Holden Deliverable

D5.3 – DL model for CSI, holography and point cloud specific 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	31/07/2025
Actual date of submission	31/07/2025
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
Ying Liu	AALTO	ying.2.liu@aalto.fi
Yuqing Song	AALTO	yuqing.song@aalto.fi
Stephan Sigg	AALTO	stephan.sigg@aalto.fi

Change history				
Version	**Date**	**Status**	**Partner**	**Description**
1.0	30.5.2025	Final	Aalto	First final draft

Abstract

Contents

1 Introduction

HOLDEN (Holographic Learning for Distributed Ethical Perception) represents a pioneering effort to advance RF-based sensing technologies while embedding privacy-preserving principles at the core of its design. This deliverable focuses on developing a unified deep learning model capable of processing diverse data types, including Channel State Information (CSI), holographic imaging data, and 3D point clouds, to enable ubiquitous perception in applications such as smart living, gesture-based human-computer interaction, and non-invasive logistics. By integrating ethical and privacy constraints, HOLDEN aims to redefine RF-based sensing, ensuring it meets both technical and societal demands. This section provides the context for our approach, reviews existing methodologies, and outlines the motivation for our proposed model, setting the stage for the objectives discussed in the subsequent section.

1.1 Background and Prior Work

In our earlier deliverable, **Holden D5.1: Data structure definition for massive RF data input**, we identified 3D point clouds as a highly effective data structure for RF sensing applications. Point clouds offer several advantages, including computational efficiency, which reduces processing overhead; canonical representation capabilities, which provide a standardized format for diverse RF data; and data volume efficiency, which minimizes storage and transmission requirements. Additionally, point clouds are well-suited for advanced AI/ML techniques, particularly Graph Convolutional Networkss (GCNs), which leverage the inherent relational structure of point cloud data for robust analysis. This deliverable builds on these findings, exploring how point clouds, alongside CSI and holographic data, can be processed to achieve high-accuracy sensing while adhering to ethical constraints.

3D point clouds are particularly valuable in gesture recognition scenarios, where variations in angle, position, and orientation pose significant challenges. When a person performs a gesture in front of a radar, angular deviations relative to the radar plane alter features such as distance, Doppler, and angle, often reducing recognition accuracy. Traditional approaches mitigate this by augmenting datasets with gestures collected from multiple perspectives, but this increases data volume and computational complexity, making it less practical for real-time applications. Point clouds address this issue by preserving the geometric integrity of gesture trajectories, enabling consistent recognition across varying conditions.

1.2 Traditional RF Sensing Methods

Historically, RF sensing data has been processed using conventional signal processing techniques, such as matched filtering, time-frequency analysis, and Kalman filtering. These methods are effective for specific tasks, such as basic target detection or tracking in controlled environments. However, they face significant limitations when applied to high-dimensional, sparse, or noisy RF data, as is common in ubiquitous perception scenarios. For instance, matched filtering struggles to handle complex motion patterns, while time-frequency analysis often fails to capture subtle spatial relationships in sparse datasets. Kalman filtering, while useful for tracking, relies on linear assumptions that do not hold in dynamic, nonlinear environments like gesture recognition or continuous-space sensing. Furthermore, these methods depend heavily on handcrafted features, which limits their adaptability to diverse sensing contexts and reduces their generalization to new scenarios.

1.3 Modern AI/ML Approaches

Recent advancements in AI/ML have introduced more sophisticated models for RF sensing data, particularly for point cloud processing. PointNet [1] and its successor, PointNet++ [2], directly process

point cloud data without requiring conversion to grids or voxels, preserving critical geometric information. PointNet employs local feature extractors and global feature pooling to learn point distributions and structures, while PointNet++ enhances this with hierarchical feature aggregation, capturing both local and global patterns. Similarly, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)-based models, such as the Pantomime system [3], combine PointNet++ with temporal processing to achieve robust gesture recognition, with reported accuracies exceeding $95\,\%$ and mean Area Under the ROC Curve (AUC) values above $99\,\%$ across varied environments.

