D6.5 – Multi-subject gesture recognition benchmark test and prototype

Grant Agreement Number	101099491
Action Acronym	HOLDEN
Action Title	Ethical Design of Holography with Dense wireless Networks (HOLDEN)
Funding Scheme	HORIZON-EIC-2022-PATHFINDEROPEN-01
Version date of the Annex I against which the assessment will be made	13/12/2022
Start date of the project	1/6/2023
Due date of the deliverable	31/5/2026
Actual date of submission	22/5/2026
Responsible	AALTO
Contributors	AALTO
Dissemination level	Public

European Innovation Council

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement No: 101099491.

Authors in alphabetical order

Full Name	Organisation	E-mail
Yuqing Song	AALTO	yuqing.song@aalto.fi
Dariush Salami	AALTO	dariush.salami@aalto.fi

Change History

Version	Date	Status	Author (Company)	Description
0.1	22/5/2026		AALTO	Version 1 Complete

Executive Summary

This deliverable reports two connected project outcomes for multi-subject gesture recognition using millimeter-wave (mmWave) radar sensing. The first outcome is a single-radar fine-grained gesture-recognition benchmark designed to evaluate whether one radar can separate and recognize subtle gestures when multiple subjects stand at controlled spatial positions. The benchmark defines two-, three-, and four-person layouts, assigns subjects to fixed positions, and specifies balanced gesture-combination protocols for repeatable testing. The final test results show that the model correctly estimates the number of persons in all test scenes and achieves 99.65% slot-level gesture accuracy. The scene-level accuracy reaches 99.02%, demonstrating that the angle-ordered slot formulation can reliably associate gestures with different persons in multi-person radar scenes.

The second outcome is a distributed multi-radar arm-gesture recognition prototype. It combines multiple range-constrained mmWave radars to produce a global point-cloud representation, separates individual subject trajectories, extracts gesture segments, and classifies gestures with graph-based spatiotemporal learning. This multi-radar outcome demonstrates that distributed radar sensing can support simultaneous multi-target tracking and gesture recognition in crowded indoor environments, thereby complementing the single-radar benchmark. In this work, we propose algorithmic methods for aggregating, cleaning, and temporally sequencing wireless sensing data collected by spatially distributed, range-constrained mmWave radars. We present a joint multi-target tracking and gesture recognition framework in which local radar measurements are fused into a global point cloud representation of the monitored environment. Individual subject trajectories are isolated from this representation, enabling accurate tracking, person counting, and gesture recognition from point cloud sequences. We evaluate the system in a 5m x 5m area under varying radar densities and coverage conditions. In the optimal configuration with eight radars, the system achieves tracking errors of 20cm for a single subject and 35 cm for six simultaneously moving subjects, while achieving 95.85% gesture recognition accuracy despite five interfering subjects.

Together, the two outcomes form a staged benchmark-and-prototype pipeline: first evaluate fine-grained recognition under a minimal single-radar setting, then extend the system to robust multi-subject operation through multi-view radar fusion.

Table of Contents

Introduction

Gesture recognition provides a natural interface for contactless human-computer interaction, robot control, smart-space interaction, and assisted living applications. Compared with cameras, mmWave radar can operate under poor lighting conditions and can preserve visual privacy because it observes sparse reflections rather than identifiable visual appearance. However, fine-grained gesture recognition remains challenging in multi-subject environments. A single radar has limited angular resolution and limited field of view, while multiple subjects may occlude one another or produce overlapping point-cloud reflections.

MmWave radar sensing has emerged as a powerful modality for device-free perception of human activity, enabled by advances in compact radar hardware and signal processing. Nevertheless, the sensing range of mmWave radar devices is inherently limited by factors such as field of view, bandwidth, antenna configuration, and transmission power [1]. As a result, a single radar typically covers only a small portion of an indoor environment, and performance may further degrade in the presence of occlusions caused by walls, furniture, or human bodies [2]–[4].

This deliverable therefore organizes the work into two complementary outcomes: a single-radar fine-grained gesture recognition benchmark and a distributed multi-radar prototype for multi-target tracking and arm-gesture recognition.

1.1. Single radar gesture recognition

The first outcome is a single-radar fine-grained benchmark. It uses a controlled room setup with one mmWave radar, MMWCAS, and multiple standing participants to evaluate how well fine gestures can be recognized when subjects are placed at different spatial positions.

This single-radar setting is useful for understanding the baseline capability and limitations of radar-based gesture recognition. Since the radar observes the scene from only one viewpoint, the captured point clouds are strongly affected by the subject's position, orientation, distance from the radar, and potential occlusion by other participants. These factors make fine-grained gesture recognition difficult, especially when multiple subjects are present in the same sensing area.

By evaluating gesture recognition under different spatial configurations, the single-radar benchmark provides insight into how subject placement and view-dependent sensing affect recognition performance. It also serves as a reference point for assessing the benefits of multi-radar sensing.

1.2. Multi radar gesture recognition

The second outcome is a distributed multi-radar prototype for arm-gesture recognition. It addresses the limitations of single-view sensing by combining multiple radar viewpoints. In this setting, data collected from spatially distributed mmWave radars are temporally aligned and transformed into a shared coordinate system, enabling consistent multi-view perception.

Collaborative mmWave radar sensing introduces fundamental challenges. Radar measurements are viewpoint-dependent and must be interpreted relative to each device's position and orientation. Therefore, accurate spatiotemporal alignment and fusion of point cloud data are required. Moreover, real-world indoor environments often involve multiple users performing different activities simultaneously, which significantly complicates perception [5].

The proposed multi-radar framework jointly supports multi-target tracking and gesture recognition. By fusing point cloud measurements from distributed radars into a global representation, the system extends the effective sensing range, improves robustness through redundant observations, and enables simultaneous person counting, multi-target tracking, and multi-target gesture recognition in realistic multi-user scenarios.

The main contributions are as follows:

1. A single-radar fine-grained gesture recognition benchmark that evaluates the feasibility and limitations of recognizing gestures from one radar viewpoint in multi-subject spatial configurations.

2. A distributed multi-radar sensing framework that jointly supports multi-target tracking and arm-gesture recognition by fusing measurements from multiple mmWave radar devices.

3. A spatiotemporal alignment and fusion method that transforms local radar observations into a shared global coordinate system, enabling consistent multi-view perception across distributed radars.

4. A spatiotemporal adaptive clustering method that exploits temporal continuity in point cloud data to improve multi-target separation and trajectory tracking.

5. A comprehensive experimental evaluation across multiple scenarios and locations, involving up to eight distributed radar devices. The system achieves tracking errors as low as 20 cm and a gesture recognition accuracy of 95.85% under simultaneous multi-person motion.

The collected dataset and source code are available at https://github.com/wenliangwanru/MultiTarget-Tracking-GestureRecognition.

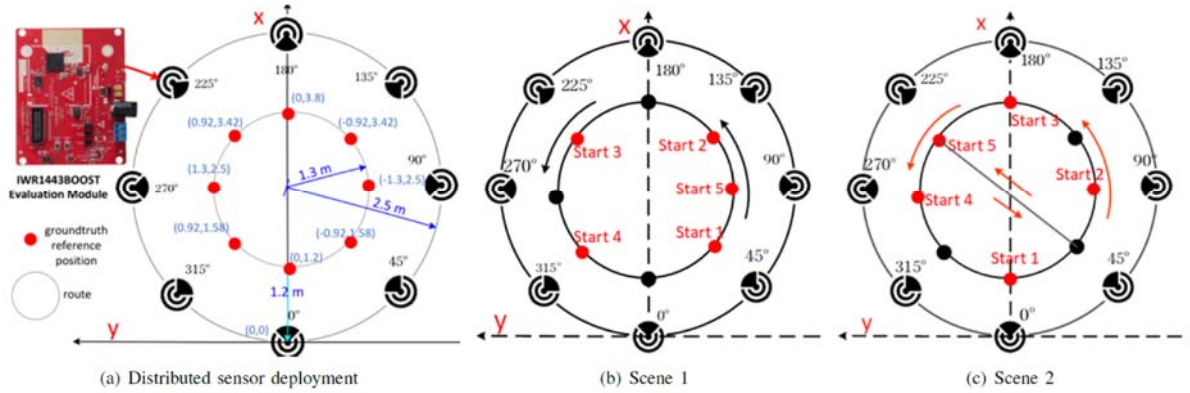


Fig. 1: Experimental validation. (a) Distributed sensor deployment. (b) Scene 1: Subjects follow a predefined path for multi-target walking-only tracking. (c) Scene 2: Subjects walk along the full path back to the start, then perform gestures.

