Holden Deliverable

D5.4 – Direction-agnostic RF sensing from point cloud data

Grant Agreement number	101099491
Action Acronym	HOLDEN
Action Title	Ehtical Design of Holography with Dense wireless Networks
Type of action	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 delierable	30/11/2025
Actual date of submission	28/11/2025
Lead beneficiary for the deliverable	AALTO
Dissemination level of the deliverable	Public

Action coordinator's scientific representative

Prof. Stephan Sigg
AALTO – KORKEAKOULUSÄÄTIÖ,
Aalto Unviersity School of Electrical Engineering, Department of Information and Communications Engineering
stephan.sigg@aalto.fi

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.

European Innovation Council

Authors in alphabetical order		
Name	**Beneficiary**	**e-mail**
Dariush Salami	AALTO	dariush.salami@aalto.fi
Yuqing Song	AALTO	yuqing.song@aalto.fi

Change history				
Version	**Date**	**Status**	**Partner**	**Description**
1.0	24.08.2025	Final	Aalto	First final draft

Abstract

This deliverable (D5.4: Direction-agnostic Radio Frequency (RF) sensing from point cloud data) addresses one of the major bottlenecks in RF based human perception: severe performance degradation when the relative orientation between the sensor and the user deviates from the angles seen during training. Using sparse 3D point clouds derived from mmWave Frequency Modulated Continuous Wave (FMCW) radars, we demonstrate that robust gesture and activity recognition across previously unseen observation directions can be achieved with minimal angular supervision.

We propose and systematically evaluate a multi-faceted approach combining (i) aggressive on-the-fly geometric augmentations, (ii) viewpoint-conditioned neural architectures that explicitly fuse acquisition angle embeddings with point cloud features, (iii) diffusion-based synthetic generation of missing angular viewpoints, and (iv) replay-based continual learning strategies to incrementally accommodate new orientations without catastrophic forgetting. Experiments on two large multi-angle datasets show that Experience Replay (ER) and A-GEM significantly outperform regularization-only methods (EWC, SI, LwF), reaching final test accuracies of $85.0\,\%$ and $81.9\,\%$ respectively when trained on only a subset of directions, compared with $59.7\,\%$ for naive fine-tuning.

The explicit modeling of sensor viewpoint, rather than strict rotation invariance, emerges as particularly effective for sparse, torso-dominated mmWave point clouds. By deliberately discarding orientation-specific details in favour of motion invariants, the resulting representations are not only highly generalizable but also inherently privacy-preserving, as personally identifiable spatial reconstructions become unfeasible.

These findings provide a practical, data-efficient foundation for scalable RF sensing layers in dense wireless networks. They directly support the HOLDEN vision of ubiquitous, ethically compliant holographic perception that operates without cameras and without requiring exhaustive multi-angle data collection from every user.

Contents

1 Introduction

The HOLDEN project (Ethical Design of Holography with Dense Wireless Networks) envisions a future where Radio Frequency (RF) signals, emitted ubiquitously by communication infrastructures such as 5G and 6G networks, are used not only for data transmission but also for environmental perception and human-centered applications. HOLDEN aims to establish the theoretical, algorithmic, and ethical foundations for holographic perception systems that can operate on top of dense wireless networks. Through the convergence of radio propagation modeling, sensing, and machine learning, the project investigates how electromagnetic fields can be repurposed to form an ethically compliant sensing layer for the physical world, one that enables ubiquitous perception without relying on conventional imaging modalities such as cameras.

The motivation behind this vision arises from the increasing demand for context-aware technologies that can interact naturally with humans and their surroundings while maintaining high standards of privacy. RF-based sensing, particularly using Millimeter-Wave (mmWave) radars, provides a unique opportunity in this regard: it offers rich information about motion, distance, and material properties while inherently respecting visual anonymity. However, achieving robust and generalizable sensing performance in realistic environments requires overcoming fundamental technical challenges associated with signal directionality, multipath propagation, and data diversity.

Among these challenges, direction dependence in RF sensing has proven to be one of the most persistent bottlenecks. When mmWave radars are used for gesture recognition or human activity classification, the measured signal characteristics, including Doppler shifts, range profiles, and micro-motion signatures, vary significantly with the orientation of the target relative to the radar aperture. Consequently, models trained on data captured from a specific viewpoint often fail to generalize to other observation angles. This limitation severely restricts scalability in practical systems, where users may approach or interact with devices from arbitrary directions. Addressing this issue requires moving beyond conventional angle-specific modeling toward direction-agnostic learning paradigms that can abstract away spatial dependencies while preserving discriminative features of the underlying motion patterns.

Direction-agnostic RF sensing seeks to build representations that remain stable across multiple perspectives. The goal is not to discard spatial diversity, but to extract features that encode invariant relationships within the reflected signal structure. Point cloud data representations derived from radar measurements serve as an ideal foundation for this pursuit. Unlike range-Doppler or range-angle maps that are tightly coupled to antenna geometry and orientation, point cloud structures offer a geometric abstraction that captures spatial relations independently of the sensing angle. By learning directly from point clouds, it becomes possible to construct models that focus on the intrinsic dynamics of gestures rather than on their apparent spatial manifestation in a particular coordinate frame.

Beyond the technical motivation, the direction-agnostic paradigm directly contributes to HOLDEN's broader ethical and privacy-oriented objectives. In traditional sensing systems, rich spatial reconstructions may inadvertently reveal personally identifiable features such as body shape or behavioral patterns. Direction-agnostic modeling, in contrast, inherently limits the granularity of identifiable information by emphasizing relational rather than positional data. This aligns with the principles of *privacy by design* and ensures that perception systems derived from HOLDEN can operate responsibly within both industrial and societal contexts. Moreover, such generalization reduces the need for extensive personal data collection across all possible orientations, thereby minimizing both the data footprint and potential ethical risks associated with large-scale RF datasets.

The specific goal of this deliverable is to develop and validate machine learning models capable of recognizing gestures from point cloud data captured at one or a few angles, and to evaluate their ability to generalize across all predefined radar orientations. The deliverable thus demonstrates a practical instantiation of the direction-agnostic concept within the broader HOLDEN architecture. The outcome

is expected to inform subsequent stages of the project, where the integration of such models into multi-node and networked RF sensing systems will be explored. Achieving robust recognition with limited directional supervision not only enhances adaptability and scalability but also represents a step toward sustainable and ethically sound RF-based perception technologies.

This deliverable D5.4 builds directly on two previous outcomes of WP5:

- D5.1: Data structure definition for massive RF data input to DL which specified the unified storage, preprocessing, and frame-to-point-cloud conversion pipelines used throughout WP5;
- D5.3: Deep learning models for CSI-based holography and point-cloud-specific data, which delivered the baseline point-cloud encoders and initial multi-angle datasets that serve as the starting point for the direction-agnostic extensions presented here.

The present work therefore completes the progression from raw massive RF datasets (D5.1) via modality-specific deep models (D5.3) to fully direction-agnostic perception models required for real-world dense-network deployments.

1.1 Problem Statement

Direction-agnostic RF sensing aims to achieve robust perception independent of the relative orientation between the transmitter–receiver pair and the observed target. In contrast to conventional angle-specific models that perform well only within a narrow range of spatial configurations, direction-agnostic models are expected to retain consistent recognition accuracy even when trained on limited angular data. This property is critical for practical and ethical RF sensing, where the user's position and orientation with respect to the sensor are rarely fixed or predictable.

In this deliverable, we define the direction-agnostic sensing problem using point cloud representations derived from mmWave radar data. While the experiments are performed using mmWave radars, the concept extends to all RF modalities where spatial or temporal backscattered features can be converted into point clouds, such as WiFi Channel State Information (CSI), Ultra-Wide-Band (UWB), or sub-THz sensing. The abstraction from raw RF signal domains to spatially meaningful point clouds allows us to address the angular dependency challenge in a modality-independent manner.

1.2 Direction-Agnostic Sensing from Point Clouds

Point clouds provide a flexible intermediate representation that encodes the three-dimensional spatial distribution of reflection points generated by RF scattering events. Each point represents a spatial location where energy is reflected, with attributes such as range, angle, and intensity embedded in Cartesian form. When the radar or transmitter-receiver configuration changes, the global coordinates of these reflections rotate accordingly, yet the geometric and temporal relations among the points remain correlated to the underlying motion. The central idea of direction-agnostic sensing is to learn these intrinsic relations while becoming invariant to the extrinsic transformations induced by angle variations.

Formally, let $\mathbf{P}_i = \{\mathbf{p}_{i1}, \mathbf{p}_{i2}, \ldots, \mathbf{p}_{iM_i}\}$ denote the i-th point cloud corresponding to one radar measurement, where $\mathbf{p}_{ij} \in \mathbb{R}^3$ are the Cartesian coordinates of the j-th point. When the same gesture is observed from a different radar angle θ', the corresponding point cloud can be expressed as $\mathbf{P}'_i = \mathbf{R}_{\theta' \leftarrow \theta}\mathbf{P}_i + \mathbf{t}$, where $\mathbf{R}_{\theta' \leftarrow \theta}$ and $\mathbf{t}$ represent the spatial rotation and translation induced by viewpoint change. Direction-agnostic learning aims to find a function f_ϕ that preserves semantic consistency under such geometric transformations, i.e.,

$$f_\phi(\mathbf{P}_i) = f_\phi(\mathbf{R}_{\theta' \leftarrow \theta}\mathbf{P}_i + \mathbf{t}), \quad \forall \theta, \theta' \in \Theta.$$

Achieving this property requires the model to extract relational and temporal features that describe the motion pattern itself, rather than its appearance from a specific orientation.

1.3 Limitations of Angle-Specific Models

Gesture recognition models trained on mmWave radar data often rely on angle-specific features extracted from range–Doppler or range–angle domains. Such representations are inherently tied to the radar aperture and the user's position, leading to direction-dependent feature spaces. When evaluated from unseen viewpoints, these models suffer from a substantial drop in accuracy due to three primary factors:

- **Viewpoint-specific scattering patterns:** Small angular deviations result in non-linear changes in phase, Doppler, and amplitude distributions, which typical convolutional or recurrent architectures fail to normalize.
- **Angle-induced overfitting:** Models implicitly encode angle-specific priors during training, creating feature entanglement between class identity and spatial orientation.
- **High data collection burden:** To achieve uniform coverage, data must be collected for every angle of interest, drastically increasing both the recording effort and storage requirements.

These limitations restrict scalability, especially in multi-user or mobile settings where orientation is inherently dynamic. Moreover, such dependency undermines the goal of privacy-compliant sensing since it forces the collection of extensive raw spatial data across many angles.

1.4 Dataset Specification

As shown in Fig. 1, the dataset used in this deliverable comprises radar-based gesture recordings collected from eight distinct azimuthal directions around the subject [13]. Each gesture instance corresponds to a unique time–space distribution of reflection points captured by a mmWave radar. The radar operates in Frequency Modulated Continuous Wave (FMCW) mode, and each frame is converted into a 3D point cloud by projecting range, azimuth, and Doppler information into Cartesian coordinates. The eight radar positions uniformly cover the 360° horizontal plane, providing angular diversity for studying direction-dependent effects.

Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_8\}$ denote the discrete set of radar orientations. The dataset can be decomposed as:

$$\mathcal{D} = \bigcup_{\theta \in \Theta} \mathcal{D}_\theta, \quad \text{where } \mathcal{D}_\theta = \{(\mathbf{P}_i, y_i)\}_{i=1}^{N_\theta}.$$

Each $\mathcal{D}_\theta$ contains N_θ point clouds for various gesture classes $y_i \in \mathcal{Y}$. The training subset $\mathcal{D}_{\text{train}}$ is drawn from one or few selected angles $\Theta_{\text{train}} \subset \Theta$, while the test subset $\mathcal{D}_{\text{test}}$ is drawn from the complementary set $\Theta_{\text{test}} = \Theta \setminus \Theta_{\text{train}}$. This split enforces a cross-angle generalization challenge: the model must classify gestures at unseen angles without access to those orientations during training.