Despite these advances, non-graph-based AI/ML models have notable drawbacks. PointNet and PointNet++ require significant computational resources to handle variations in data orientation, often necessitating extensive data augmentation. This increases both the data volume and training time, making them less efficient for real-time applications. RNNs and LSTMs, while effective for temporal sequences, struggle to fully exploit the spatial relationships within sparse point clouds, limiting their ability to capture fine-grained geometric features. These limitations highlight the need for a more flexible and efficient approach to handle the complexity of RF sensing data in HOLDEN's diverse use cases.

1.4 Graph Neural Networks

To overcome the limitations of traditional and non-graph-based AI/ML models, GCNs offer a robust and versatile solution. By representing point clouds as graphs, where points are nodes and edges are formed via nearest neighbor searches, GCNs effectively capture both local and global geometric relationships. Advanced variants, such as Dynamic Graph Convolutional Neural Network (DGCNN) [4], dynamically construct adjacency graphs to account for temporal and spatial dynamics, enabling high-resolution feature extraction. For example, the Tesla-Rapture system [5] demonstrated the efficacy of GCNs in gesture recognition, while DGCNN-based approaches [6] achieved state-of-the-art performance on spatiotemporal 3D event clouds by leveraging dynamic graph convolutions. The computational efficiency, adaptability to sparse data, and ability to model complex relationships make GCNs ideally suited for processing CSI, holographic data, and point clouds in HOLDEN's three technical paths: continuous-space sensing, discrete-space sensing, and signal processing.

Graph Neural Networkss (GNNs) are a class of neural networks tailored for processing graph-structured data. Unlike traditional neural networks that assume data lies in Euclidean spaces (e.g., images or sequences), GNNs operate on non-Euclidean structures, making them suitable for applications such as social network analysis, molecular chemistry, and recommendation systems.

1.4.1 Core Concepts

GNNs leverage the topology of a graph $G = (V, E)$, where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. Each node $v \in V$ is associated with a feature vector $\mathbf{x}_v$, and the goal is to learn a node embedding $\mathbf{h}_v$ that encodes both the node's features and its structural context within the graph.

The key mechanism in GNNs is Message Passing Neural Network (MPNN), where nodes exchange information with their neighbors iteratively. This process typically involves three steps:

1. **Message Computation**: For each edge $(u, v) \in E$, a message is computed based on the features of the source node u and, optionally, the edge attributes.

2. **Aggregation**: Each node aggregates messages from its neighbors to form a combined representation.

3. **Update**: The node's representation is updated using the aggregated messages and its previous state.

1.4.2 Mathematical Formulation

The general framework for a GNN layer can be described as follows. Let $\mathbf{h}_v^{(k)}$ denote the embedding of node v at layer k. The update rule for a GNN layer is:

$$\mathbf{h}_v^{(k+1)} = \phi\left(\mathbf{h}_v^{(k)}, \bigoplus_{u \in \mathcal{N}(v)} \psi(\mathbf{h}_u^{(k)}, \mathbf{h}_v^{(k)}, \mathbf{e}_{u,v})\right), \tag{1}$$

where $\mathcal{N}(v)$ is the set of neighbors of node v, ψ is the message function computing information from neighbor u to node v, $\mathbf{e}_{u,v}$ represents edge features (if available), $\bigoplus$ is an aggregation function (e.g., sum, mean, or max), and ϕ is the update function, often parameterized by a neural network.

A popular instantiation of this framework is the graph convolution model, such as the GCN. In a GCN, the update rule simplifies to:

$$\mathbf{h}_v^{(k+1)} = \sigma\left(\sum_{u \in \mathcal{N}(v) \cup \{v\}} \frac{1}{\sqrt{\deg(u)\deg(v)}} \mathbf{h}_u^{(k)} \mathbf{W}^{(k)}\right), \tag{2}$$

where $\deg(v)$ is the degree of node v, $\mathbf{W}^{(k)}$ is a learnable weight matrix for layer k, and σ is a non-linear activation function (e.g., ReLU).

This formulation normalizes the contributions of neighbors based on node degrees, ensuring stable training. Advanced variants of GNNs, such as Graph Attention Networks (GATs) [7], introduce attention mechanisms to weigh neighbor contributions dynamically:

$$\alpha_{u,v} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_u || \mathbf{W}\mathbf{h}_v]))}{\sum_{w \in \mathcal{N}(v)} \exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_w || \mathbf{W}\mathbf{h}_v]))}, \tag{3}$$

where $\alpha_{u,v}$ is the attention coefficient, $\mathbf{a}$ is a learnable vector, and $||$ denotes concatenation. This allows the model to focus on more relevant neighbors.