2. Related Work

2.1. MmWave Radar for Device-Free Indoor Perception

MmWave radar sensing has become a central modality for device-free indoor perception [6], enabling robust sensing under varying lighting conditions and preserving visual privacy [7]. Prior work has demonstrated the use of mmWave radar for fundamental perception tasks such as indoor localization and positioning [8], orientation estimation [9], and general activity recognition [10], forming the foundation for intelligent human–environment interaction systems [11]. These capabilities have further enabled applied sensing systems in domains such as contact tracing [12] and remote physiological monitoring [13], highlighting the practical relevance of radar-based sensing in healthcare and smart environments.

2.2. Fine-Grained and Micro-Motion Sensing

Early radar sensing research primarily focused on large-scale body motion and coarse activity recognition, including fall detection [14] and full-body movement analysis. With advances in radar resolution, signal processing, and learning-based methods, recent work has shifted toward fine-grained and micro-motion sensing. This includes gesture recognition and person identification [15], respiration and breathing monitoring [16], [17], sleep and posture-aware sensing [18], and people counting in dense and crowded environments [19].

These advances have been enabled in part by the adoption of deep learning architectures for radar signal interpretation, point cloud processing, and spatiotemporal modeling. Techniques ranging from end-to-end neural processing pipelines to representation learning [20] and sequence modeling [21], [22] have significantly improved robustness and generalization [23]. However, most of these systems assume controlled sensing conditions, limited interference, and isolated tasks.

2.3. Single-Radar Fine-Grained Gesture Recognition

Single-radar gesture recognition has been widely investigated as a practical and low-cost solution for contactless human–computer interaction. A single mmWave radar can capture sparse point-cloud reflections generated by hand, arm, and body movements, allowing gesture recognition without cameras or wearable devices. Existing single-radar studies have explored both coarse and fine-grained gestures, including mid-air hand gestures, arm movements, and micro-motion patterns. These works typically rely on handcrafted signal features, range-Doppler representations, point-cloud features, or deep neural networks to classify temporal gesture patterns from one radar viewpoint.

Despite its simplicity, single-radar fine-grained gesture recognition remains challenging in multi-subject environments. Since the radar observes the scene from only one direction, the received reflections are highly dependent on the subject's position, orientation, and distance from the radar. Fine gestures may generate weak or sparse reflections, and their spatial-temporal patterns can vary significantly across different locations in the sensing area. In addition, when multiple subjects are present, their reflections may overlap or interfere with one another, making it difficult to isolate the gesture of a specific target.

Therefore, single-radar benchmarks are important for understanding the baseline capability of radar-based fine gesture recognition, as well as the limitations caused by viewpoint dependency, limited angular resolution, occlusion, and multi-subject interference.

2.4. Multi-Radar and Distributed Radar Sensing

Another line of research has focused on multi-radar systems and distributed sensing infrastructures [24]. Radar calibration, coordinate alignment, and spatial registration are critical prerequisites for multi-device sensing [25]. For example, [26] proposes a radar position calibration framework that aligns radar coordinates with room coordinates, enabling consistent spatial reasoning across multiple devices. Other works investigate sensor placement strategies and multi-view geometry for improved coverage and robustness [27].

These approaches address spatial alignment and coverage but do not fully address higher-level perception tasks in multi-user environments.

2.5. Research Gap and Position of This Work

Despite these advances, most radar-based sensing systems remain limited to single-task or single-subject settings, even in multi-radar deployments [28]. Realistic indoor environments with multiple users performing concurrent activities therefore require frameworks that jointly support multi-target tracking and fine-grained activity recognition.

Addressing this gap, this work combines two complementary directions: a single-radar fine-grained gesture recognition benchmark for evaluating gesture recognition under different subject placements, and a collaborative mmWave radar sensing framework that integrates multi-radar data fusion with joint multi-target tracking and gesture recognition in multi-user environments.

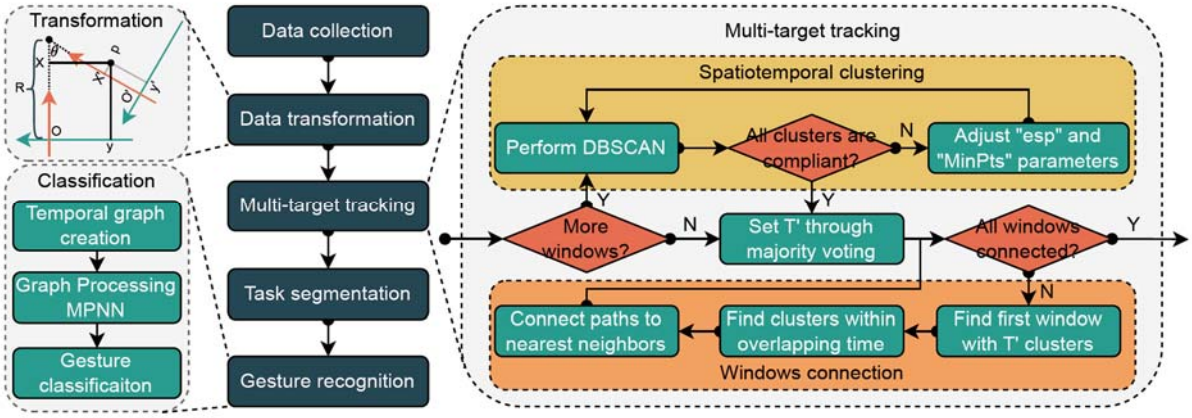


Fig. 2: System model of multi-target trajectory tracking. The individual radar coordinate systems are transformed into a canonical global representation across all radars. Classification is achieved via a graph-based scheme.

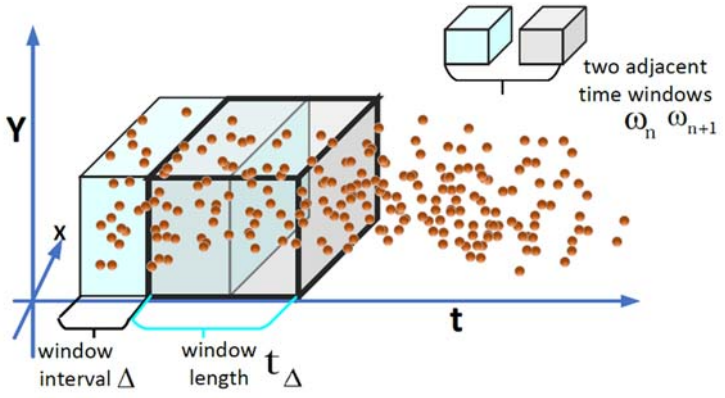


Fig. 3: Schematic diagram of a “Time Window”. It divides the point cloud into several segments with a certain length t_Δ and interval Δ.