Beyond the multi-directional setup, the dataset used in the experiments further consists of 15,967 samples of depth-based gesture recordings. After collection, each sample is processed into a sequence of 28 frames, with each frame containing a point cloud of 120 points. These point clouds represent the 3D trajectories of participants' finger movements, where each point includes not only spatial coordinates but also additional attributes such as speed and amplitude. The dataset covers 16 gesture classes and is organized into 16 contexts that vary in rotation, angle, distance, and participant sets. The default configuration employs participant set 1 and places the subject 1 meter from the radar with both rotation

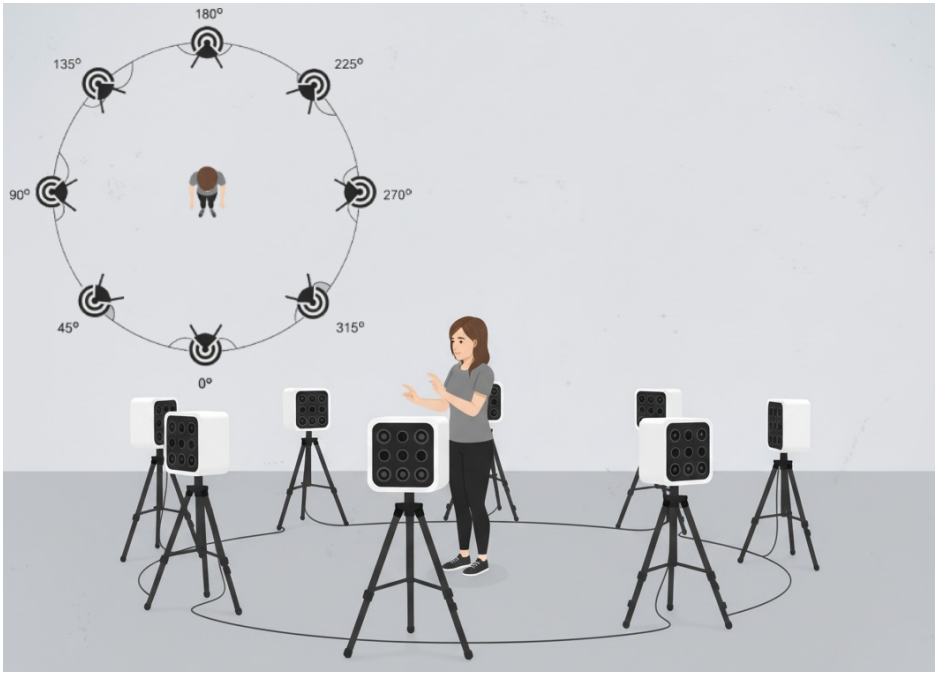

Figure 1: Positioning of the radar sensors relative to the participant, shown from side and top views. Eight identical mmWave radars are arranged equidistantly on a circle of radius $1.5\,\mathrm{m}$ centered on the participant, with all sensors mounted at a height of $1.2\,\mathrm{m}$ above the floor, ensuring full 360° azimuthal coverage at torso level.

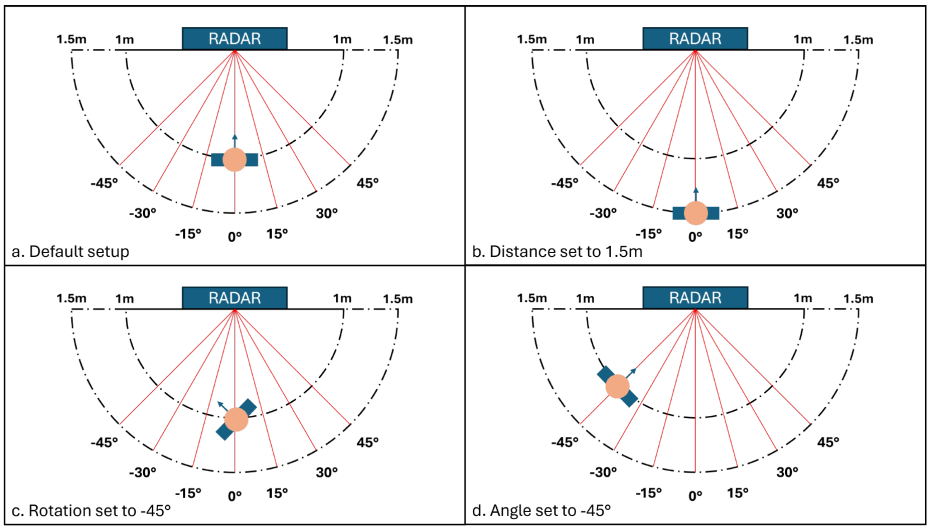

Figure 2: Illustration of the 16 data collection contexts with variations in rotation, angle, distance, and participant sets.

and angle set to 0° (see Fig. 2a). Contexts 1–15 contain only the first 12 gesture classes, while context 16 includes classes 13–16 under the default setup. Context 1 serves as the base context, and context 2 repeats the same settings but with a different participant set. In context 3, the subject stands 1.5 meters from the radar (see Fig. 2b). Contexts 4–9 adjust the rotation to 15°, 30°, 45°, -10°, -30°, and -45°, respectively (see Fig. 2c), whereas contexts 10–15 vary the angle using the same set of values (see Fig. 2d).

1.5 Research Objective and Question

The fundamental research question guiding this deliverable is:

> *How can a learning model trained on point cloud data from one or a few radar viewpoints achieve high classification accuracy across unseen angles while minimizing the required training data?*

This question encapsulates both the data efficiency and generalization requirements. The goal is to construct models that encode invariant latent representations, allowing recognition across unseen orientations without exhaustive multi-angle training. The underlying hypothesis is that geometric and temporal relations within the point cloud encode sufficient information for direction-invariant understanding of human gestures.

1.6 Mathematical Problem Formulation

Formally, let the RF sensing task be defined over the dataset $\mathcal{D} = \{(\mathbf{P}_i, y_i, \theta_i)\}_{i=1}^{N}$, where each $\mathbf{P}_i \in \mathbb{R}^{M_i \times 3}$ is a point cloud representing a single radar frame, y_i is its corresponding gesture label, and $\theta_i \in \Theta$ denotes the radar angle. The goal is to learn a classifier

$$f_\phi : \mathbb{R}^{M \times 3} \to \mathcal{Y},$$

parameterized by ϕ, that minimizes the expected loss across the entire angular domain:

$$\min_{\phi} \ \mathbb{E}_{(\mathbf{P},y,\theta)\sim\mathcal{D}} \left[\mathcal{L}(f_\phi(\mathbf{P}), y)\right],$$

subject to the angular invariance constraint:

$$\forall \mathbf{P},\ \forall \theta, \theta' \in \Theta : \quad f_\phi(\mathbf{P}) = f_\phi(\mathbf{R}_{\theta' \leftarrow \theta}\mathbf{P} + \mathbf{t}).$$

In practice, exact invariance is unattainable; hence, the objective is relaxed to minimizing the expected discrepancy between predictions under angular transformations:

$$\min_{\phi} \ \mathbb{E}_{\mathbf{P},\theta,\theta'} \left[\mathcal{L}_{\text{invar}}\big(f_\phi(\mathbf{P}), f_\phi(\mathbf{R}_{\theta' \leftarrow \theta}\mathbf{P})\big)\right],$$

where $\mathcal{L}_{\text{invar}}$ penalizes deviation between outputs for rotated versions of the same gesture. This formulation enforces that the learned representation remains stable under rotation while maintaining discriminative power for gesture classification.

1.7 Scope of the Problem

The scope of this problem extends beyond angle generalization in mmWave sensing. A solution to the defined objective would enable scalable and transferable RF perception across modalities and scenarios. In multi-sensor deployments, a direction-agnostic model can serve as a shared inference backbone for distributed sensing nodes without requiring per-node retraining. Furthermore, this formulation aligns with the HOLDEN project's goal of ethical and privacy-preserving perception: by operating on geometric abstractions instead of raw field patterns, the model inherently reduces personal identifiability and sensitivity of the captured data.

Thus, the problem addressed in this deliverable is both technically and ethically motivated: to construct a data-efficient, generalizable, and direction-agnostic RF sensing framework using point cloud representations that remain robust to viewpoint transformations across multiple radar modalities and configurations.

For the additional dataset, we aim to employ continual learning to accommodate data arriving incrementally from different orientations. Such an approach eliminates the need for repeatedly retraining the model from scratch and instead enables it to acquire new directional knowledge with modest computational cost while preserving previously learned capabilities. This facilitates a scalable and adaptive cross-direction sensing framework that evolves with incoming data.

2 Related Work

The fusion of RF sensing and machine learning techniques provides promising avenues for developing direction-agnostic sensing systems utilizing point cloud data, particularly within constrained environments and adhering to privacy requirements. This section reviews key literature pertaining to RF-based gesture recognition, methodologies for handling viewpoint variations in point cloud data, ethical and privacy considerations, and identifies the remaining gaps in achieving direction-agnostic RF point cloud sensing under privacy constraints.

2.1 RF Sensing for Gesture Recognition using mmWave Radars

RF technology, spanning from $30\,\mathrm{Hz}$ to $300\,\mathrm{GHz}$, has recently been adapted for gesture interaction, enabling device-free sensing through various modalities, including radar [8]. Radar sensing is particularly advantageous because it provides 3D spatial information, exhibits robustness against challenging conditions such as adverse weather and lighting, and possesses the capability to penetrate non-metallic surfaces. Furthermore, mmWave sensing operates using non-ionizing waves, ensuring safety for the human body. The proliferation of miniaturized, radar-on-chip mmWave devices, such as those from Texas Instruments (TI) and Google Soli, has transformed radar sensing into a commodity hardware suitable for integration into Internet of Things (IoT) and wearable devices. Compared to sub-$6\,\mathrm{GHz}$ RF systems, which are limited by larger wavelengths (over $5\,\mathrm{cm}$) and smaller bandwidths, mmWave sensing benefits from high bandwidths (typically $4\,\mathrm{GHz}$ to $7\,\mathrm{GHz}$), high directivity, and small antenna apertures, often only a few centimeters in size. mmWave radars can capture spatial information through sparse 3D point clouds, which are derived from range, angle, and Doppler data. This conversion from raw Analog to Digital Conversion (ADC) data to point clouds results in a massive reduction in data volume, transitioning from gigabytes to mere megabytes, which facilitates rapid data processing and transfer necessary for real-time machine learning algorithms. Gesture recognition approaches utilizing mmWave radar are typically categorized as either model-driven, which restrict classification to a limited set of gestures, or data-driven, which are scalable to larger gesture vocabularies. Many data-driven solutions

employ deep learning architectures combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) modules, to process features like Doppler or range-Doppler. However, the interpretability of these Doppler features can be limited, making the distinction of simultaneous movements difficult. The **Pantomime** [8] system addressed this by using sparse mmWave radar point clouds directly, positioning itself as a medium-resolution, medium-range, high-frequency sensing approach ($77\,\mathrm{GHz}$) capable of privacy-aware recognition that is robust to visibility issues. Pantomime demonstrated the efficacy of a hybrid deep learning architecture, combining PointNet++ [10] and LSTM, for extracting spatio-temporal features directly from sparse point clouds, achieving high accuracy in classifying a complex set of 21 gestures. Other models, such as RadHAR, have similarly used mmWave point clouds for human activity recognition. More recently, the **Tesla** model [12], which is based on a Message Passing Neural Network (MPNN) graph convolution approach, was specifically developed to process sparse mmWave radar point clouds and is designed for real-time operation on resource-constrained embedded devices like the Raspberry Pi 4 (in the **Tesla-Rapture** system).