GNNs, specifically GCNs, are employed to process point cloud data derived from different modalities of RF sensing by representing the data as a graph where points are nodes and their spatial relationships are edges. After detecting walls and excluding regions outside the area of interest to ensure privacy preservation, the point cloud is transformed into a graph structure, with node features capturing RF signal attributes (e.g., amplitude, phase) and edge weights reflecting spatial proximity or signal correlation. GCNs layers aggregate information from neighboring points, enabling the model to learn robust representations of the environment while respecting privacy constraints by omitting sensitive data. This approach is effective for tasks such as object detection or scene reconstruction in applications like indoor mapping or autonomous navigation.

1.5 Privacy-Preserving Sensing

A cornerstone of HOLDEN's mission is to ensure that RF-based sensing adheres to strict ethical and privacy standards. Ubiquitous perception, while enabling transformative applications, raises significant privacy concerns, particularly when RF signals penetrate physical barriers like walls, potentially capturing data from individuals in adjacent spaces. Such unintended data collection risks violating personal privacy, especially in domestic or workplace settings. To address this, our proposed model incorporates privacy-preserving mechanisms that automatically filter out data from beyond designated sensing areas, ensuring that only relevant targets within the intended environment are processed. This privacy-by-design approach aligns with HOLDEN's commitment to ethical sensing and establishes a framework for deploying RF-based technologies in privacy-sensitive contexts, such as smart homes or public spaces.

This deliverable presents a unified deep learning model designed to process CSI, holographic imaging data, and 3D point clouds, supporting HOLDEN's three core technical paths: (1) continuous-space sensing for real-time tracking of human motion and position, (2) discrete-space sensing for fixed-point detection in specific locations, and (3) signal processing and representation learning for complex activity recognition. The model prioritizes privacy and ethical compliance by integrating mechanisms to exclude data from unauthorized areas, enabling applications in smart living, gesture-based interaction, and non-invasive logistics. By building on the strengths of GCNs and addressing the limitations of prior approaches, this model aims to achieve unprecedented accuracy and robustness while respecting societal constraints. The next section outlines the specific objectives of the proposed scheme, detailing its technical and ethical goals and the strategies to achieve them.

2 Objectives

The Deep Learning (DL) model should ideally be designed to address the multifaceted challenges of ubiquitous perception through RF-based sensing, while prioritizing ethical compliance and privacy preservation. By leveraging advanced GNN processing tailored for point cloud data derived from diverse RF modalities, such as radar and WiFi, the model aims to redefine the boundaries of RF-based perception. The objectives outlined below encapsulate the ambitious goals of enabling accurate, privacy-centric, and high-fidelity sensing capabilities that align with the HOLDEN project's vision of a paradigm shift from technology-centric to privacy-centric perception systems.

2.1 Multitarget recognition and tracking

The DL model seeks to achieve robust multitarget recognition and tracking by processing RF signals to identify and monitor multiple subjects or objects within a sensing environment. Unlike traditional RF sensing approaches that often struggle with distinguishing overlapping signals in complex scenarios, the proposed GNN-based framework exploits the structural and relational properties of point cloud data. This enables the model to differentiate between multiple targets based on their spatial and temporal signatures, even in cluttered environments with dynamic motion patterns. By modeling RF data as graphs, where nodes represent measurement points and edges capture spatial or temporal relationships, the model can effectively disentangle overlapping RF reflections, ensuring high precision in tracking trajectories and recognizing distinct entities. This capability is critical for applications such as automated logistics, where precise tracking of multiple objects in real-time is essential, and smart living environments, where accurate identification of human activities enhances personalized services.