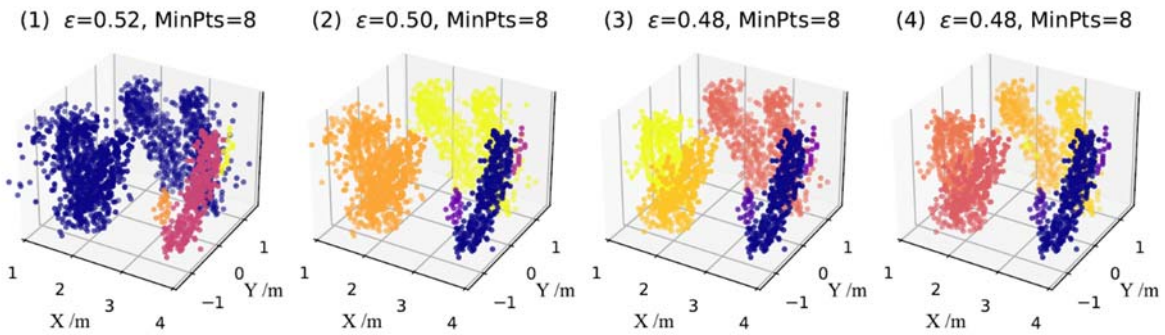


Fig. 4: Clustering parameter optimization. Parameters ϵ and MinPts are iteratively adjusted to split multi-target clusters into single-person trajectories.

3. Fine Gesture Recognition Methodology

3.1. Data collection setup

The single-radar benchmark is designed to test fine-grained gesture recognition under controlled multi-person spatial layouts. One radar is installed at the front of the scene and observes two, three, or four subjects. Each subject stands at a predefined position, and the benchmark defines fixed combinations so that spatial separation, angular separation, and gesture-class balance can be evaluated systematically.

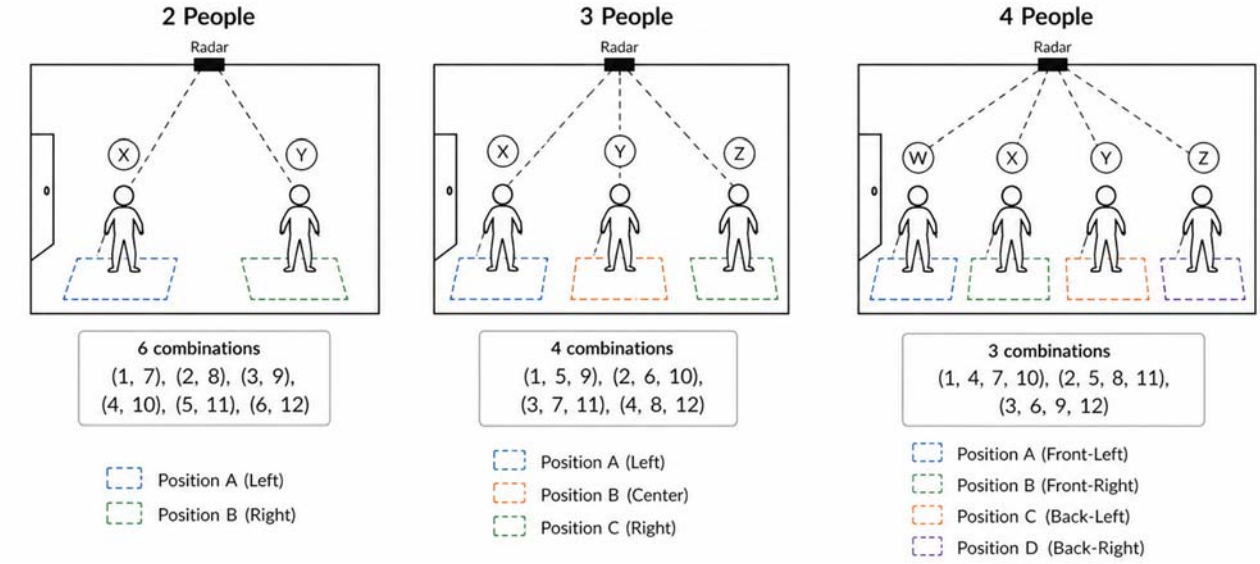


Fig. 5: Single-radar multi-subject fine-grained gesture benchmark protocol. The study defines two-, three-, and four-person layouts with controlled positions and predefined combinations. Results are pending.

In the two-person benchmark, two subjects are placed on the left and right sides of the sensing area, denoted as subject A and subject B. Six predefined position combinations are used: (1, 7), (2, 8), (3, 9), (4, 10), (5, 11), and (6, 12). These combinations are designed to evaluate gesture recognition performance when two participants are located at symmetric or corresponding positions across the sensing space.

In the three-person benchmark, three subjects are positioned on the left, center, and right sides of the sensing area, denoted as subjects A, B, and C. Four predefined combinations are used: (1, 5, 9), (2, 6, 10), (3, 7, 11), and (4, 8, 12). This setup evaluates whether the radar can distinguish fine gestures when multiple subjects are distributed across three spatial regions.

In the four-person benchmark, four subjects are arranged in front-left, front-right, back-left, and back-right positions, denoted as subjects A, B, C, and D. Three predefined combinations are used: (1, 4, 7, 10), (2, 5, 8, 11), and (3, 6, 9, 12). This scenario represents a denser multi-subject setting and is used to examine the robustness of single-radar gesture recognition under stronger spatial overlap and potential inter-subject interference.

3.2. End-to-end gesture recognition framework

The primary objective is not only to classify the gesture occurring in the scene, but to infer the number of active persons and assign a gesture label to each person-specific slot. To make the output representation independent of semantic position names such as left, center, or right, each person is indexed using an angle-ordered slot ID. Given a target location in the XOY plane, the polar angle with respect to the radar origin is computed as atan2(y, x), and all targets in the same scene are sorted in ascending angular order. The sorted order defines slot 0, slot 1, slot 2, and slot 3. The model therefore predicts existence and gesture labels for

a maximum of four angle-ordered persons, without regressing the continuous angle.

The proposed model, referred to as MP-SlotNet, uses a shared spatio-temporal encoder followed by two task-specific branches. The shared encoder consists of a PointNet-style frame encoder and a bidirectional GRU temporal encoder. For each frame, a point-wise multilayer perceptron maps the F-dimensional point features to a 256-dimensional latent representation, followed by max pooling over the point dimension. The resulting sequence of frame-level features is passed to a two-layer bidirectional GRU to model temporal gesture dynamics. The multi-person branch uses four learnable person queries in a Transformer decoder to extract slot-specific representations. Two prediction heads are attached to each slot representation: an existence head that predicts whether the slot is occupied, and a gesture head that predicts one of the 12 gesture classes. A separate single-person auxiliary head is applied to the temporally pooled shared representation and is used only for auxiliary gesture classification on single-person clips.

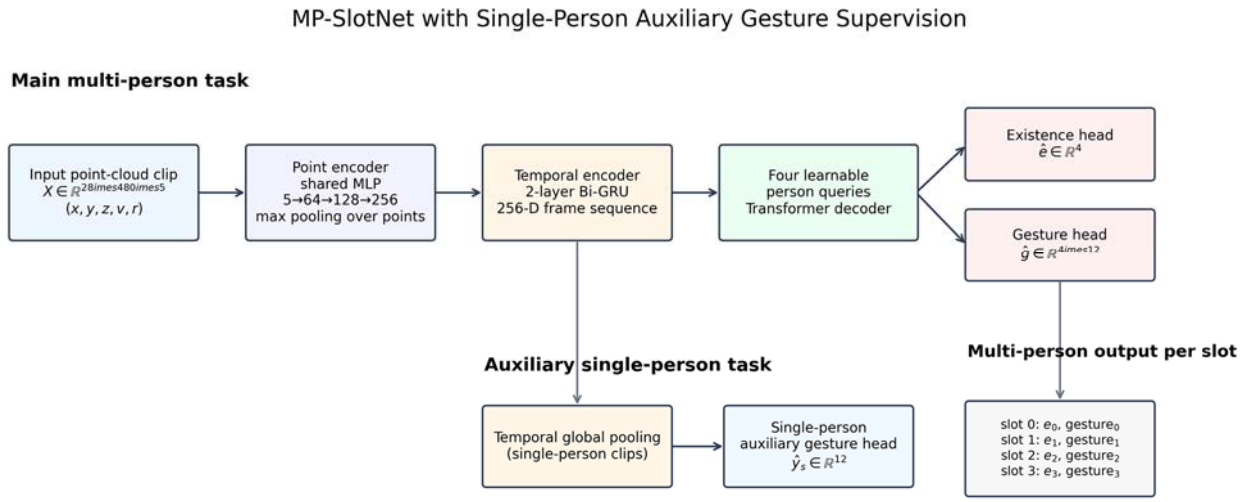


Fig. 6: Architecture of MP-SlotNet with a multi-person angle-slot branch and a single-person auxiliary branch.