2.2 Processing Dynamic Point Clouds and Addressing Viewpoint Variations

Point cloud processing methods in computer vision are broadly classified into multi-view (projection onto 2D planes), volumetric (voxelization), and direct processing. Approaches involving projection or voxelization often incur information loss and lead to increased computational and memory demands, hindering real-time performance. Direct processing models, such as PointNet [9] and PointNet++ [10], were groundbreaking for classifying stationary 3D shapes by ensuring permutation invariance to the input point order. PointNet++ further refined this by introducing hierarchical feature learning using Set Abstraction (SA) layers to capture local structures within the point set. However, gesture recognition relies on capturing temporal evolution, treating the data as 4D point clouds (3 spatial dimensions plus time). To address this, various architectures integrate temporal modeling:

- **Hybrid Architectures**: Systems like Pantomime [8] and PointGest [15] combine spatial feature extractors (PointNet/PointNet++) operating frame-wise with temporal modeling layers (LSTMs). This approach balances the need to preserve fine-grained spatial features (by applying PointNet++ on aggregated point clouds) with capturing the directionality of movement (via temporal features from frame sequences).
- **Graph Convolution Networks and Temporal Modeling**: Novel approaches, such as the **Tesla** model [12], avoid the recurrent processing of RNNs by reflecting the temporal evolution of the gesture directly within the graph structure using a Temporal Graph K-Nearest Neighbor (K-NN) algorithm. This graph is then processed using MPNN layers in a single forward pass, significantly improving computational efficiency for real-time applications.

A significant challenge in gesture recognition, especially utilizing RF sensing, is the dependence on the observation angle [13]. RF sensing systems are typically optimized for the boresight angle ($0°$), and accuracy degrades notably as the articulation angle increases (e.g., up to $20\,\%$ drop at -45 and 45). Addressing these viewpoint variations involves several technical strategies:

- **Data Augmentation**: A foundational technique to increase model generalizability. During training, techniques such as random translation (up to $10\,\mathrm{cm}$), scaling (0.8 to 1.25 factor), point-wise translation (jitter), and clipping are applied on-the-fly to make the model resilient to geometric transformations and positional variations.
- **Rotation Normalization and Spatial Transformers**: Preprocessing can normalize the point cloud position and angle. Pantomime applies rotation and translation based on the cluster centroid to

normalize the input data to a reference condition (e.g., $1.5\,\mathrm{m}$ distance, 0° angle). The Tesla model incorporates a spatial transformer network namely TFNet module to dynamically transform skewed input points into a rigid, canonical orientation, simplifying subsequent recognition layers.

- **Attention and Transformer Networks**: The use of attention mechanisms inherently helps focus on relevant features, improving robustness. For instance, the Tesla model integrates a multi-head self-attention mechanism within its MPNN layer, which enhances performance and helps filter noisy points without explicit outlier removal preprocessing.
- **Training on Multiple Angles (Angle-Agnostic Models)**: Recent work specifically addresses the need for rotation-resilient sensing by training models on data collected simultaneously from multiple angles [13]. Approaches include:
  1. **Angle-Invariant Prediction**: Aggregating features from multiple angles using symmetric pooling operators, such as Max pooling, Vote pooling (element-wise summation of class scores), or Attention pooling (using a multi-head self-attention mechanism to weight angle representation vectors).
  2. **Gesture Orientation Tracking**: This superior method involves training a separate encoder function dedicated specifically to each observation angle. During inference, the system selects the appropriate encoder based on the detected orientation, achieving exceptional resilience even when few angles are available, a capability significantly outperforming traditional models.

2.3 Ethical and Privacy Considerations in Sensing Technologies

The deployment of sensing technologies must inherently address ethical and privacy concerns. Vision-based systems, which utilize cameras (e.g., RGB-depth sensors like Microsoft Kinect), are known to raise significant privacy concerns due to their image capture capabilities [12, 14, 16]. In contrast, RF sensing, particularly systems utilizing mmWave radar, are often highlighted as a more privacy-preserving alternative. Furthermore, the radiation utilized in mmWave radar sensing is non-ionizing, meaning it is not considered dangerous to the human body. The contrast with cameras is crucial in environments sensitive to monitoring, such as in elderly care or smart home scenarios, where the psychological impact of monitoring on privacy-enhancing behaviors has been documented. This intrinsic characteristic of RF sensing makes it highly desirable for integration into environments where user privacy must be paramount, such as in smart homes or industrial settings.

2.4 Research Gaps in Direction-Agnostic RF Point Cloud Sensing

Despite advances in RF-based gesture recognition and point cloud processing, several critical gaps remain, particularly concerning direction-agnostic performance coupled with privacy constraints.

Most existing models, even advanced ones like Pantomime, are optimized for gestures articulated directly in front of the sensor (0° boresight angle). When the observation angle deviates or changes dynamically, recognition accuracy degrades severely. While solutions exist for training on multiple angles (e.g., Gesture Orientation Tracking), these models must be explicitly tailored to the properties of **sparse RF point clouds** generated by mmWave radar, which differ significantly from dense point clouds derived from RGB-D sensors. The efficacy of rotation-invariant feature transformations (like PCA) is compromised for human subjects because reflections from the large torso tend to be similar across different gestures, muddying the features needed for classification.

Moving towards angle-agnosticism often necessitates multi-device sensing to cover various viewpoints, as demonstrated in recent research. Centralized processing of raw multi-angle data, however,

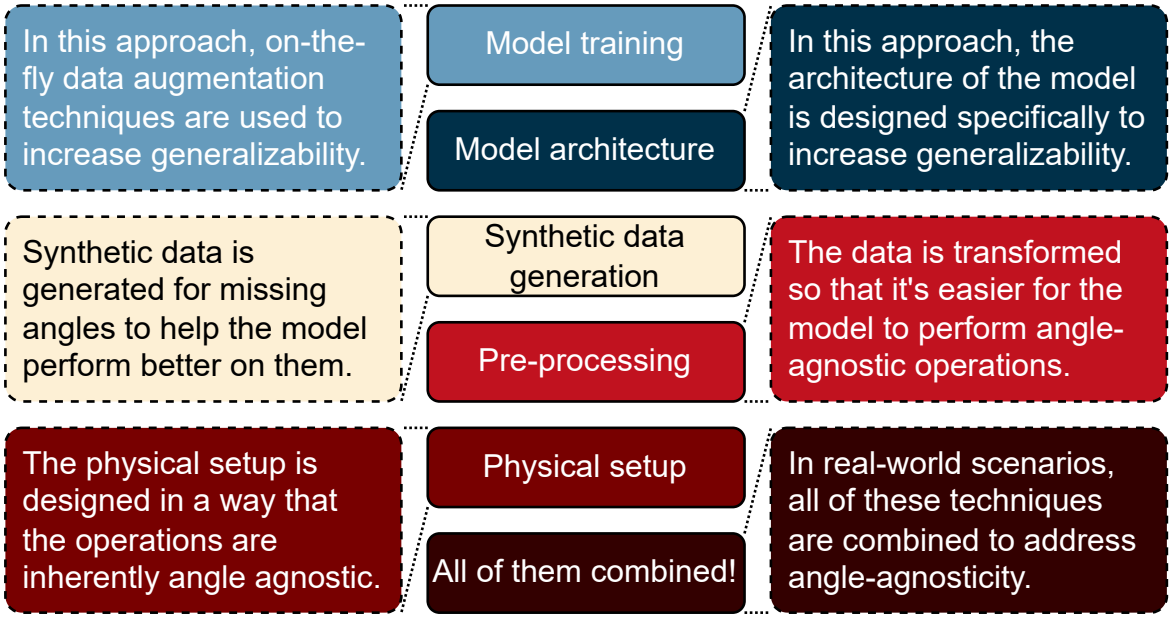

Figure 3: Different approaches towards angle-agnostic processing of RF sensing data

incurs high communication costs (due to transferring large raw point clouds) and poses scalability and privacy risks. The need exists for robust, direction-agnostic models specifically tailored for sparse RF point clouds that can seamlessly integrate with a **distributed processing architecture** where only minimal, non-raw data (e.g., gesture probability vectors or consensus votes) is exchanged, thereby significantly reducing communication load (by up to $99.3\,\%$) and enhancing privacy assurance.

The core identified gap is the absence of models that simultaneously achieve high direction-agnostic performance specifically on challenging sparse mmWave point clouds while inherently supporting privacy-preserving mechanisms necessary for distributed deployment in real-world constrained and sensitive applications.

3 Methodology

The methodologies developed in this deliverable reuse and extend the data ingestion and point-cloud generation pipelines standardised in D5.1 as well as the baseline PointNet++-LSTM and MPNN encoders delivered in D5.3. All experiments are consequently performed on the exact same data structures and pre-processing chain, ensuring full reproducibility and direct comparability with the results reported in the previous deliverables.

Our approach to achieving direction-agnostic RF sensing from point cloud data integrates multiple complementary techniques spanning the stages of data acquisition, pre-processing, model design, and training. The methodology is structured to systematically address the challenges associated with angle variability in real-world RF sensing scenarios, ensuring robust and generalizable model performance.

(a) Original (b) Permuted (c) Translated (d) Rotated (e) scaled

Figure 4: Different augmentation techniques that can be utilized to train a more generalizable model.

3.1 Model Training

A critical component of our methodology focuses on the training of models capable of generalizing across a wide range of orientations. In this approach, on-the-fly data augmentation strategies are employed to artificially increase the diversity of the training dataset. These augmentation techniques include random rotations, jittering of point coordinates, and simulated noise perturbations, all of which expose the model to a variety of input configurations that it may encounter during deployment. By incorporating these transformations during training, the model learns representations that are invariant to specific point cloud orientations, thereby enhancing its ability to generalize to previously unseen angles.

Since direction-agnostic RF sensing relies on stable feature extraction under arbitrary orientations and measurement conditions, the augmentation pipeline is designed to expose the model to a broad range of plausible geometric distortions that are shown in Fig.4. All transformations are applied independently for every mini-batch, ensuring that the network never encounters the same representation twice and is continuously challenged to generalize beyond the canonical geometry of the measurements.

A moderate random translation of up to $10\,\mathrm{cm}$ is applied to simulate small variations in the subject's position with respect to the radar coordinate frame. These perturbations force the model to learn translation-invariant features and reduce sensitivity to minor misalignments between measurement sessions. In addition, the point cloud is randomly scaled within the range $[0.8, 1.25]$, emulating variations in subject size, radar–target distance, and amplitude calibration differences. This scaling augmentation encourages the model to focus on structural patterns and motion signatures rather than absolute geometric dimensions.

To further mimic fine-grained sensor noise, a point-wise jitter is injected by adding Gaussian perturbations with $\mu = 0$ and $\sigma = 0.01$. This subtle distortion prevents the network from memorizing exact point locations and improves resilience to phase noise, quantization artifacts, and minor nonidealities in the radar signal processing chain. Complementing this, a random clipping of $0.03\,\mathrm{cm}$ removes a small spatial portion of the point cloud. Such partial occlusion approximates real-world situations in which parts of the target reflections may be attenuated or missing due to multipath, shadowing, or antenna pattern variations.

Finally, the point cloud is randomly shuffled while preserving the spatial coordinates and temporal ordering of each point. This ensures that the network does not rely on the specific ordering of points in memory, but instead learns representations grounded in geometry and physical consistency. Collectively, these augmentation techniques provide a diverse, continuously changing training distribution that enhances the model's ability to perform consistently across unseen orientations, positions, and sensing conditions [8, 9, 12].