2.2 High-dimensional RF-data processing

The DL model is designed to handle the complexity of high-dimensional data derived from CSI, holographic imaging, and point cloud representations generated by RF sensing modalities. CSI data, which captures the amplitude and phase of RF signals across multiple subcarriers, provides rich information about the propagation environment but is inherently high-dimensional and noisy. Similarly, holographic RF data, obtained through wavefront processing, offers detailed 3D spatial information but requires sophisticated algorithms to reconstruct meaningful representations. Point cloud data, derived from these modalities, further complicates processing due to its unstructured and sparse nature. The GNN-based approach addresses these challenges by transforming high-dimensional RF data into graph structures, where nodes represent spatial points or signal features, and edges encode relationships such as proximity or signal correlation. This enables efficient processing of complex data structures, allowing the model

to extract meaningful patterns for ubiquitous perception. For instance, in a smart living scenario, the model can reconstruct a 3D map of a room, identifying furniture, walls, and human occupants, thereby enabling context-aware services.

2.3 Privacy-by-Design via Area-Limited Sensing

A cornerstone of the HOLDEN project is its commitment to ethical and privacy-compliant sensing, and the DL model incorporates a dedicated privacy-preserving module to achieve this goal. RF signals, such as those from WiFi or radar, can penetrate walls, potentially capturing sensitive information about individuals or objects in adjacent spaces, which raises significant privacy concerns. The proposed privacy-by-design module leverages GNN processing to identify structural boundaries, such as walls, within the point cloud data. By analyzing spatial patterns and signal attenuation characteristics, the module isolates the designated sensing area (e.g., a specific room or apartment) and filters out data originating from beyond these boundaries. This process involves segmenting the point cloud into regions based on geometric and signal-based features, ensuring that only data from the intended area is processed further. For example, in a multi-apartment setting, the model can restrict its perception to a single apartment, preventing unintended surveillance of neighboring spaces. This approach not only mitigates privacy risks but also aligns with ethical guidelines by ensuring that sensing is confined to consented areas.

2.4 Unprecedented Accuracy in Activity Recognition

The DL model aims to set a new benchmark in recognizing and classifying complex activities and motions by leveraging high-dimensional tensor processing within the GNN framework. RF-based sensing generates massive-dimensional data that captures subtle variations in motion, such as gestures, gait, or object manipulation. Traditional signal processing methods often struggle to discern these nuances due to the complexity and variability of the data. The proposed GNN-based approach overcomes these limitations by modeling RF data as high-dimensional tensors, where each dimension represents a different aspect of the signal, such as time, frequency, or spatial coordinates. The GNN then processes these tensors to extract hierarchical features, enabling the model to distinguish between intricate activities, such as typing on a keyboard versus waving a hand, or complex motions, such as a person walking while carrying an object. This high-fidelity recognition is achieved through iterative graph convolutions that capture both local and global patterns in the data, ensuring robustness to noise and environmental variations. The model's performance will be validated against application-level benchmarks in scenarios such as logistics (e.g., distinguishing between different types of package handling), smart living (e.g., recognizing specific household activities), and free-space gesture interaction (e.g., detecting precise hand movements for device control).

These objectives collectively advance the HOLDEN project's vision of redefining RF-based perception through a privacy-centric lens. By integrating advanced GNN processing with privacy-preserving mechanisms, the DL model not only achieves unprecedented accuracy in multitarget recognition and activity distinction but also ensures that these capabilities are deployed in an ethically responsible manner. The model's ability to process diverse RF data modalities and filter out privacy-sensitive information positions it as a transformative tool for applications requiring ubiquitous perception while respecting user privacy and ethical boundaries.