4. Multi Radar System Overview

We consider multi-radar scenarios, with a deployment as depicted in Fig. 1(a). The principle of simultaneous tracking and gesture recognition is shown in Fig. 2. The operations of data collection, data transformation, multi-target tracking, extraction of gesture data, and gesture recognition are described in the following sections.

Trajectory identification for multiple targets is achieved through clustering and matching. A clustering algorithm is used to isolate gestures for individual targets, which are then fed into a pre-trained gesture recognition model, as illustrated in the "Multi-target tracking" module in Fig. 2.

4.1. Data Collection

We deploy eight TI AWR1443 mmWave radars operating in the 77 GHz to 81 GHz band with a 4 GHz bandwidth and a fixed frame rate of 30 fps. All radars are connected via USB to a single laptop for synchronized data collection, while subsequent processing treats each device independently. Point cloud data are acquired using the mmWave Data Collector [5].

To aggregate measurements from distributed radars, point clouds are transformed from local radar coordinate systems into a shared global reference frame. This requires spatial alignment and time synchronization across devices. Each local point $P' = (x', y')$ is mapped to the global coordinate system via a rotation by angle θ, determined by the relative orientation of the radar, followed by a translation. The transformation, illustrated in Fig. 2, is given by:

$$x = x'\cos\theta - y'\sin\theta + (1-\cos\theta)R, \tag{1}$$

$$y = x'\sin\theta + y'\cos\theta - \sin\theta \cdot R \tag{2}$$

4.2. Multi-Target Tracking

Fig. 2 illustrates the proposed multi-target tracking pipeline. Each point cloud is timestamped and synchronized across radars during data fusion. We introduce a sliding time window, shown in Fig. 3, to temporally align measurements and segment the continuous point cloud into overlapping spatiotemporal subsets, enabling iterative clustering and trajectory extraction.

All points within a time window are clustered to separate multiple moving targets. While Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is commonly used for spatial clustering, we extend it to a spatiotemporal domain by incorporating time as an additional dimension. This exploits temporal continuity and avoids merging spatially close but temporally disconnected points. Specifically, for each window w_n, the point set is defined as:

$$D = \left\{ (x_i, y_i, z_i) \in \mathbb{R}^3 \,\middle|\, i = 1, ..., m \right\}, \tag{3}$$

Clustering is performed using a combined spatial-temporal distance metric:

$$d = \left\| (x_1, y_1, t_1) - (x_2, y_2, t_2) \right\|. \tag{4}$$

MmWave point clouds exhibit non-uniform density due to distance-dependent resolution and signal attenuation. To account for this, we propose a spatiotemporal adaptive clustering strategy in which DBSCAN parameters are iteratively adjusted within each time window. As illustrated in Fig. 4, parameters are initialized using the k-distance graph [29], [30] and progressively refined until individual trajectories emerge as distinct clusters.

Convergence is determined by evaluating whether the spatial extent of each cluster within short temporal segments, such as 0.1 s, matches the expected cross-sectional size of a human body. Finally, spurious

clusters caused by limb motion or partial occlusions are removed based on cluster duration and volume. Small clusters with volumes below one-third of the average cluster volume are discarded. The final number of trajectories is determined across all windows using majority voting, yielding robust multi-target tracking results.

In this way, trajectories and target counts are obtained for each time window. The trajectories are then merged across windows to reconstruct complete multi-target trajectories. The trajectory merging procedure is summarized in Algorithm 1.

Algorithm 1 Trajectory Merging Across Time Windows

Step	Operation	Description
Input	Sequence of time windows with extracted trajectories	The algorithm takes a sequence of time windows, where trajectories have already been extracted from each window.
Output	Merged multi-target trajectories	The algorithm outputs complete trajectories after merging trajectory segments across time windows.
1	Initialization	Select the first time window with (T) detected trajectories as the initialization.
2.1	Temporal overlap detection	For each pair of consecutive time windows, find the temporal overlap between the two windows.
2.2	Overlapping-point extraction	Extract trajectory points in the overlapping interval.
2.3	Distance computation	Compute pairwise trajectory distances between trajectory segments in adjacent windows.
2.4	Trajectory matching	Match trajectories according to the minimum distance.
2.5	Trajectory merging	Merge the matched trajectories across consecutive windows.
3	Iteration	Repeat the process until trajectories from all windows are merged.

4.3. Gesture Recognition

Gesture extraction is based on the observation that users typically pause to perform gestures, which results in reduced point cloud density compared to walking, as motion is largely confined to the limbs. Low-density intervals in the time domain are therefore used to segment gesture phases from trajectories.

In multi-user scenarios, gesture data are further separated across targets using K-means clustering, with the number of clusters determined by the preceding multi-target tracking results. After extracting the gesture time segment, gestures are classified using a graph convolutional neural network with message passing, inspired by [4], [5].

Each gesture point cloud is defined as: $X = \{x_i\}_{i=1}^{n} \subset \mathbb{R}^F$, where each point includes spatial features and a frame index f_i^s encoding temporal order. To model temporal dynamics, we construct an intermediate graph by connecting each point only to its nearest neighbors in the subsequent frame:

$$F_{xi} = \left\{ x_j \in X \,\middle|\, f_j^s = f_i^s + 1 \right\}, \tag{5}$$

Euclidean distance is computed using features excluding time. Feature updates are then performed via message passing:

$$h_i^l = \Gamma_{j:(i,j)\in\varepsilon} M_\theta \left(h_i^{l-1}, h_j^{l-1} \right), \tag{6}$$

where Γ is a channel-wise max operator and M_θ is a learnable function capturing local spatiotemporal structure.

For gesture recognition, only the three radars facing each target are used. Gestures are first classified independently for each radar to maintain robustness to missing views, and the final label is obtained by aggregating radar-specific representations:

$$P = \rho\left(\sum_{i=1}^{m} \phi_i\left(R_i\right)\right), \tag{7}$$

where R_i denotes the representation from radar i, ϕ_i is a learnable projection, ρ normalizes the output, and m is the number of available radars.

5. Results of single radar evaluation

This section evaluates the performance of single-radar fine-grained gesture recognition in multi-subject scenarios. The goal is to investigate whether a single mmWave radar can recognize gestures from multiple people located at different spatial positions, and to analyze how recognition performance changes as the number of participants increases.

The evaluation focuses on two aspects. First, we describe the collected dataset, including the gesture classes, participant layout, spatial configurations, and data collection procedure. Second, we report the multi-person gesture recognition results under two-person, three-person, and four-person benchmark settings.

5.1. Dataset

The single-radar dataset was collected using TI's MMWCAS radar in a controlled indoor environment. Participants were asked to stand at predefined spatial positions and perform a set of fine-grained gestures. The radar captured sparse point-cloud reflections generated by hand and arm movements, which were then used for gesture recognition.

The gesture set contains 12 predefined gesture classes. These gestures are designed to cover different arm and hand motion patterns, including directional movements, circular movements, and other fine-grained dynamic gestures. A visual illustration of the 12 gestures is shown in Fig. 7.

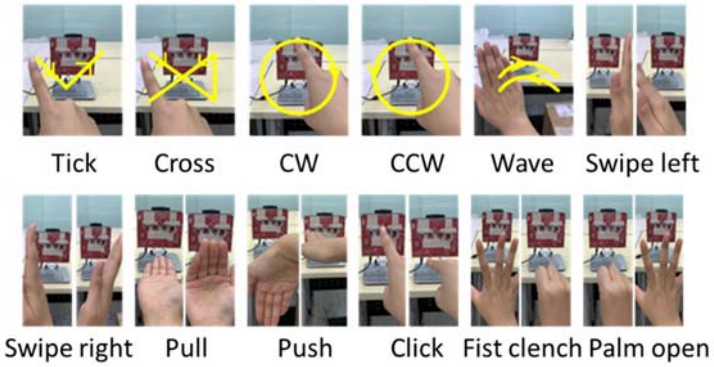


Fig. 7: Illustration of the 12 gesture classes used in the single-radar benchmark.

