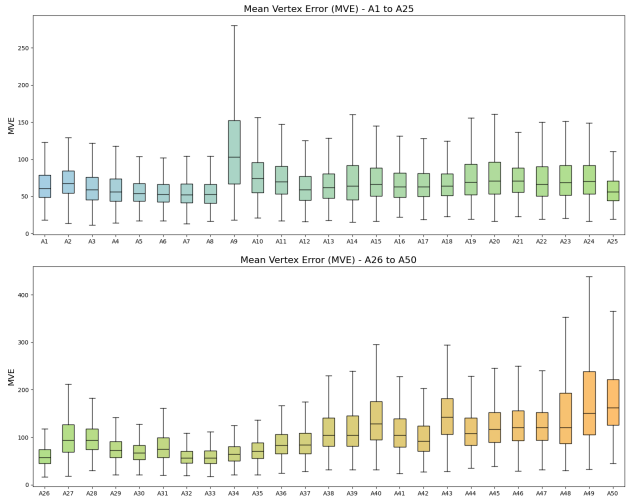


Figure 7. Mean Vertex Error (MVE) across different action types. Dynamic non-in-place actions exhibit higher MVE, highlighting their increased difficulty and suggesting the need for more advanced motion modeling and stronger prior knowledge.

training set. This suggests that the proposed RT-Mesh is able to capture meaningful spatio-temporal body dynamics that transfer to novel motion patterns, making it suitable for generalizable human monitoring. However, the model still struggles with more complex motions, such as side lunges or highly dynamic non-in-place boxing, which require a deeper understanding of human kinematics and body coordination. We believe that the future integration of stronger human priors and more expressive motion models will improve generalization to such challenging actions.

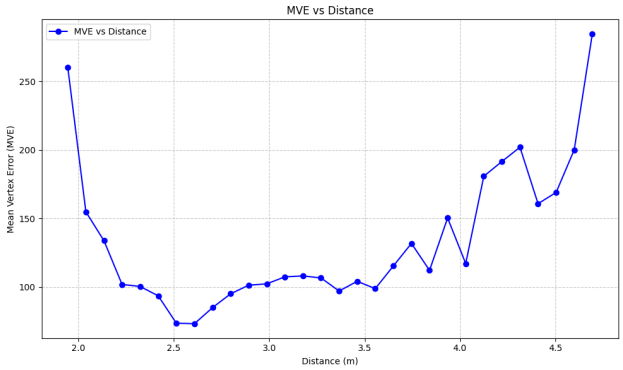


Figure 8. Mean Vertex Error (mm) versus sensing distance (m). The optimal sensing range is around 2.0 to 4.0 m.

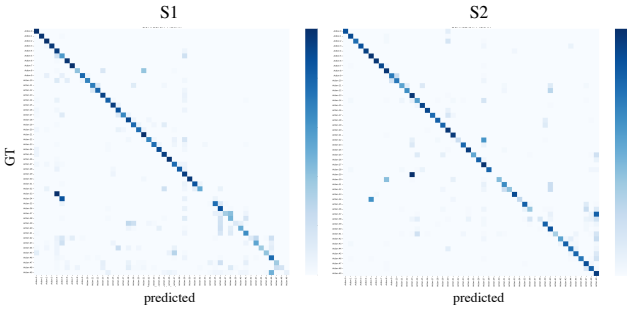


Figure 9. Confusion Matrix visualization for RT-based HAR using the best performing AGCN, under S1 and S2 splits.

5.3. HMR Performance Across Actions and Sensing Distances

As shown in Fig. 7, we report Mean Vertex Error (MVE) for all 50 actions using box plots grouped by action type. Our models perform well on simple in-place and sit-in-place exercises, with errors typically below 100,mm. In contrast, non-in-place dynamic actions are noticeably more challenging, as they involve more complex motion coordination and kinematics and therefore require richer motion understanding and stronger priors, which we leave for future study.

Fig. 8 further plots MVE as a function of sensing distance, illustrating how radar-based HMR accuracy varies with subject–sensor range. Performance generally degrades at larger distances, where fewer foreground points are captured, and also at very close ranges, where the human body can exceed the field of view. The best accuracy is obtained at medium distances, approximately 2.0–4.0 m. We expect that more advanced algorithmic designs could further extend the effective sensing range.

References

[1] 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*, pages 889–894, 2022. 5

[2] 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*, pages 3501–3510, 2022. 5, 6

[3] Howard Chu. Mdb: A memory-mapped database and backend for openldap. In *Proceedings of the 3rd International Conference on LDAP, Heidelberg, Germany*, page 34, 2011. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[5] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, pages 579–583. IEEE, 2015. 6

[6] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1323–1333, 2024. 5

[7] 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*, pages 14204–14213, 2021. 4, 5, 6

[8] Gavin Henry. Howard chu on lightning memory-mapped database. *Ieee Software*, 36(06):83–87, 2019. 1

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

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3, 6

[11] Catherine Morgan, Emma L Tonkin, Alessandro Masullo, Ferdian Jovan, Arindam Sikdar, Pushpajit Khaire, Majid Mirmehdi, Ryan McConville, Gregory JL Tourte, Alan Whone, et al. A multimodal dataset of real world mobility activities in parkinson's disease. *Scientific data*, 10(1):918, 2023. 2

[12] M Mahbubur Rahman, Ryoma Yataka, Sorachi Kato, Pu Wang, Peizhao Li, Adriano Cardace, and Petros Boufounos. Mmvr: Millimeter-wave multi-view radar dataset and benchmark for indoor perception. In *European Conference on Computer Vision*, pages 306–322. Springer, 2024. 5

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[14] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019. 6

[15] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[16] 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*, pages 269–282, 2021. 4

[17] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *arXiv preprint arXiv:2305.10345*, 2023. 5

[18] Ryoma Yataka, Adriano Cardace, Perry Wang, Petros Boufounos, and Ryuhei Takahashi. Retr: Multi-view radar detection transformer for indoor perception. *Advances in Neural Information Processing Systems*, 37:19839–19869, 2024. 4

[19] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 4

[20] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2049–2058, 2024. 6