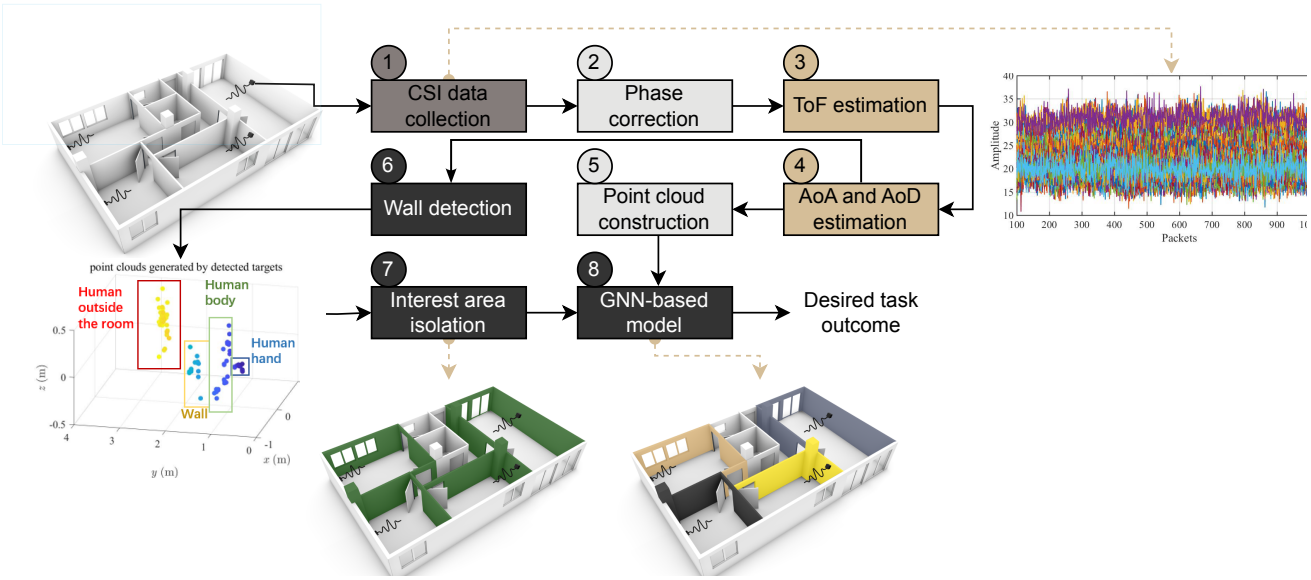


Figure 1: System model of the proposed scheme for privacy-aware processing of wifi-sensing data through GNNs based on CSI. The block shown in dark brown represents the RF sensing medium, the light boxes correspond to pre-processing steps, the brown boxes represent the angle estimation operations, and the black boxes represent the privacy-aware processing steps.

3 Methodology

The methodology for developing the DL model is centered on a GNN-MPNN framework tailored for processing point cloud data derived from diverse RF sensing modalities, such as radar and WiFi, while embedding privacy-by-design principles. The approach begins with the preprocessing of high-dimensional RF data, including CSI, holographic wavefronts, and point clouds, to construct graph representations where nodes encapsulate spatial or signal-based features and edges represent their relational dependencies. A privacy-preserving module is integrated to identify structural boundaries, such as walls, by analyzing signal attenuation and geometric patterns, ensuring that subsequent processing is restricted to data within the designated sensing area. This is followed by the application of GNN-based convolutional layers to extract hierarchical features for multitarget recognition and complex activity classification, leveraging high-dimensional tensor processing to capture intricate spatial and temporal patterns. The methodology is designed to balance unprecedented accuracy in ubiquitous perception with ethical compliance, validated through rigorous testing against application-level benchmarks in logistics, smart living, and free-space gesture interaction scenarios.

3.1 Point Cloud Generation from WiFi CSI Representation

The transformation of CSI data obtained from WiFi signals into a 3D point cloud is a critical step in enabling RF-based ubiquitous perception. CSI data, which captures the amplitude and phase of WiFi signals across multiple subcarriers and antennas, provides rich information about the propagation environment, including reflections, scattering, and attenuation caused by objects and subjects. This subsection outlines a systematic methodology to convert high-dimensional CSI data into a 3D point cloud representation, leveraging signal processing and geometric modeling techniques. The resulting point cloud encapsulates spatial coordinates of scatterers in the environment, facilitating subsequent GNN processing for multitarget recognition and tracking. The overall approach is shown in Fig. 1. The steps are further elaborated in the following part of this section.

3.1.1 CSI data collection

The CSI data is typically collected from a WiFi system with multiple antennas, operating over a set of subcarriers in an Orthogonal Frequency Division Multiplexing (OFDM) framework. For a system with N_t transmit antennas, N_r receive antennas, and K subcarriers, the CSI matrix for a single time instance is represented as a complex-valued tensor $\mathbf{H} \in \mathbb{C}^{N_r \times N_t \times K}$, where each element $H_{i,j,k}$ denotes the channel response between the i-th receive antenna, j-th transmit antenna, and k-th subcarrier. The CSI matrix can be expressed as:

$$H_{i,j,k} = |H_{i,j,k}| e^{j\phi_{i,j,k}}, \tag{4}$$

where $|H_{i,j,k}|$ is the amplitude and $\phi_{i,j,k}$ is the phase of the channel response. The goal is to extract spatial information from $\mathbf{H}$ to construct a 3D point cloud $\mathbf{P} = \{(x_m, y_m, z_m)\}_{m=1}^{M}$, where each point (x_m, y_m, z_m) represents the coordinates of a scatterer in the environment.