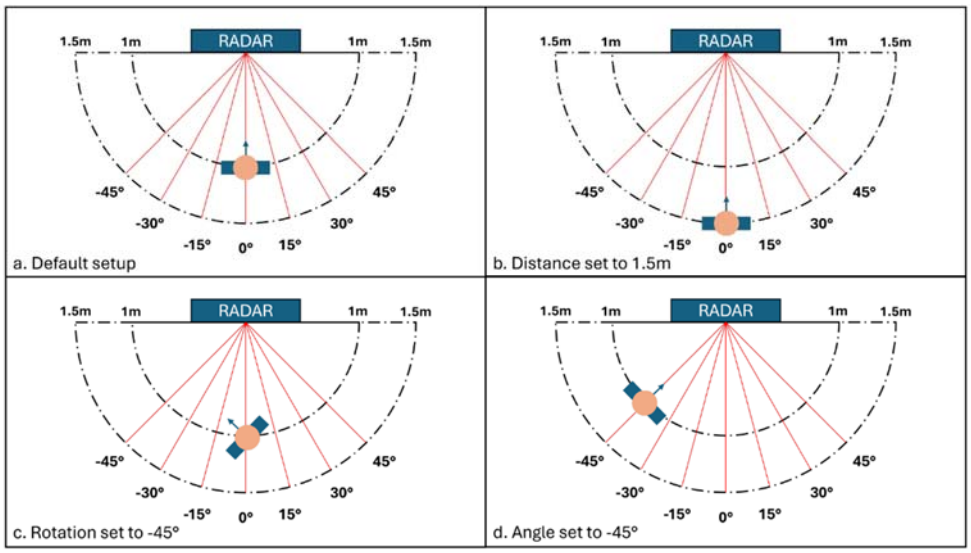


Fig. 8: Illustration of settings for single subject gesture data collection

A large-scale single-person gesture dataset was collected to learn robust gesture representations. The dataset includes 12 primary gestures recorded under a base setting of 1 m distance, 0° observation angle, and 0° body rotation, with 600 samples per gesture. Additional data were collected to improve robustness across different sensing conditions, including a 1.5 m distance, observation angles of ±15°, ±30°, and ±45°, body rotations of ±15°, ±30°, and ±45°, and new participants, with 50 samples per gesture/setting.

Each input sample is represented as a temporal point-cloud clip X ∈ R^{T×N×F}. In the reported experiment, T = 28 frames, N = 480 points per frame after sampling, and F = 5 point-level features. The five features correspond to spatial coordinates and radar-derived attributes, e.g., x, y, z, Doppler velocity, and range. The multi-person dataset contains two-, three-, and four-person samples. The single-person dataset is used as an auxiliary source of gesture supervision to improve the discriminative quality of the shared spatio-temporal encoder.

Table 1. Dataset splits and label formats used in the experiment.

Dataset subset	Split	Number of samples	Label format
Single-person auxiliary set	Train	10,896 clips	gesture label y ∈ {0,…,11}
Single-person auxiliary set	Validation	2,336 clips	gesture label y ∈ {0,…,11}
Multi-person main set	Train	437 clips	exist ∈ {0,1}^4; gesture_slots ∈ {0,…,11}^4
Multi-person main set	Validation	100 clips	exist ∈ {0,1}^4; gesture_slots ∈ {0,…,11}^4
Multi-person main set	Test	102 clips	exist ∈ {0,1}^4; gesture_slots ∈ {0,…,11}^4

5.2. Angle-ordered slot labels

For a sample containing M persons, where 1 ≤ M ≤ 4, the label is converted into a fixed four-slot representation. The binary existence vector e ∈ {0,1}^4 indicates whether each angle-ordered slot is occupied. The gesture vector g ∈ {0,…,11}^4 stores the corresponding gesture class for each valid slot. Empty slots are ignored during gesture-loss computation. For example, a two-person sample with ordered gestures [6, 1] is encoded as e = [1, 1, 0, 0], while the first two entries of the gesture-slot vector contain the two gesture labels after zero-based indexing.

Fig. 9: Angle-ordered slot labeling. Slot IDs are assigned by sorting persons according to atan2(y, x) around

the radar origin.

5.3. Experimental setup

We train the model using Python, Pytorch on the GPU of Nvidia A40 48GB. Training is formulated as a multi-task optimization problem. The multi-person main task supervises slot existence and slot-wise gesture classification, while the single-person auxiliary task encourages the shared encoder to learn robust gesture representations from a larger set of single-person samples.

For the multi-person branch, binary cross-entropy is used for the existence logits, and cross-entropy is used for gesture classification on valid slots only. The multi-person loss is defined as $L_{multi} = \lambda_{exist} L_{exist} + \lambda_{gesture} L_{gesture}$. For single-person auxiliary supervision, the loss L_{single} is a standard cross-entropy loss over the 12 gesture classes. The final joint objective is $L = L_{multi} + \alpha L_{single}$, with $\alpha = 0.3$ in the reported experiment.

The training procedure consists of three stages. First, the shared encoder and the single-person gesture head are pretrained for 20 epochs using only single-person samples. Second, the multi-person branch is trained while freezing the shared encoder for the first five multi-person epochs. Third, all modules are jointly fine-tuned using the multi-person main loss and the single-person auxiliary loss. The model is optimized with AdamW using a learning rate of 0.001 and weight decay of 0.0001. The batch size is 8, the number of training epochs is 200, and the decision threshold for slot existence is 0.5.

5.4. Multi people gesture recognition results

Three metrics were used to evaluate multi-person recognition. **People-count accuracy** measures whether the number of occupied slots is correctly predicted. **Slot-level gesture accuracy** evaluates gesture classification over all valid person slots. **Scene-level accuracy** is the strictest metric, where a scene is considered correct only if both the number of persons and all corresponding slot-level gestures are correctly predicted.

The training process shows stable convergence, as shown in Fig. 10, 11, 12 and 13. During single-person pretraining, the loss decreased from **1.1200** at the first epoch to **0.0933** at the 20th epoch, while the validation accuracy reached a peak of **96.79%** at epoch 11 and remained at **96.15%** at the end of pretraining. This indicates that the shared encoder learned discriminative single-person gesture representations. In the subsequent multi-person training stage, the multi-person loss dropped rapidly and the validation scene-level accuracy increased from **26.00%** at epoch 1 to **81.00%** at epoch 2, **94.00%** at epoch 10, and **97.00%** at epoch 11. The scene-level accuracy reached **99.00%** at epoch 27 and then remained close to saturation. Meanwhile, the single-person auxiliary accuracy during joint training fluctuated between approximately **70% and 83%**, suggesting that the shared representation was mainly optimized toward the primary multi-person task.

The best validation results were **100.00%** for people-count accuracy, **99.28%** for slot-level gesture accuracy, and **99.00%** for scene-level accuracy. On the final test set, the model achieved **100.00%** people-count accuracy, **99.65%** slot-level gesture accuracy, and **99.02%** scene-level accuracy, demonstrating that the angle-ordered slot formulation can reliably associate gestures with different persons in multi-person radar scenes.

The confusion matrix in Fig. 14 was constructed by collecting all valid multi-person slots from the test set and comparing each slot's ground-truth gesture label with its predicted label. Rows represent true classes, columns represent predicted classes, diagonal entries indicate correctly classified samples, and off-diagonal entries indicate misclassifications. Only one off-diagonal error was observed: one sample from class 7 was misclassified as class 1. Therefore, the slot-level gesture accuracy is **281/282 = 99.65%**, indicating highly accurate person-wise gesture recognition with minimal inter-class confusion.