3.2 Model Architecture

Complementing the training strategies, the design of the model architecture itself is oriented towards improving direction-agnostic performance. Architectural components are specifically chosen to capture the intrinsic geometric properties of the point cloud while minimizing sensitivity to rotation and orientation. For instance, layers that aggregate local and global spatial features, along with normalization and

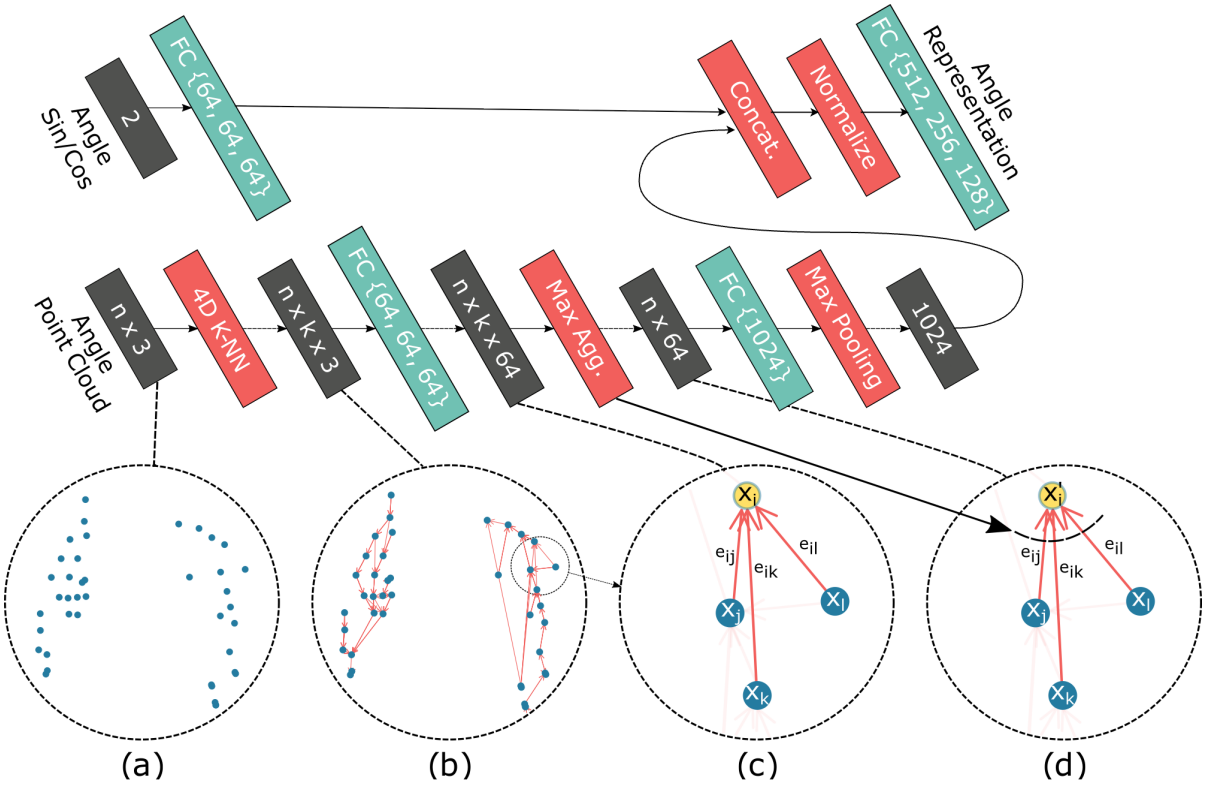

Figure 5: Schematic of the model. (a) Point cloud produced by the a radar for a given gesture; (b) graph construction using K-NN to capture the temporal dependencies within the sequence; (c) computation of edge features for all incident edges associated with a central node i; (d) generation of node-level representations by aggregating the corresponding edge features.

attention mechanisms, are employed to ensure that the learned embeddings remain consistent across different angular perspectives. The architecture is thereby intrinsically predisposed to generalize beyond the specific configurations observed during training, reducing the reliance on explicit data augmentation alone.

Unlike the baseline model [13], our proposed architecture integrates the angle of acquisition directly into the learning process as shwon in Fig. 5. While traditional approaches attempt to enforce strict rotation invariance, we argue that providing the model with explicit viewpoint information offers a more principled way to handle directional variations in RF measurements. From a perceptual standpoint, humans instinctively use viewpoint cues to interpret partially visible structures; knowing where an observation originates helps us mentally reconstruct occluded regions and reason about the occluder–target geometry. In a similar manner, radar point clouds inherently contain angle-dependent distortions, shadowing effects, and variations in reflection intensity. Explicit encoding of the sensor angle allows the model to contextualize these geometric differences rather than forcing it to infer them implicitly from the point distribution alone.

The motivation for this design choice also stems from the nature of RF sensing itself. Since the radar captures only a projection of the target that depends on both the angle and scattering characteristics, relying solely on the point cloud geometry can lead to ambiguous interpretations. Two distinct angles may produce similar point-level structures, while small shifts in direction can yield significant changes in visibility. Incorporating angular information therefore provides an additional axis of disambiguation that helps the model learn a more stable and semantically meaningful latent representation. This is particularly important for direction-agnostic sensing, where the goal is to develop a model that generalizes smoothly across all orientations without sacrificing sensitivity to fine-grained motion or structural cues.

To embed this angle information effectively, we first compute the sine and cosine of the acquisition angle. This representation ensures continuity across the full 360° range and avoids discontinuities inherent in raw angle values. The two-dimensional angle encoding is then passed through a dedicated Multi-Layer Perceptron (MLP), which extracts a compact, non-linear embedding that captures relationships between different viewpoints. This angular embedding is intentionally kept lightweight to avoid overpowering the geometric features extracted from the point cloud, while still providing enough expressive capacity to model complex dependencies.

In parallel, the point cloud data is processed using an MPNN-based encoder [13, 16], which maps the raw coordinates to a high-dimensional latent space through layers of neighborhood aggregation and feature propagation. The outputs of the angular MLP and the MPNN encoder are then concatenated to form a fused representation. Because these two embeddings may exist at different numerical scales, we apply normalization to enforce zero mean and unit variance prior to downstream processing. This step stabilizes training, prevents dominance of one modality over the other, and ensures that gradients are distributed uniformly across the fused vector.

Finally, the normalized fused representation is passed through an additional MLP responsible for extracting fine-grained cross-modal features and shaping the joint latent space. This network learns how geometric structures interact with viewpoint-specific characteristics and produces a viewpoint-aware embedding suitable for the downstream task, such as classification of motion patterns or recognition of subtle variations in the sensed target. By explicitly modeling the acquisition angle alongside the point cloud geometry, the proposed architecture achieves a richer and more robust understanding of the underlying physical scene, ultimately improving performance in direction-agnostic RF sensing.

Let $\mathbf{P} \in \mathbb{R}^{n\times 3}$ denote the input point cloud containing n points in Cartesian coordinates. The point cloud is first processed by a geometric encoder [13], resulting in a fixed-dimensional embedding

$$\mathbf{z}_{\text{pc}} \in \mathbb{R}^{1024},$$

which compactly summarizes the spatial and temporal characteristics of the point cloud.

In parallel, the acquisition angle θ is encoded using a continuous periodic representation. Instead of using the raw angle value, we compute its sine and cosine as we discussed before,

$$\mathbf{a} = [\sin\theta,\ \cos\theta] \in \mathbb{R}^2,$$

ensuring a smooth embedding across the entire 360° range. This two-dimensional vector is passed through a MLP, producing a viewpoint-aware feature representation

$$\mathbf{z}_\theta = f_{\text{angle}}(\mathbf{a}),$$

where $f_{\text{angle}}(\cdot)$ denotes the angle-encoding network.

The two embeddings are then fused through concatenation,

$$\mathbf{z}_{\text{fusion}} = [\mathbf{z}_{\text{pc}} \parallel \mathbf{z}_\theta]\,,$$

forming a joint representation that integrates geometric and angular cues. Since the two components may lie at different scales, we apply feature-wise normalization to the fused vector to enforce zero mean and unit variance, thereby stabilizing gradient propagation and preventing one modality from dominating the other.

A second MLP is subsequently applied to the normalized fused representation, mapping it to a compact latent space:

$$\mathbf{z} = f_{\text{fusion}}(\text{Norm}(\mathbf{z}_{\text{fusion}}))\,.$$

This final embedding $\mathbf{z}$ captures both the structural properties of the gesture and the directional information associated with the measurement. It is used as the input to the downstream classification layer (or any other task-specific head), enabling the model to perform viewpoint-aware recognition and to generalize more effectively across different acquisition angles.

3.3 Synthetic Data Generation

To further mitigate the challenges posed by incomplete angular coverage, synthetic data generation is employed as an auxiliary strategy. Missing angles in the dataset are compensated for by generating synthetic point cloud representations corresponding to those unobserved orientations. These synthetic samples enable the model to encounter a more uniform angular distribution during training, improving performance on angles that would otherwise be underrepresented. The generation process is carefully designed to preserve the physical realism of the point clouds, maintaining the statistical properties and spatial relationships present in real-world measurements.

An advanced approach to this augmentation leverages diffusion models to synthesize point clouds for unseen angles [3]. Consider a dataset where radar measurements are available from 7 distinct angles $\{\theta_1, \theta_2, \ldots, \theta_7\}$ for each object, while the goal is to generate data for an eighth unseen angle θ_8. To train the diffusion model, we create pairs from the existing angles: for a given object's point clouds $\{P_{\theta_1}, P_{\theta_2}, \ldots, P_{\theta_7}\}$, where each $P_{\theta_i} \in \mathbb{R}^{N\times 3}$ represents a point cloud with N points in 3D space, we form all possible ordered pairs $(P_{\theta_i}, P_{\theta_j})$ for $i \neq j$. This results in $7 \times 6 = 42$ pairs per object, allowing the model to learn transformations across diverse angular shifts.

The diffusion model is conditioned on an encoded representation of the input point cloud P_{θ_i} and the target angle θ_j (or the angular difference $\Delta\theta = \theta_j - \theta_i$). For encoding, we employ a MPNN to process the point cloud into a latent feature vector. The MPNN operates on the point cloud as a graph, where points are nodes and edges connect neighboring points based on spatial proximity. The message passing update for a node v at layer l is given by:

$$h_v^{(l+1)} = \phi\left(h_v^{(l)}, \bigoplus_{u\in\mathcal{N}(v)} \psi(h_v^{(l)}, h_u^{(l)}, e_{vu})\right),$$

where $h_v^{(l)}$ is the feature vector of node v at layer l, $\mathcal{N}(v)$ are its neighbors, $\bigoplus$ is an aggregation function (e.g., mean or max), ψ is the message function, ϕ is the update function, and e_{vu} are edge features (e.g., distances). After L layers, a global pooling yields the latent encoding $z_i = \text{MPNN}(P_{\theta_i}) \in \mathbb{R}^D$.

The diffusion model follows a Denoising Diffusion Probabilistic Model (DDPM) [5] framework, where the forward process gradually adds Gaussian noise to the target point cloud P_{θ_j} over T timesteps:

$$q(P_t|P_{t-1}) = \mathcal{N}(P_t; \sqrt{1-\beta_t}P_{t-1}, \beta_t\mathbf{I}),$$

with $P_0 = P_{\theta_j}$ and variance schedule $\{\beta_t\}_{t=1}^T$. The reverse process learns to denoise, parameterized by a neural network $\epsilon_\phi(P_t, t, z_i, \Delta\theta)$ that predicts the noise ϵ added at timestep t, conditioned on z_i and $\Delta\theta$. The training objective minimizes:

$$\mathcal{L} = \mathbb{E}_{t,\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\epsilon - \epsilon_\phi(\sqrt{\bar{\alpha}_t}P_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, z_i, \Delta\theta)\|_2^2\right],$$

where $\bar{\alpha}_t = \prod_{s=1}^t(1-\beta_s)$.

During inference, starting from pure noise $P_T \sim \mathcal{N}(0, \mathbf{I})$, we iteratively denoise to generate $\hat{P}_{\theta_j}$ conditioned on z_i and $\Delta\theta$. To synthesize the unseen angle θ_8, we select an input P_{θ_k} (e.g., the closest available angle), compute $z_k = \text{MPNN}(P_{\theta_k})$, set $\Delta\theta = \theta_8 - \theta_k$, and sample $\hat{P}_{\theta_8}$ from the diffusion model. This pairwise training ensures the model captures angular invariances and generates realistic point clouds for unobserved orientations, enhancing the dataset's completeness.