3.1.2 Phase correction

Raw CSI data is often noisy due to hardware imperfections, multipath effects, and synchronization errors. To ensure reliable spatial information, preprocessing is applied to correct phase distortions and mitigate noise. The phase $\phi_{i,j,k}$ is affected by Carrier Frequency Offset (CFO) and Sampling Time Offset (STO), which introduce linear phase shifts across subcarriers [8]. The corrected phase $\tilde{\phi}_{i,j,k}$ is obtained by estimating and removing these offsets:

$$\tilde{\phi}_{i,j,k} = \phi_{i,j,k} - \phi_{\text{CFO}} - 2\pi f_k \tau_{\text{STO}}, \tag{5}$$

where ϕ_{CFO} is the phase offset due to CFO, f_k is the frequency of the k-th subcarrier, and τ_{STO} is the time offset. The amplitude is denoised using a low-pass filter to reduce high-frequency noise, yielding a cleaned CSI matrix $\tilde{\mathbf{H}}$.

3.1.3 ToF estimation

To derive spatial coordinates, the CSI data is used to estimate the ToF of signal paths, which corresponds to the distance traveled by the signal from transmitter to scatterer to receiver. The ToF for a given path l can be inferred from the phase differences across subcarriers. The frequency-domain CSI for a single antenna pair can be modeled as a superposition of L paths:

$$H_{i,j}(f_k) = \sum_{l=1}^{L} a_l e^{-j2\pi f_k \tau_l}, \tag{6}$$

where a_l is the complex amplitude of the l-th path, and τ_l is the ToF. The ToF values are estimated using a super-resolution algorithm, such as the MUltiple SIgnal Classification) (MUSIC) algorithm [9], which decomposes the CSI data into a signal subspace to identify dominant path delays. The ToF τ_l is related to the distance d_l via the speed of light c:

$$d_l = \frac{c\tau_l}{2}, \tag{7}$$

where the factor of two accounts for the round-trip distance (transmitter to scatterer to receiver).

3.1.4 AoA and AoD estimation

To construct a 3D point cloud, the spatial location of each scatterer is determined using AoA and AoD in addition to ToF. For a Uniform Linear Array (ULA) with N_r receive antennas, the AoA θ_l for the l-th path is estimated from the phase differences across antennas. The steering vector for the receive array is:

$$\mathbf{a}_r(\theta_l) = [1, e^{-j2\pi \frac{d}{\lambda} \sin(\theta_l)}, \ldots, e^{-j2\pi \frac{(N_r-1)d}{\lambda} \sin(\theta_l)}]^T, \tag{8}$$

where d is the antenna spacing, and λ is the wavelength. Similarly, the AoD ϕ_l is estimated using the transmit array's steering vector. The CSI matrix $\tilde{\mathbf{H}}$ is processed using a 2D MUSIC algorithm to jointly estimate (θ_l, ϕ_l) for each path, providing angular information that complements the distance d_l.

3.1.5 Point cloud construction

Using the estimated ToF, AoA, and AoD, the 3D coordinates of each scatterer are computed. For a scatterer at position (x_m, y_m, z_m), the distance d_l corresponds to the total path length from the transmitter at (x_t, y_t, z_t) to the scatterer and from the scatterer to the receiver at (x_r, y_r, z_r):

$$d_l = \sqrt{(x_m - x_t)^2 + (y_m - y_t)^2 + (z_m - z_t)^2} + \sqrt{(x_m - x_r)^2 + (y_m - y_r)^2 + (z_m - z_r)^2}. \tag{9}$$

The AoA and AoD provide directional constraints:

$$\sin(\theta_l) = \frac{y_m - y_r}{\sqrt{(x_m - x_r)^2 + (y_m - y_r)^2}}, \quad \sin(\phi_l) = \frac{y_m - y_t}{\sqrt{(x_m - x_t)^2 + (y_m - y_t)^2}}. \tag{10}$$

These equations are solved numerically for each path l to obtain the coordinates (x_m, y_m, z_m), forming the point cloud $\mathbf{P}$. To enhance accuracy, multiple transmitter-receiver pairs are used to triangulate scatterer positions, reducing ambiguity in 3D localization.