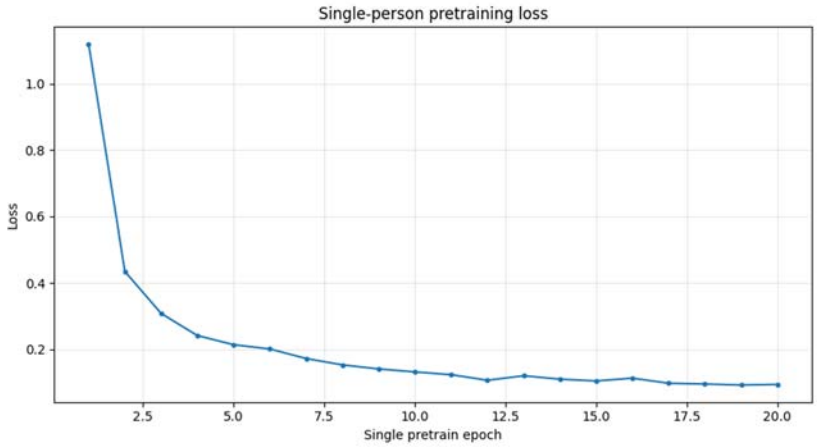


Fig. 10: Single person pretraining loss process

Fig. 11: Single person pretraining accuracy process

Fig. 12: Multi person training loss process

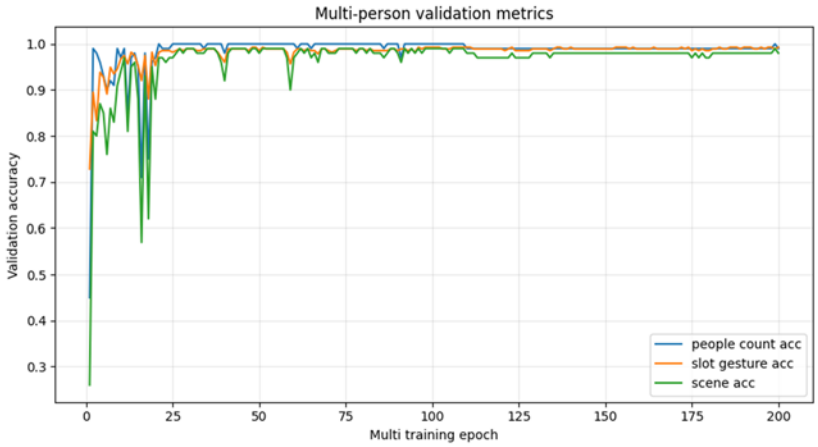


Fig. 13: Multi person training accuracy process

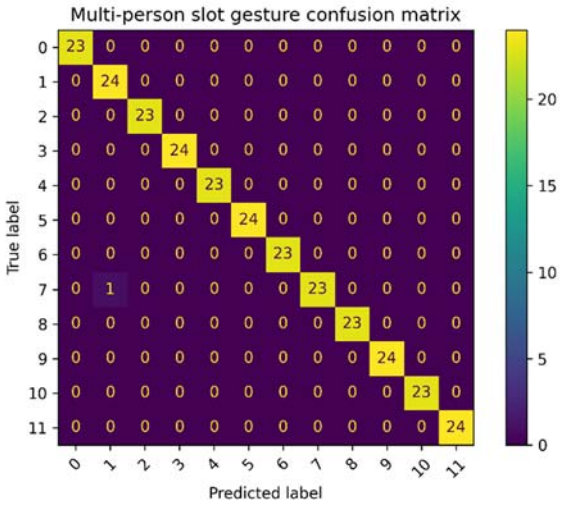


Fig. 14: Confusion matrix of multi person slot gesture recognition

6.Results of multi radar evaluation

In this section, the overall performance of the 8-radar system is evaluated under different settings for tracking, counting, and gesture recognition.

6.1.Experimental Setup

The dimensions of the experimental scenes are illustrated in Fig. 1(a). Eight radars are arranged in a circular area with a diameter of 2.5 m and mounted on tripod stands approximately 1.2 m above the ground. At the center of this area, a circular path with a radius of 1.3 m is designated as the predefined walking route for the targets. Consequently, the minimum distance between each radar and the target is approximately 1.2 m. The environment is equipped with eight TI AWR1443 radars.

We first evaluate a multi-person walking scenario to assess trajectory tracking and target counting accuracy. This is followed by a second scenario that combines tracking and gesture recognition. For the walking-only evaluation, five volunteers participated, with randomized starting locations across trials. Experiments were conducted with four and five targets in Scene 1. In the four-target case, participants started at positions 1 to 4, while in the five-target case, participants started at positions 1 to 5. All participants walked eight synchronized steps in the same direction.

Movement was synchronized using a metronome set to 0.8 s, and ground markings ensured a step length of 0.7 m for repeatable conditions.

For simultaneous trajectory tracking and gesture recognition, the five targets started at positions 1 to 5 in Scene 2. Mobility was again synchronized by a metronome in the same way as described above. Subjects remained stationary for approximately 3 s before performing the designated gestures. This experiment was repeated 10 times, resulting in a total of 500 gesture samples collected for recognition.

6.2. Radar-Count Ablation for Multi-Target Tracking

We treat the radar-density study as an ablation over sensing coverage. Using the configurations in Fig. 15, we evaluate radar subsets ranging from 2 to 8 radars under otherwise identical four-target and five-target trials. Table 2 reports the resulting tracking error and counting accuracy. This ablation isolates how performance changes when viewpoints and redundant coverage are removed from the full eight-radar deployment.

Fig. 16(a) and Fig. 16(b) show that adding more radars consistently improves the positioning-error distribution, with the largest performance gain observed when increasing the number of radars from two to three. For four targets, the probability of achieving tracking errors below 0.2 m increases from 30% to 83% as the radar count grows from 2 to 8. For five targets, this probability increases from 18% to 76%.

The eight-radar configuration achieves the best average tracking error of 0.19 m for both target counts. Compared with the two-radar configuration, the eight-radar setup improves tracking accuracy by 38.7% and 51.3%, and improves counting accuracy by 22% and 29% in the four-target and five-target cases, respectively. The five-target setting remains more challenging, confirming that increased radar density can partially, but not fully, compensate for crowding and occlusion.

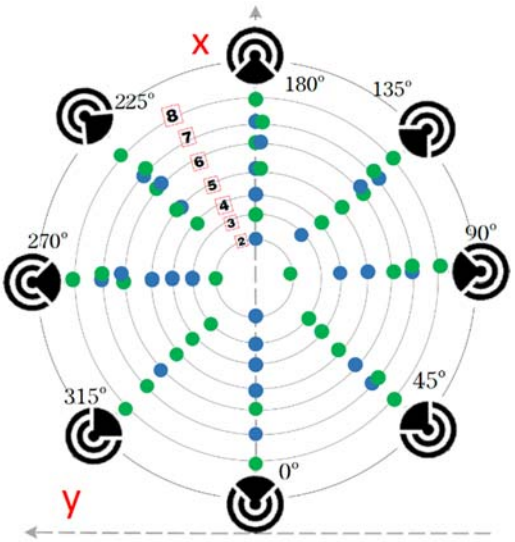
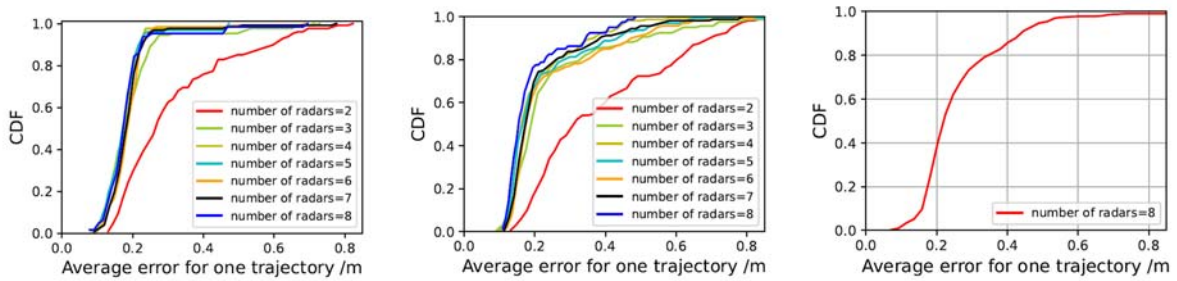


Fig. 15: Radar configurations with varying numbers of radars. Red boxes indicate the selected radar count, while blue and green denote different radar groups. For example, in the three-radar case, configurations [0◦, 270◦, 135◦] and [0◦, 225◦, 90◦] are used in separate experiments.