3.4 Pre-processing

Prior to model ingestion, the raw point cloud data undergoes a sequence of pre-processing transformations designed to enforce consistency across measurements originating from different sensing angles. The overarching goal is to condition the input space such that the subsequent learning modules operate on a representation that is as decoupled as possible from viewpoint-induced variability.

First, each point cloud is normalized with respect to its spatial extent to ensure that variations in absolute scale do not propagate into the learning pipeline. This step is complemented by centering the points around their geometric centroid, which removes global translational offsets introduced by the measurement setup or subject placement. A uniform scaling operation is then applied to map all point clouds into a canonical coordinate range, enabling the model to treat inputs from different recording sessions in a homogeneous manner.

To further attenuate angular dependencies, a rotational alignment procedure is employed. This alignment relies on estimating a stable reference axis from the underlying point distribution, typically derived from principal components or motion trajectories, and reorienting each point cloud accordingly. The effect is a partial neutralization of viewpoint-specific distortions while preserving the intrinsic spatial structure that encodes the relevant RF scattering properties.

In addition to these deterministic transformations, the pipeline may incorporate controlled perturbations such as random jittering or slight rotations to regularize the model and prevent overfitting to residual orientation cues which is already discussed in the previous subsection. These augmentations broaden the effective support of the training data and encourage the extraction of features that remain consistent under small geometric variations.

Collectively, the pre-processing steps reshape the raw RF point cloud into a standardized, angle-agnostic format. This ensures that the encoder operates on data where the dominant sources of variability are tied to the physical phenomenon of interest rather than sensor placement, ultimately improving the generalization capability of the downstream learning components.

3.5 Physical Setup

The sensing infrastructure consists of five FMCW radars mounted on the ceiling of the measurement area, as illustrated in Fig. 6. The radars are positioned in a distributed layout across a rectangular region, with inter-radar distances of approximately $5\,\mathrm{m}$ and $2\,\mathrm{m}$ along the longer and shorter axes, respectively. The ceiling height is fixed at $5\,\mathrm{m}$, placing all radars at a significantly elevated vantage point relative to the user.

This top-down configuration is deliberately chosen to mitigate common challenges encountered in ground-level or body-level sensor placements. When radars are positioned around the user, the human body introduces substantial *occlusion*, *cluttering* from multipath reflections, and *shadowing* effects that degrade the quality and completeness of the recorded point clouds. By contrast, mounting the radars on the ceiling ensures an unobstructed line-of-sight to the user throughout the sensing area. As a result, the captured RF scattering patterns are more stable, less noisy, and more consistent across different motion trajectories.

An additional advantage of this overhead arrangement is that the distance between the user and each radar remains nearly constant, even as the user moves within the horizontal plane. Let the ceiling height be $h = 5\,\mathrm{m}$ and let the user's horizontal displacement relative to a given radar be d. The slant range $R(d)$ from the radar to the user is

$$R(d) = \sqrt{h^2 + d^2}. \tag{1}$$

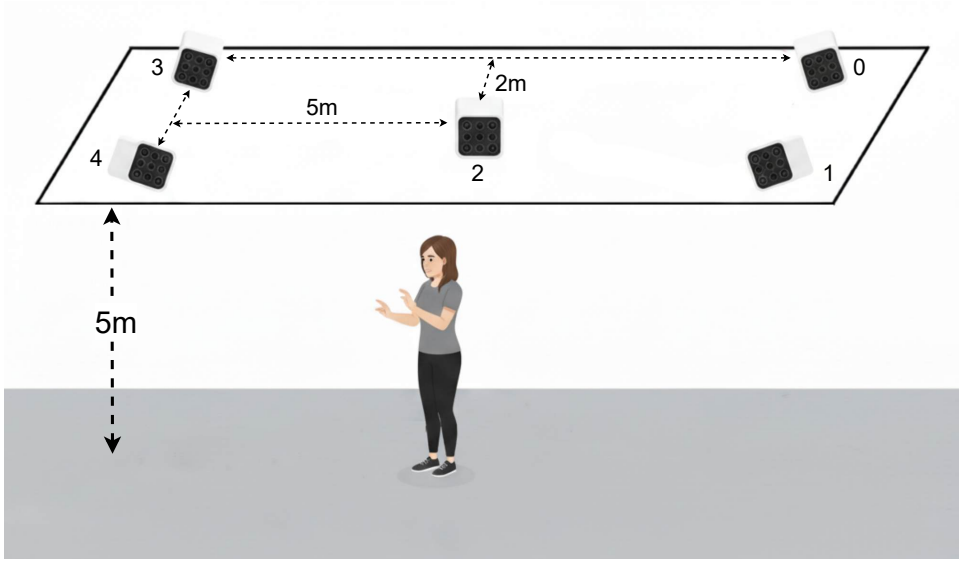

Figure 6: The schematic view of the setup where the radars or the RF sensors are mounted on the ceiling to have a better angle-agnostic features.

Because $h \gg d$ for typical indoor movements, the relative change in range is minimal. For example, for displacements of $d = 2\,\text{m}$ and $d = 5\,\text{m}$, the corresponding distances are

$$R(2) = \sqrt{5^2 + 2^2} = \sqrt{29} \approx 5.39\,\text{m}, \tag{2}$$

$$R(5) = \sqrt{5^2 + 5^2} = \sqrt{50} \approx 7.07\,\text{m}. \tag{3}$$

Compared to the baseline $R(0) = 5\,\text{m}$, these variations correspond to increases of only 7.8% and 41.4% respectively, despite the user moving over a large floor area. Within the central region of interest—where most gestures and activities are performed—the actual horizontal displacements are typically much smaller (e.g., $d \leq 2\,\text{m}$), resulting in less than 8% change in range. This small variation preserves the signal strength and radar cross-section properties, ensuring high-quality point cloud measurements across the entire usable space.

Overall, the ceiling-mounted multi-radar configuration provides wide spatial coverage, reduces orientation-dependent artefacts, and maintains stable sensing conditions. This design enables users to move freely within an extended area while allowing the model to operate under highly consistent geometric conditions, which is critical for reliable direction-agnostic RF sensing.

3.6 Continual Learning Approach

The section is divided into two parts. Subsection 3.6.1 demonstrates the architecture of the base network, to which continual learning methods are applied. Subsection 3.6.2 presents the formulas and algorithm details of the compared methods.

3.6.1 Base Model Architecture

There are three main components in the base network—the set abstraction module, the GRU layer, and the classifier (see Figure 7). The set abstraction module is adapted from the set abstraction layers

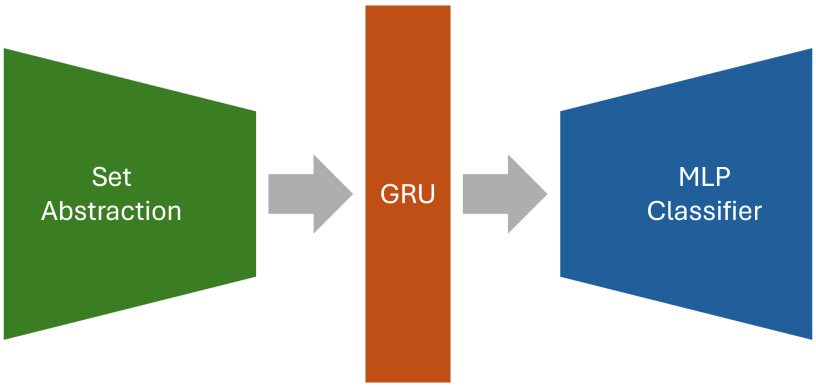

Figure 7: Base model architecture.

of PointNet++. It takes a sequence of frames as input and applies sampling, grouping, and feature extraction repeatedly. This results in a more compact sequence of frames, but it still contains important features of the original sequence. The GRU layer then processes the new sequence and outputs a vector that captures the geometric and temporal features across all frames. This vector is passed to the classifier to predict the class to which the original sequence belongs.

3.6.2 Continual Learning Methods

Baseline methods include None [17], Joint [17], elastic weight consolidation (EWC) [6], synaptic intelligence (SI) [19], learning without forgetting (LwF) [7], experience replay (ER) [2, 11], and averaged gradient episodic memory (A-GEM) [1].

None

None, which serves as the lower baseline, is the naive approach where the model incrementally learns the training data of each context [18]. In the None method, there is no attempt to mitigate catastrophic forgetting.

Joint

Joint is the higher baseline, where all the training data are available at the beginning [18]. This is similar to the traditional deep learning settings.

Elastic Weight Consolidation

For EWC, a regularization term $\mathcal{L}_{\text{EWC}}(\theta)$ is added to the total loss [6]. During training, in context K, $\mathcal{L}_{\text{EWC}}(\theta)$ can be formulated as:

$$\mathcal{L}_{\text{EWC}}(\theta) = \lambda \sum_{k=1}^{K-1} \left(\frac{1}{2} \sum_{i=1}^{N_{\text{params}}} F_i^{(k)} \left(\theta_i - \hat{\theta}_i^{(k)} \right)^2 \right),$$

where λ is the regularization strength (set to 10 based on hyperparameter tuning; see Subsection 4.3 for details), $\hat{\theta}_i^{(k)}$ is the i-th parameter after the training of context k, θ_i is the current value of the i-th

parameter, and $F_i^{(k)}$ is the Fisher information for parameter θ_i [6, 18]. The Fisher information of the i-th parameter measures the importance of the i-th parameter to the input-output mapping of the model in context k:

$$F_i^{(k)} = \frac{1}{|S^{(k)}|} \sum_{x \in S^{(k)}} F_{i,x}^{(k)},$$

with

$$F_{i,x}^{(k)} = \sum_o p_{\hat{\theta}^{(k)}}(o \mid x) \left(\frac{\delta}{\delta \theta_i} \log p_\theta(o \mid x) \Big|_{\theta = \hat{\theta}^{(k)}} \right)^2,$$

where $S^{(k)}$ is the input set in context k and $p_\theta(o \mid x)$ is the probability—predicted by the model with parameter θ—that input x belongs to class o [6, 18]. Thus, EWC effectively discourages significant changes in important parameters by minimizing the difference between the model parameters and context-optimized parameters $\hat{\theta}_i^{(k)}$ of each context k [6].

Synaptic Intelligence

In SI, the loss function includes a surrogate loss $\mathcal{L}_{\text{SI}}(\theta)$ [19]. The surrogate loss during context $K > 1$ is defined as

$$\mathcal{L}_{\text{SI}}(\theta) = \lambda \sum_i^{N_{\text{params}}} \Omega_i^{(K-1)} \left(\tilde{\theta}_i - \theta_i \right)^2,$$

where λ is the regularization strength (set to 0.1 based on hyperparameter tuning; see Subsection 4.3 for details), $\tilde{\theta}_i$ is the i-th parameter at the beginning of the current context, θ_i is the current value of the i-th parameter, and $\Omega_i^{(K-1)}$ is the importance of the i-th parameter [18, 19]. The importance $\Omega_i^{(K-1)}$ is updated after the training of each context as follows:

$$\Omega_i^{(K-1)} = \sum_{k=1}^{K-1} \frac{\omega_i^{(k)}}{\left(\hat{\theta}_i^{(k)} - \hat{\theta}_i^{(k-1)} \right)^2 + \xi},$$

where $\hat{\theta}_i^{(k)}$ is the i-th parameter after the training of context k, ξ is set to 0.1 to avoid zero division, and $\omega_i^{(k)}$ is updated after each iteration according to the formula

$$\omega_i^{(k)} = \sum_{t=1}^{N_{\text{iters}}} \left(\theta_i^{(k)}[t] - \theta_i^{(k)}[(t-1)] \right) \frac{-\delta}{\delta \theta_i} \mathcal{L}^{(k)}[t],$$

where $\theta_i^{(k)}[t]$ is the i-th parameter after the t-th iteration of context k, and $\mathcal{L}^{(k)}[t]$ is the total loss (including the surrogate loss) at the t-th iteration of context k [18, 19]. Similar to EWC, SI also discourages radical changes in important parameters. However, while EWC uses Fisher information to estimate the importance of parameters, SI calculates the importance by accumulating the contribution of each parameter to the loss reduction over the course of training [19].