3.1.6 Wall detection

To align with privacy-by-design principles, range-Doppler images are processed using the YOLOv9 model to identify and filter data outside the designated sensing area. Walls and structural boundaries are detected by analyzing range-Doppler images for static, high-magnitude reflections with consistent spatial patterns, indicative of high material density and geometric continuity. The YOLOv9 model leverages its advanced object detection capabilities to accurately distinguish walls from other targets, such as humans, within these images. Data corresponding to regions beyond detected walls are excluded, ensuring that the processed data only includes information from the intended area, such as a specific room or apartment. This step is crucial for preventing unintended data capture from neighboring spaces.

This methodology ensures that CSI data from WiFi is transformed into a 3D point cloud with high spatial fidelity, enabling subsequent GNN-based processing for ubiquitous perception while adhering to privacy constraints.

3.2 Point Cloud Generation from mmWave FMCW Radars

The generation of a 3D point cloud mmWave Frequency Modulated Continuous Wave (FMCW) radar data is a pivotal process, enabling high-resolution spatial perception while adhering to privacy-by-design principles. mmWave FMCW radars transmit a continuous signal with linearly varying frequency, allowing precise measurement of range, velocity, and angle of detected targets through the analysis of reflected signals. This subsection details a comprehensive methodology to transform raw radar data into a 3D

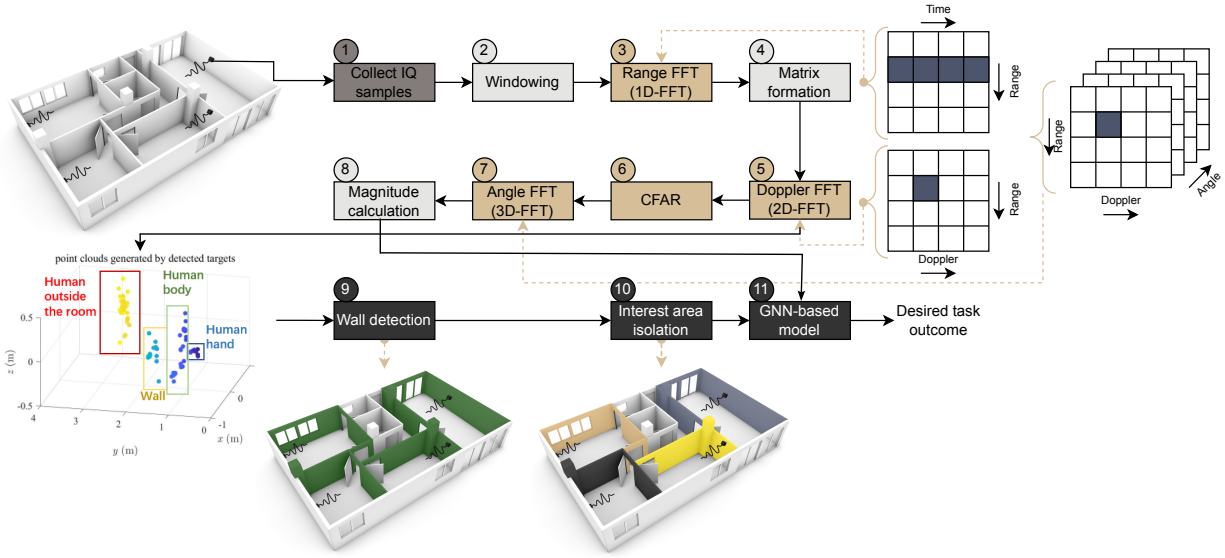


Figure 2: System model of the proposed scheme for privacy-aware processing of RF-sensing data through GNNs.The block shown in dark brown represents the RF sensing medium, the light boxes correspond to pre-processing steps, the brown boxes represent the FFT operations, and the black boxes represent the privacy-aware processing steps.

point cloud $\mathbf{P} = \{(x_m, y_m, z_m)\}_{m=1}^{M}$, where each point (x_m, y_m, z_m) represents the spatial coordinates of a scatterer, such as a human target or object, facilitating subsequent GNN processing for multitarget recognition and tracking [3, 5, 10].