(a) Tracking errors with 4 targets in Scene 1. (b) Tracking errors with 5 targets in Scene 1. (c) Tracking errors with 5 targets in the Scene 2.

Fig. 16: Tracking errors in scene 1 and scene 2.

Table 2: Radar-count ablation results

Number of Radars	Four Targets: Tracking Error	Four Targets: Counting Accuracy	Five Targets: Tracking Error	Five Targets: Counting Accuracy
2	0.31 m	73%	0.39 m	62%
3	0.20 m	92%	0.25 m	82%
4	0.18 m	94%	0.21 m	91%
5	0.18 m	94%	0.22 m	91%
6	0.19 m	94%	0.24 m	88%
7	0.19 m	95%	0.22 m	91%
8	0.19 m	95%	0.19 m	91%

6.3. Joint Tracking and Gesture Recognition

In Scene 2, five targets walk simultaneously and then stop to perform gestures. The trajectory tracking cumulative distribution function (CDF) is shown in Fig. 7(c). With eight radars, the probability of achieving a tracking error below 0.2 m is 38%, which is substantially lower than the 76% obtained in Scene 1 under the same radar configuration and target count.

The average tracking error in Scene 2 is 0.26 m, while correct target counts are obtained in 94% of time windows. Since the only difference between the two scenarios is the presence of target occlusion and different spatial distribution, these results demonstrate that occlusions and target spatial distribution significantly degrade tracking accuracy, even with identical radar deployments.

6.4. Gesture Recognition Accuracy

As shown in Fig. 17(a), using all three radars facing each target yields a gesture recognition accuracy of 95.85% over ten gestures. Reducing the number of contributing radars generally degrades performance. An exception is observed when using only the radar directly in front of the user, which outperforms the two-radar configuration located at 45° and 315°.

Fig. 17(b) reports recognition accuracy as a function of radar count. For the one-radar and two-radar cases, radars are randomly removed, and each configuration is evaluated over ten trials. The mean accuracy and standard deviation are reported.

The confusion matrices show misclassifications mainly between gestures e–f, corresponding to vertical circle clockwise and counterclockwise gestures, and h–i, corresponding to push and pull gestures with both hands. The former confusion is caused by the limited elevation resolution of the radar [3], while the latter is due to smaller radar cross sections and similar spatiotemporal patterns that differ mainly in temporal dynamics.

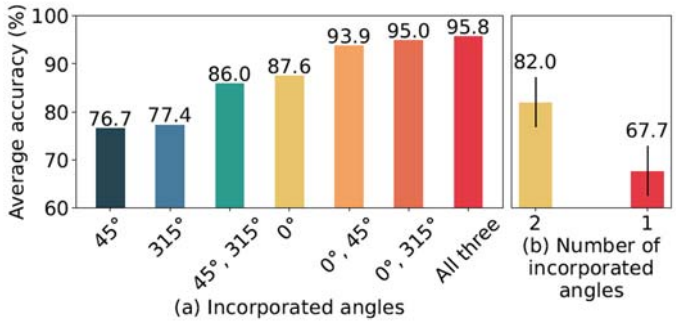


Fig. 17: (a) Gesture recognition accuracy versus the number of radars relative to each target. (b) Gesture recognition accuracy with randomly removed radars; each configuration is evaluated over 10 trials.

7. Discussion

The two outcomes address different stages of the same research problem. The single-radar benchmark is useful for isolating the fine-grained recognition problem. Because the radar viewpoint is fixed, it helps identify which gestures are intrinsically difficult to separate and how subject placement affects the point-cloud signature. This is valuable before adding the complexity of distributed synchronization and multi-radar fusion.

The multi-radar prototype addresses scalability. By observing the environment from multiple angles, it reduces the dependence on a single viewpoint and enables joint tracking and gesture recognition. However, the system still depends on accurate radar calibration, temporal synchronization, and reliable clustering. Crossing trajectories, dynamic entry and exit of subjects, and very close subject spacing remain important deployment challenges.

A practical next step is to use the single-radar benchmark to identify difficult fine-grained gesture pairs, then test whether multi-radar fusion improves those same pairs. This would create a direct experimental link between the two outcomes and strengthen the benchmark as a project deliverable.

8. Conclusion

This deliverable integrates two project achievements into a coherent benchmark-and-prototype framework. The first achievement is a single-radar fine-grained gesture-recognition benchmark with defined two-, three-, and four-person layouts and controlled gesture combinations. The second achievement is a distributed multi-radar arm-gesture recognition prototype that fuses point clouds, tracks multiple subjects, segments gesture intervals, and recognizes gestures using graph-based spatiotemporal learning.

The document deliberately presents the single-radar study first, followed by the multi-radar study. This ordering reflects the development path from controlled fine-grained recognition to robust multi-subject operation. Results for the single-radar benchmark remain open, while the multi-radar prototype already demonstrates strong performance in multi-target tracking and gesture recognition.

9.References

[1] M. Nagenborg, S. Cammers-Goodwin, Alexander Paulus, and Thomas Eibert, “Privacy and Ethical Constraints in Static Environments,” European Innovation Council, EU, 3.4, Nov. 2024.

[2] Sage Cammers-Goodwin, Michael Nagenborg, and Stefano Savazzi, “Anticipated social implications for RF Holography in dynamic environments following privacy by design approach,” European Innovation Council, EU, 4.4, May 2025.

[3] Sage Cammers-Goodwin *et al.*, “Ethics Status Monitor (ESM) Version 3,” European Innovation Council, EU, 8.11, July 2025.

[4] “Radar Sensor for Smart TVs,” Texas Instruments. Accessed: Nov. 18, 2025. [Online]. Available: https://www.ti.com/video/6325071464112

[5] Ciano Aydin, Sage Cammers-Goodwin, Emanuele Ferro, Moniruzzaman Kiron, Daniele Piazza, and Luca Possati, “Functional Requirements, Privacy Profiles for the Scenarios,” European Innovation Council, EU, 6.1, Nov. 2024.

[6] A. C. Schembri and C. Lucas, Eds., *Sociolinguistics and deaf communities*, First published. Cambridge: Cambridge University Press, 2015. doi: 10.1017/CBO9781107280298.

[7] M. Nagenborg, “Surveillance and persuasion,” *Ethics Inf. Technol.*, vol. 16, no. 1, pp. 43–49, Mar. 2014, doi: 10.1007/s10676-014-9339-4.

[8] S. Zuboff, “Surveillance Capitalism or Democracy? The Death Match of Institutional Orders and the Politics of Knowledge in Our Information Civilization,” *Organ. Theory*, vol. 3, no. 3, p. 263178772211292, July 2022, doi: 10.1177/26317877221129290.

[9] P.-P. Verbeek and D. Tijink, *Guidance Ethics Approach: An ethical dialogue about technology with perspective on actions*. ECP | Platform voor de InformatieSamenleving, 2020. [Online]. Available: https://ecp.nl/wp-content/uploads/2020/11/Guidance-ethics-approach.pdf

[10] T. Swierstra, D. Stemerding, and M. Boenink, “Exploring Techno-Moral Change: The Case of the ObesityPill,” in *Evaluating New Technologies: Methodological Problems for the Ethical Assessment of Technology Developments.*, vol. 3, P. Sollie and M. Düwell, Eds., in The International Library of Ethics, Law and Technology, vol. 3. , Dordrecht: Springer Netherlands, 2009. doi: 10.1007/978-90-481-2229-5.