Learning without Forgetting

Similar to other regularization methods, the loss function of LwF also includes a regularization loss—$\mathcal{L}_{\text{LwF}}(\theta)$ [7]. According to the original paper [7], during context K, $\mathcal{L}_{\text{LwF}}(\theta)$ is defined as knowledge distillation loss [4]:

$$\mathcal{L}_{\text{LwF}}(\theta) = -\lambda \sum_o p_{\hat{\theta}^{(K-1)}}^T(o \mid x) \log \left[p_\theta^T(o \mid x) \right],$$

where λ is the regularization strength (set to 1 based on hyperparameter tuning; see Subsection 4.3 for details), θ is the current parameter, $\hat{\theta}^{(K-1)}$ is the parameter in the end of the last context, and $p_\theta^T(o \mid x)$ is the softmax with temperature, which follows the formula

$$p_\theta^T(o \mid x) = \text{softmax}(z_o^{(x,\theta)}/T) = \frac{\exp\left[z_o^{(x,\theta)}/T\right]}{\sum_j \exp\left[z_j^{(x,\theta)}/T\right]}.$$

LwF effectively utilizes knowledge distillation to discourage changes in the input-output mapping of the model from the older version of itself [7].

Experience Replay

In ER, a memory buffer with a finite budget is used to store training data from past contexts [11]. During training, the final training dataset of a context is the union of the memory buffer and the dataset of that context. This is different from the algorithm in [2], where a minibatch is sampled from the memory buffer to be combined with the current minibatch. After the training of each context, the memory buffer is updated according to Algorithm 1. ER is a simple and straightforward approach to handle catastrophic forgetting, but it is one of the most effective ones [2, 18].

Algorithm 1: Update Memory Buffer

Input: $\mathcal{D}_{new}$ (new dataset), $\mathcal{M}$ (memory buffer), B (total budget), C (number of contexts so far)

$new\_size \leftarrow \lfloor B/C \rfloor$; `// Compute new size per context`

foreach $context\_index \in \mathcal{M}$ **do**

  $\mathcal{M}[context\_index] \leftarrow \mathcal{M}[context\_index][: new\_size]$; `// Truncate samples`

end

$samples \leftarrow \text{RandomSample}(\mathcal{D}_{new},\ new\_size)$; `// Sample subset`

$new\_context\_index \leftarrow C - 1$; `// Assign new context ID`

$\mathcal{M}[new\_context\_index] \leftarrow samples$; `// Store samples in memory buffer`

Averaged Gradient Episodic Memory

Similar to ER, A-GEM also maintains a memory buffer with a finite budget, and it also utilizes Algorithm 1 to manage the buffer [1]. During training, besides the current gradient $g_{current}$ obtained from the loss on the current training batch, A-GEM also computes the replay gradient g_{replay} using the loss on a randomly sampled batch from the memory buffer [1]. The final gradient is computed according to Algorithm 2. If the vector angle between $g_{current}$ and g_{replay} is greater than 90 degrees (i.e., the dot product of $g_{current}$ and g_{replay} is negative), $g_{current}$ is projected to be more aligned with g_{replay} [1]. Otherwise, $g_{current}$ is left unchanged [1]. A-GEM efficiently uses the alignment of $g_{current}$ and g_{replay} to ensure parameters are updated in the correct direction, which effectively mitigates catastrophic forgetting [1].

Algorithm 2: Gradient Calculation and Projection in A-GEM [1]

Input: B_n (current batch), $\mathcal{M}$ (episodic memory), θ (parameters)

$B_M \leftarrow$ Sample minibatch from $\mathcal{M}$; // Sample memory batch

$\gamma \leftarrow 1 \times 10^{-7}$; // Small constant for numerical stability

$g_{current} \leftarrow \nabla_\theta \mathcal{L}(B_n)$; // Compute gradient on current batch

$g_{replay} \leftarrow \nabla_\theta \mathcal{L}(B_M)$; // Compute reference gradient

if $\langle g_{current}, g_{replay} \rangle \geq 0$ **then**

  $g_{final} \leftarrow g_{current}$; // No projection needed

end

else

  $g_{final} \leftarrow g_{current} - \frac{\langle g_{current}, g_{replay} \rangle}{\|g_{replay}\|^2 + \gamma} g_{replay}$; // Project gradient

end

return g_{final}

3.7 Integrated Approach

In practice, real-world deployments combine all of the above strategies to achieve robust direction-agnostic RF sensing. Model training, architecture design, synthetic data generation, pre-processing transformations, and carefully engineered physical setups are leveraged in conjunction to address angular variability comprehensively. This integrated methodology ensures that the resulting models can reliably interpret point cloud data across diverse orientations, supporting high-fidelity RF sensing applications with minimal angular bias.

Overall, our methodology emphasizes a multi-level approach to angle-agnostic RF sensing, where algorithmic, data-driven, and hardware considerations are systematically combined. This comprehensive framework establishes a robust foundation for subsequent evaluation, deployment, and real-world utilization of direction-invariant RF sensing models.

4 Results

This section covers the results of the experiments. The section is divided into 4 subsections. Subsection 4.1 describes the setup of the experiments. Subsection 4.2 introduces the evaluation metrics that are used to assess the performance of the methods. Subsection 4.3 demonstrates the setup for hyperparameter tuning. Subsection 4.4 presents and analyzes the results.

4.1 Experiments Setup

The dataset is split into a training dataset with 60% of the data, a validation dataset with 20% of the data, and a test dataset with 20% of the data. In each epoch, the model is trained on the training dataset, and the evaluation metrics are collected with the validation dataset. After all epochs are finished, the model is tested using the test dataset to evaluate its final performance.

The model is trained for 14 epochs for each context. Since there are 16 contexts, the total number of epochs is 224. For the Joint method, where there is only one context with the full dataset, the model is also trained for 224 epochs. In the experiments, cross-entropy loss is used as the base loss function, which is standard for classification problems. For methods that require a memory buffer, the buffer size is set to 300 samples, which is approximately 3.1% of the size of the training dataset.

4.2 Evaluation Metrics

To assess the continual learning capacities of the compared methods, three evaluation metrics are utilized in this work: **Final Test Accuracy (FTA)**, **Seen-Context Accuracy (SCA)** across contexts, and **Current-Context Accuracy (CCA)** across contexts. All metrics are based on accuracy, which is formularized as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}.$$

FTA evaluates the model based on its accuracy when tested on the test dataset. SCA captures the performance of the model on the union of the validation datasets of the seen contexts, including the current one. On the other hand, CCA assesses the model using solely the validation dataset of the current context. While FTA provides an overview of the final performance of the model, SCA evaluates the ability of the model to retain learned information from the previous contexts, and CCA reflects the ability of the model to adapt to new training data. Furthermore, for each compared methods (i.e., EWC, SI, LwF, ER, and A-GEM), the differences between the evaluation metrics of the method and of the None method are also calculated, which is referred to as FTA-diff, SCA-diff and CCA-diff. For example, FTA-diff of EWC is computed as

$$\text{FTA-diff}_{\text{EWC}} = \text{FTA}_{\text{EWC}} - \text{FTA}_{\text{None}}.$$

In the experiments, for the Joint method, SCA and CCA evaluations are conducted at the end of the training. For other methods, these metrics are measured after the training of each context.

4.3 Hyperparameter Tuning

For EWC, SI, and LwF, the regularization strength λ is tuned on a logarithmic scale, and the candidate values are 0.01, 0.1, 1, 10, and 100. For each candidate value, the model is trained on 144 epochs (i.e., nine epochs per context). The evaluation metrics used to select the best hyperparameter value are Average CCA and Average SCA, which is the average of CCAs and SCAs across all contexts, respectively. The chosen value for each method is determined by the sum of the two metrics. For the hyperparameter tuning results, see Appendix C.

4.4 Results

In the remainder of this deliverable, the term **regularization-based methods** refers to EWC, SI, and LwF; the term **continual learning methods** refers to all methods apart from the Joint method and the None method; the term **incremental methods** refers to all methods except the Joint method; the term **domain-incremental phase** refers to the phase from Context 1 to Context 15; and the term **class-incremental phase** refers to Context 16.

Method	FTA	FTA-diff
Joint	**98.8%**	
ER	85.0%	+25.3%
A-GEM	81.9%	+22.2%
SI	66.7%	+7.0%
LwF	64.8%	+5.1%
EWC	63.8%	+4.1%
None	**59.7%**	**0.0%**

Table 1: Final Test Accuracies of Compared Methods and Their Differences from the None Baseline

Table 1 presents the FTA of the compared methods in descending order. From the table, it can be observed that the FTAs of all incremental methods are between the higher baseline (i.e., the FTA of the Joint method) and the lower baseline (i.e., the FTA of the None method). Among the incremental methods, ER has the highest FTA with 85.0%, followed by A-GEM with 81.9%, SI with 66.7%, LwF with 64.8%, EWC with 63.8%, and None with 59.7%. Notably, there is a significant gap between the FTAs of the regularization-based methods and of the replay-based methods. While the FTA of ER and A-GEM exceeds 80%, the FTA of EWC, SI, and LwF is less than 70%, and the FTA of None is less than 60%. The results of the FTAs highlighted the superiority of replay-based methods over the None method and regularization-based methods.

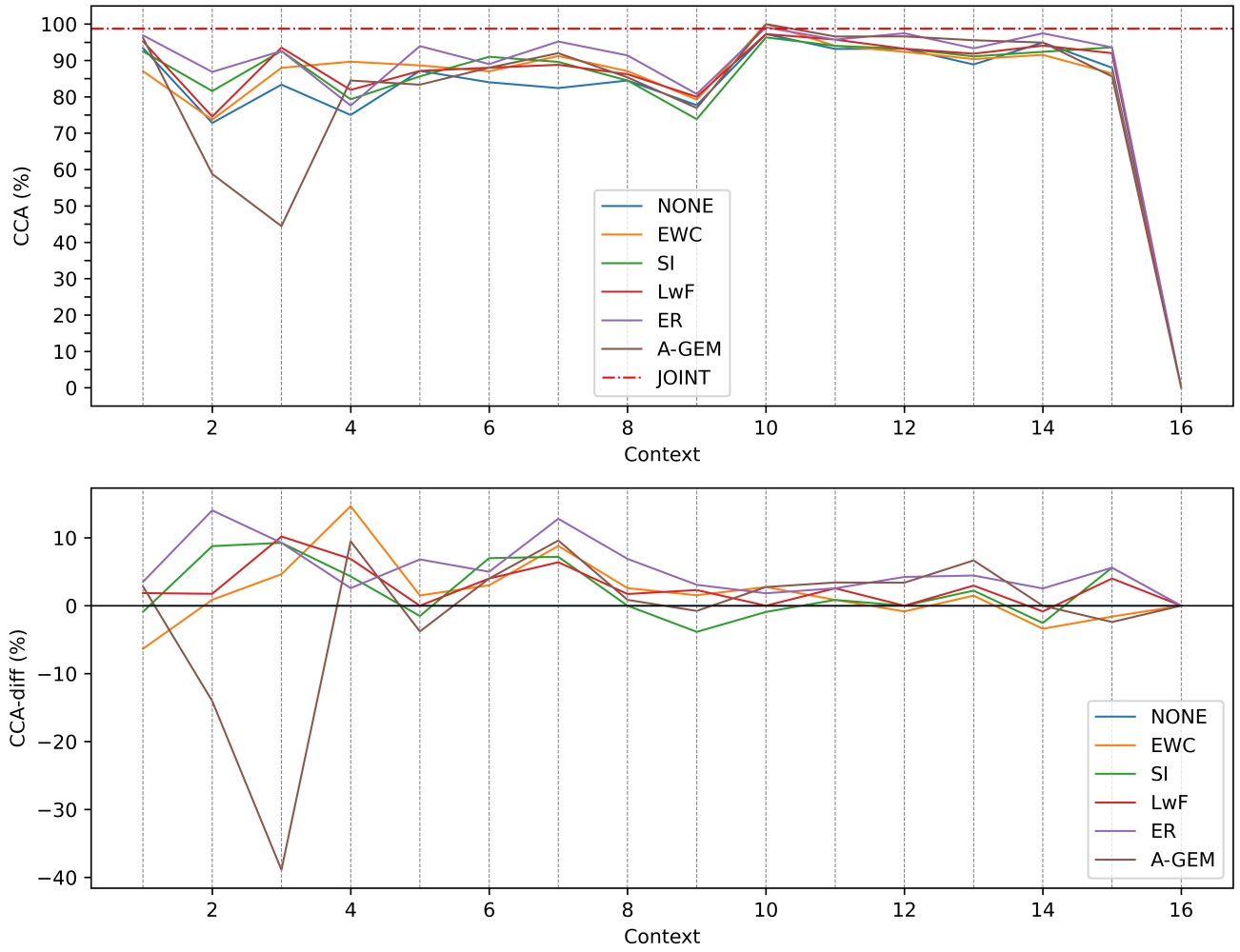

Figure 8: Current-Context Accuracies of Compared Methods in Contexts

The upper subplot of Figure 8 shows the CCA evaluations of the compared methods, and the lower subplot presents the difference in CCAs between the None method and other incremental methods. The detailed version of Figure 8 is provided in Appendix A. Apart from being surpassed by two replay-based methods in Context 10, the Joint method outperforms all other methods in all other contexts. Generally, most methods manage to maintain a CCA above 70% in the domain-incremental phase, before dropping to 0% in the last context. In particular, ER shows excellent performance by being on the top in multiple contexts and being the only incremental method to match or surpass the None method in all contexts.

A-GEM experiences a drop in contexts 2 and 3, then recovers and reaches its peak at Context 10 with a remarkable CCA of 100%. Moreover, regularization-based methods show a relatively stable progression of CCA-diffs across contexts; and they managed to exceed the None method in at least nine contexts. Among these methods, EWC stands out as the method with the best CCA in Context 4. Apart from this case, the CCA-diffs of all regularization-based methods in all contexts fall in the $\pm 10\%$ range. The results of the CCAs highlighted the superior performance of ER and the unremarkable performance of other compared methods.

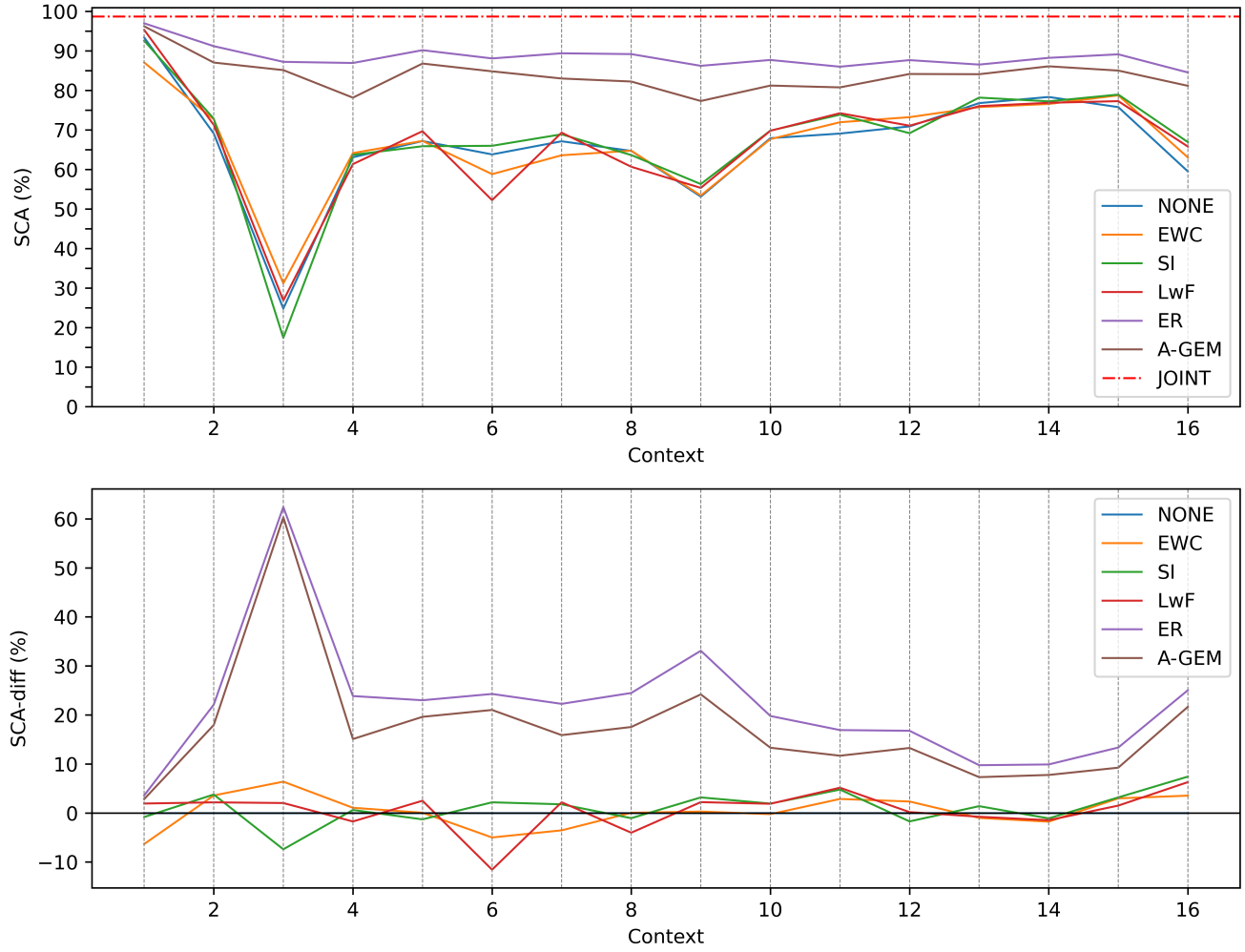

Figure 9: Seen-Context Accuracies of compared methods in contexts.

The upper subplot of Figure 9 illustrates the SCA evaluations of the compared methods, and the lower subplot presents the difference in SCAs between the None method and other incremental methods. The detailed version of Figure 9 is provided in Appendix B. In this figure, no incremental method manages to outperform the Joint method in any context. Meanwhile, ER once again shows excellent performance by surpassing all incremental methods in all contexts. This is followed by A-GEM, which demonstrates a remarkable trajectory of SCA and outperforms all regularization-based methods and the None method from the second to the last context. Furthermore, the overall trends in the SCAs of the None method and regularization-based methods aligned closely with each other. They experience a drop in the second context, reach their trough in the third context, then recover and remain in the 50%–70% region in the

subsequent contexts. Notably, these behaviors are similar to the behavior of A-GEM in the upper subplot of Figure 8. Furthermore, for most contexts, the SCA-diffs of regularization-based methods stay in the $\pm 10\%$ region, while the SCA-diffs of replay-based methods exceed 10%. The results of the SCAs further demonstrate the excellence of ER and the modest performance of regularization-based methods. In addition, they also highlight the outstanding outcomes achieved by A-GEM.

5 Discussion

The results presented in this deliverable demonstrate that direction-agnostic RF sensing from point cloud data is not only feasible but can be achieved with practical and scalable machine learning approaches that require only limited angular supervision. By combining aggressive on-the-fly geometric augmentation, viewpoint-conditioned architectures, synthetic data generation via diffusion models, and, in the continual-learning setting, replay-based mitigation of catastrophic forgetting, we have shown substantial gains in cross-angle generalization compared to conventional angle-specific baselines.

A central observation is the clear superiority of replay-based continual learning strategies (ER and A-GEM) over purely regularization-based approaches (EWC, SI, LwF). Even with a modest memory buffer of approximately $3\,\%$ of the training set, Experience Replay consistently outperformed all other incremental methods across Final Test Accuracy, Seen-Context Accuracy, and Current-Context Accuracy. This finding reinforces the broader continual-learning literature: retaining a small but representative subset of past experiences remains one of the most effective defences against catastrophic forgetting, especially when new tasks (here: new observation angles or distances) share significant feature overlap with previous ones. Regularization alone, while helpful in early contexts, proves insufficient when the model must repeatedly adapt to geometrically transformed versions of the same underlying motion patterns.

The explicit conditioning of the model on acquisition angle (via sinusoidal embedding and lightweight fusion MLP) proved surprisingly effective. Rather than enforcing strict rotation invariance – which is difficult with sparse, torso-dominated mmWave point clouds – providing the network with reliable side information about the sensor viewpoint allows it to learn compensatory mappings that preserve discriminative power. This viewpoint-aware paradigm aligns closely with human perceptual strategies and yields a more stable latent space than pure invariance-seeking approaches such as canonical alignment via PCA or heavy rotation augmentation alone. Importantly, the angle information can be obtained at inference time.

6 Conclusion

In conclusion, the continual learning section compares the performance of different continual learning methods when applied to RF-based gesture recognition. The compared continual learning methods include three regularization-based methods and two replay-based methods. Throughout the experiments, it can be observed that the two replay-based methods performed better than the three regularization-based methods. This suggests that retaining past data (i.e., maintaining a memory buffer) plays a significant role in continual learning, while having a regularization term may be insufficient to mitigate catastrophic forgetting. Surprisingly, the performance of all continual learning methods is good when it comes to new angles. However, it still has some limitations regarding the number of compared methods, the range of evaluation metrics utilized, and the benchmarking of computational efficiency due to time and resource constraints. Future work should explore more methods or develop new ones, employ more evaluation metrics, and standardize hardware to allow the measurement of training time.

References

[1] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem, Jan. 2019. arXiv preprint.

[2] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning, Jun. 2019. arXiv preprint.

[3] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10850–10869, 2023.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, Mar. 2015. arXiv preprint.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[6] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, Mar. 2017.

[7] Zhizhong Li and Derek Hoiem. Learning without forgetting, Feb. 2017. arXiv preprint.

[8] Sameera Palipana, Dariush Salami, Luis A Leiva, and Stephan Sigg. 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):1–27, 2021.

[9] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[10] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[11] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience replay for continual learning, Nov. 2019. arXiv preprint.

[12] Dariush Salami, Ramin Hasibi, Sameera Palipana, Petar Popovski, Tom Michoel, and Stephan Sigg. Tesla-rapture: A lightweight gesture recognition system from mmwave radar sparse point clouds. *IEEE Transactions on Mobile Computing*, 22(8):4946–4960, 2022.

[13] Dariush Salami, Ramin Hasibi, Stefano Savazzi, Tom Michoel, and Stephan Sigg. Angle-agnostic radio frequency sensing integrated into 5g-nr. *IEEE Sensors Journal*, 2024.

[14] Dariush Salami, Wanru Ning, Kalle Ruttik, Riku Jäntti, and Stephan Sigg. A joint radar and communication approach for 5g nr using reinforcement learning. *IEEE Communications Magazine*, 61(5):106–112, 2023.

[15] Dariush Salami, Sameera Palipana, Manila Kodali, and Stephan Sigg. Motion pattern recognition in 4d point clouds. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020.

[16] Dariush Salami and Stephan Sigg. Zero-shot motion pattern recognition from 4d point-clouds. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2021.

[17] Gido M. van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. Continual learning and catastrophic forgetting, Mar. 2024. arXiv preprint.

[18] Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, Dec. 2022.

[19] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence, Jun. 2017. arXiv preprint.