[11] R. S. Lewis, “Technological Gaze,” in *Perception and the inhuman gaze: perspectives from philosophy, phenomenology, and the sciences*, 1st ed., A. Daly, F. Cummins, J. Jardine, and D. Moran, Eds., in Routledge studies in contemporary philosophy. , New York, NY: Routledge, 2020.

[12] C. Aydin, M. González Woge, and P.-P. Verbeek, “Technological Environmentality: Conceptualizing Technology as a Mediating Milieu,” *Philos. Technol.*, vol. 32, no. 2, pp. 321–338, June 2019, doi: 10.1007/s13347-018-0309-3.

[13] P.-P. Verbeek, “Toward a Theory of Technological Mediation: A Program for Postphenomenological Research,” in *Technoscience and postphenomenology: the Manhattan papers*, J. K. B. O. Friis and R. P. Crease, Eds., in Postphenomenology and the philosophy of technology. , London: Lexington Books, 2015, pp. 189–204.

[14] B. Friedman and D. G. Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press, 2019. doi: 10.7551/mitpress/7585.001.0001.

[15] I. Van De Poel, “Translating Values into Design Requirements,” in *Philosophy and Engineering: Reflections on Practice, Principles and Process*, vol. 15, D. P. Michelfelder, N. McCarthy, and D. E. Goldberg, Eds., in Philosophy of Engineering and Technology, vol. 15. , Dordrecht: Springer Netherlands, 2013, pp. 253–266. doi: 10.1007/978-94-007-7762-0_20.

[16] D. Ihde, *Technology and the lifeworld: from garden to earth*. in The Indiana series in the philosophy of technology. Bloomington: Indiana University Press, 1990.

[17] P.-P. Verbeek, “Beyond interaction: a short introduction to mediation theory,” *Interactions*, vol. 22, no. 3, pp. 26–31, Apr. 2015, doi: 10.1145/2751314.

[18] K. Bauer and J. Hermann, “Technomoral Resilience as a Goal of Moral Education,” *Ethical Theory Moral Pract.*, vol. 27, no. 1, pp. 57–72, Mar. 2024, doi: 10.1007/s10677-022-10353-1.

[19] European Parliament and Council, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139*

and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. 2024, pp. 1–144. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[20] Susan Leigh Star, “Power, Technology and the Phenomenology of Conventions: On being Allergic to Onions,” *Sociol. Rev.*, vol. 38, no. 1 (Suppl), pp. 26–56, 1990.

[1] Liu, An, et al. "A survey on fundamental limits of integrated sensing and communication." *IEEE Communications Surveys & Tutorials* 24.2 (2022): 994-1034.

[2] Liu, Fan, et al. "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond." *IEEE journal on selected areas in communications* 40.6 (2022): 1728-1767.

[3] Palipana, Sameera, et al. "Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds." *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5.1 (2021): 1-27.

[4] Salami, Dariush, et al. "Tesla-rapture: A lightweight gesture recognition system from mmwave radar sparse point clouds." *IEEE Transactions on Mobile Computing* 22.8 (2022): 4946-4960.

[5] Salami, Dariush, et al. "Angle-Agnostic Radio Frequency Sensing Integrated Into 5G-NR." *IEEE Sensors Journal* 24.21 (2024): 36099-36114.

[6] Chityat, Inbar, et al. "Multimodal Non-Contact Sensing of Respiration and Movement in Neonates." *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2025.

[7] Jin, Can, et al. "SpDiff: A Speech Sensing System with Diffusion Model Based on mm Wave Radar." *2025 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2025.

[8] Huang, Xu, et al. "Indoor detection and tracking of people using mmwave sensor." *Journal of Sensors* 2021.1 (2021): 6657709.

[9] Joshi, Marvin, et al. "A fully passive machine learning enabled lens-based mmID system for enhanced orientation detection in 5G/mmWave and IoT applications." *2024 54th European Microwave Conference (EuMC)*. IEEE, 2024.

[10] Liu, Yunhao, et al. "Real-time continuous activity recognition with a commercial mmWave radar." *IEEE Transactions on Mobile Computing* 24.3 (2024): 1684-1698.

[11] Zhang, Jia, et al. "A survey of mmWave-based human sensing: Technology, platforms and applications." *IEEE Communications Surveys & Tutorials* 25.4 (2023): 2052-2087.

[12] Canil, Marco, Jacopo Pegoraro, and Michele Rossi. "MilliTRACE-IR: Contact tracing and temperature screening via mmWave and infrared sensing." *IEEE Journal of Selected Topics in Signal Processing* 16.2 (2021): 208-223.

[13] Kang, Wei, Chenwei Zhou, and Wen Wu. "Respiration monitoring of all occupants in a vehicle using time-division multiplexing FMCW radar based on metasurface technology." *IEEE Transactions on Microwave Theory and Techniques* 72.8 (2024): 4960-4974.

[14] Zhang, Xusheng, et al. "Waffle: A waterproof mmwave-based human sensing system inside bathrooms with running water." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7.4 (2024): 1-29.

[15] Yang, Hongliu, et al. "From Spatial Domain to Temporal Domain: Unleashing the Capability of CFAR for mmWave Point Cloud Generation." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9.2 (2025): 1-29.

[16] Zhang, Duo, et al. "From single-point to multi-point reflection modeling: Robust vital signs monitoring via mmwave sensing." *IEEE Transactions on Mobile Computing* 23.12 (2024): 14959-14974.

[17] Chang, Zhaoxin, et al. "Mmecare: Enabling fine-grained vital sign monitoring for emergency care with handheld mmwave radars." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8.4 (2024): 1-24.

[18] Adhikari, Aakriti, and Sanjib Sur. "MiSleep: Human sleep posture identification from deep learning augmented millimeter-wave wireless systems." *ACM Transactions on Internet of Things* 5.2 (2024): 1-33.

[19] Tang, Junkui, et al. "Multichannel radar forward-looking super-resolution imaging method based on structured sparsity." *IEEE Transactions on Geoscience and Remote Sensing* 63 (2025): 1-14.

[20] Rai, Prashant Kumar, Nataliya Strokina, and Reza Ghabcheloo. "Representation learning for place recognition using MIMO radar." *IEEE Open Journal of Intelligent Transportation Systems* (2025).

[21] Gao, Hong, et al. "Radar-Mamba: 4D Millimeter-Wave Point Cloud Enhancement via State Space Models." *Proceedings of the 33rd ACM International Conference on Multimedia*. 2025.

[22] Chen, Yingru, et al. "STPM: Spatial-Temporal Point Mamba for Activity Recognition Using mmWave Radar Point Clouds." *2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025.

[23] Salami, Dariush, et al. "A joint radar and communication approach for 5G NR using reinforcement learning." *IEEE Communications Magazine* 61.5 (2023): 106-112.

[24] He, Guangqiang, et al. "DRMTrack: An Extended Distributed Millimeter-Wave Radar Framework for Indoor Multi-Target Human Trajectory Tracking." *IEEE Internet of Things Journal* (2025).

[25] Li, Yi, et al. "High resolution DOA estimation of coherent distributed millimeter-wave radar." *IEEE Transactions on Aerospace and Electronic Systems* 61.3 (2025): 6098-6109.

[26] Zhang, Duo, et al. "Local: An automatic location attribute calibration approach for large-scale deployment of mmwave-based sensing systems." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7.4 (2024): 1-27.

[27] Wang, Xuanzhi, et al. "Placement matters: Understanding the effects of device placement for WiFi sensing." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.1 (2022): 1-25.

[28] Khan, Irshad, and Young-Woo Kwon. "Radar-Based Hand Gesture Recognition With Feature Fusion Using Robust CNN-LSTM and Attention Architecture." *IEEE Access* (2025).

[29] Sander, Jörg, et al. "Density-based clustering in spatial databases: The algorithm gdbscan and its applications." *Data mining and knowledge discovery* 2.2 (1998): 169-194.

[30] Rahmah, Nadia, and Imas Sukaesih Sitanggang. "Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra." *IOP conference series: earth and environmental science*. Vol. 31. No. 1. IoP Publishing, 2016.