A Detailed Current-Context Accuracies (CCAs) Results Figure

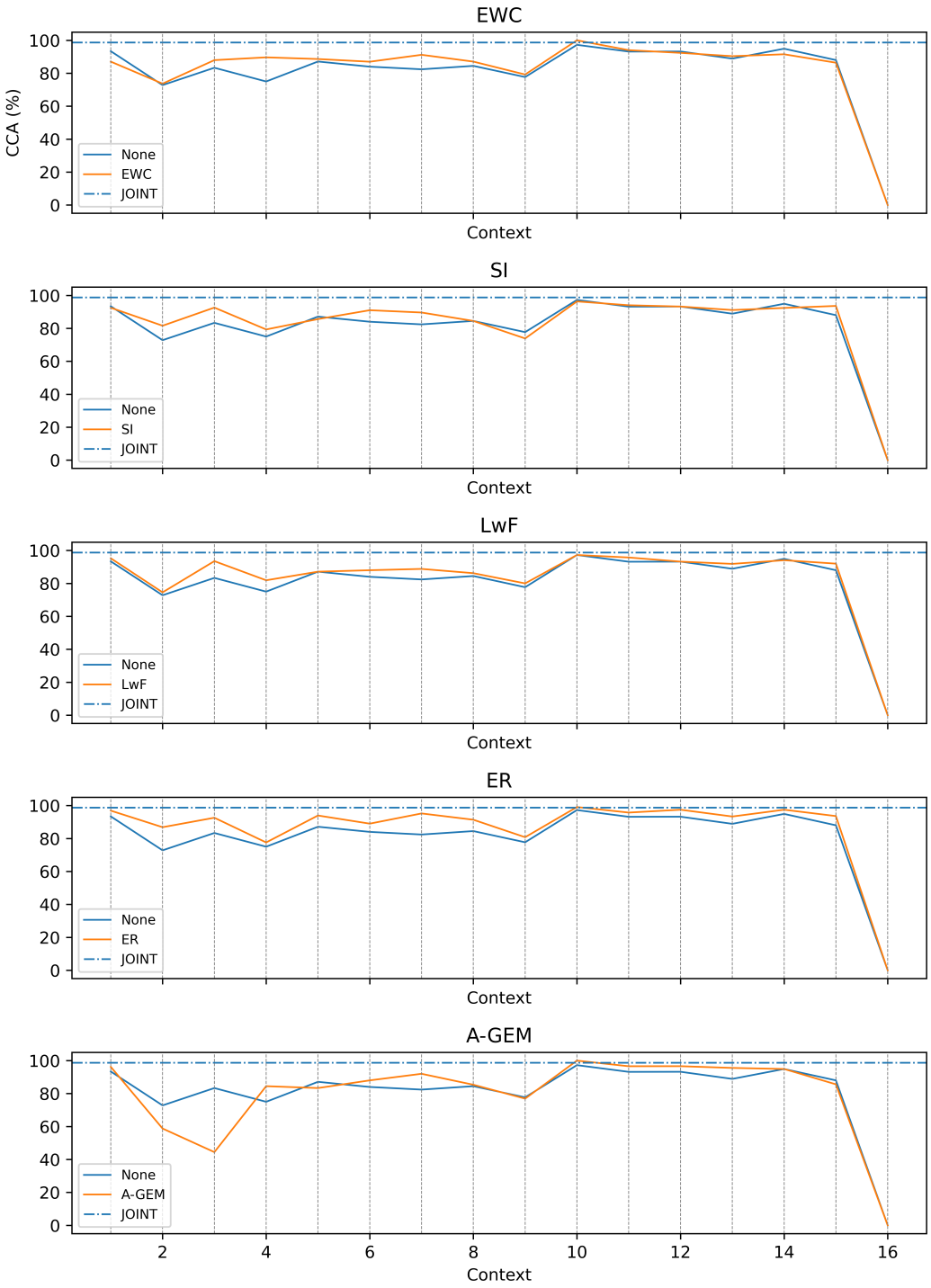

Figure 10: Detailed Current-Context Accuracies (CCAs) Results Figure

B Detailed Seen-Context Accuracies (SCAs) Results Figure

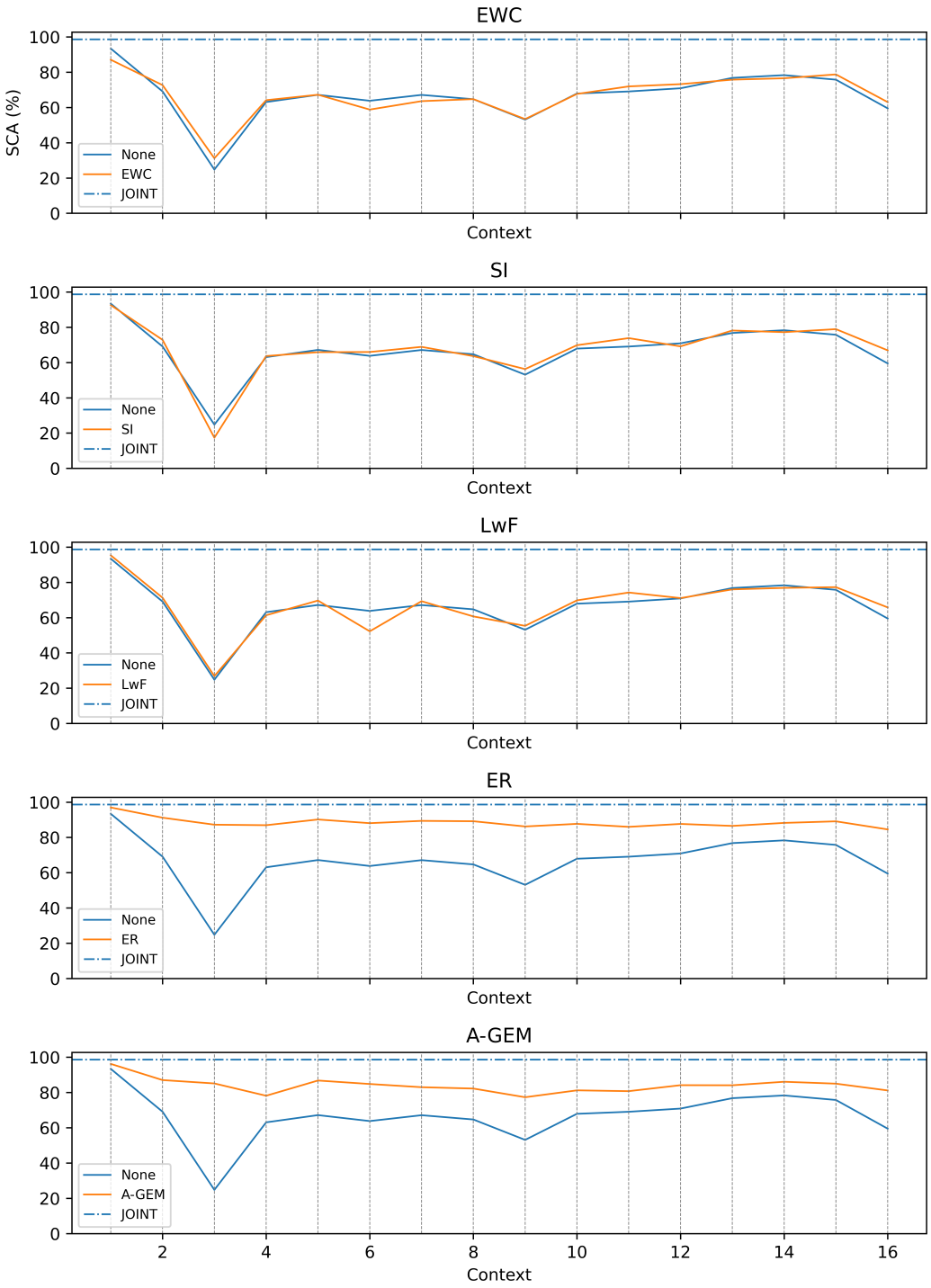

Figure 11: Detailed Seen-Context Accuracies (SCAs) Results Figure

C Hyperparameter Tuning Results

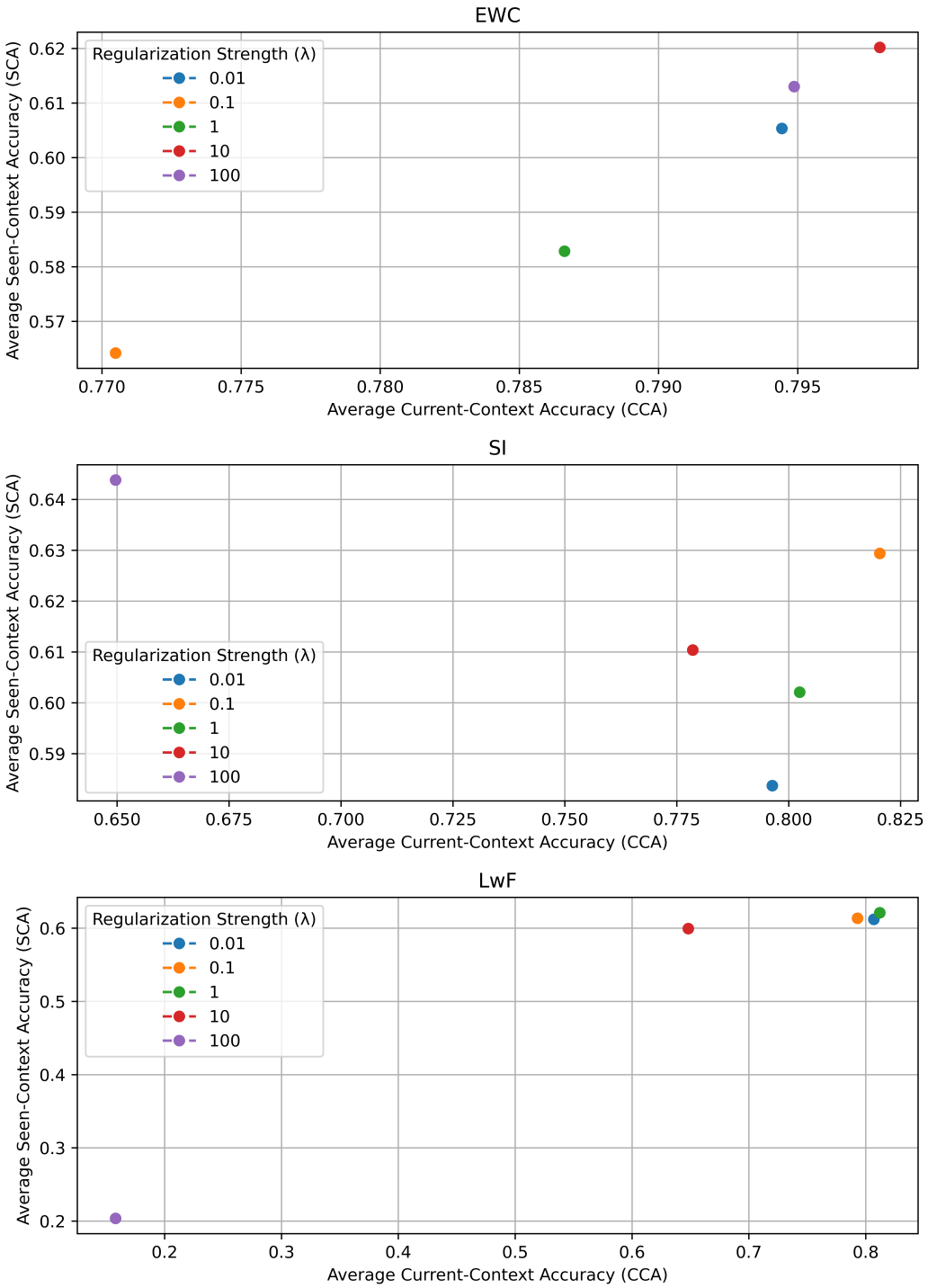

Figure 12: Hyperparameter Tuning Results