3.2.1 Collection of IQ samples

The process begins with the collection of in-phase (I) and quadrature (Q) samples from the mmWave FMCW radar, which capture the beat frequency resulting from the mixing of transmitted and received signals. For a radar with N_{chirp} chirps and N_{sample} samples per chirp, the raw data forms a complex-valued matrix $\mathbf{S} \in \mathbb{C}^{N_{\text{sample}} \times N_{\text{chirp}}}$, where each element $S_{n,c}$ represents the signal amplitude and phase for the n-th sample in the c-th chirp. The beat frequency f_b is proportional to the range R and is given by:

$$f_b = \frac{2R \cdot B}{T_c \cdot c}, \tag{11}$$

where B is the bandwidth, T_c is the chirp duration, and c is the speed of light.

3.2.2 Windowing

To reduce spectral leakage and improve range resolution, a windowing function (e.g., Hamming window) is applied to the IQ samples. The windowed signal $\mathbf{S}_w$ is computed as:

$$S_{w,n,c} = S_{n,c} \cdot w_n, \tag{12}$$

where w_n is the window function value at the n-th sample. This step enhances the signal-to-noise ratio for subsequent Fourier transforms.

3.2.3 Range FFT (1D-FFT)

A one-dimensional FFT is applied along the sample dimension to convert the windowed signal into the range domain, producing a range profile. The range spectrum $\mathbf{R} \in \mathbb{C}^{N_{\text{sample}}}$ is obtained as:

$$R_k = \sum_{n=0}^{N_{\text{sample}}-1} S_{w,n,c} e^{-j2\pi \frac{kn}{N_{\text{sample}}}}, \tag{13}$$

where k indexes the range bins, and the range resolution is $\Delta R = \frac{c}{2B}$. This step identifies the distances of reflectors from the radar.

3.2.4 Matrix formation

The range profiles across multiple chirps are stacked into a 2D matrix $\mathbf{M} \in \mathbb{C}^{N_{\text{sample}} \times N_{\text{chirp}}}$, where each column corresponds to a chirp's range profile. This matrix serves as the input for further processing to extract angular and Doppler information.

3.2.5 Doppler FFT (2D-FFT)

A second 1D-FFT is applied along the chirp dimension to estimate the Doppler shift, which corresponds to the velocity of moving targets. The Doppler spectrum $\mathbf{D} \in \mathbb{C}^{N_{\text{chirp}}}$ is computed as:

$$D_l = \sum_{c=0}^{N_{\text{chirp}}-1} M_{k,c} e^{-j2\pi \frac{lc}{N_{\text{chirp}}}}, \tag{14}$$

where l indexes the Doppler bins, and the velocity resolution is $\Delta v = \frac{\lambda}{2N_{\text{chirp}}T_{\text{PRI}}}$, with λ being the wavelength and T_{PRI} the pulse repetition interval.

3.2.6 CFAR

CFAR [11] processing is employed to detect significant peaks in the range-Doppler map, filtering out noise and clutter. The detection threshold T is adaptively set based on the local noise level μ and a scaling factor α:

$$T = \mu(1 + \alpha). \tag{15}$$

Peaks exceeding T are selected as candidate targets, providing initial range and Doppler estimates.

3.2.7 Angle FFT (3D-FFT)

For radars with multiple receive antennas, a 1D-FFT is performed across the antenna dimension to estimate the AoA. The steering vector for a uniform linear array with N_r antennas is:

$$\mathbf{a}(\theta) = [1, e^{-j2\pi \frac{d}{\lambda} \sin(\theta)}, \ldots, e^{-j2\pi \frac{(N_r-1)d}{\lambda} \sin(\theta)}]^T, \tag{16}$$

where d is the antenna spacing. The angle spectrum is computed to resolve θ, enhancing the spatial resolution of detected targets

3.2.8 Magnitude calculation

The magnitude of the 3D range-Doppler-angle data is calculated to emphasize the strength of reflections, aiding in target identification. The magnitude $|M_{k,l,\theta}|$ is derived as:

$$|M_{k,l,\theta}| = \sqrt{\text{Re}(M_{k,l,\theta})^2 + \text{Im}(M_{k,l,\theta})^2}, \tag{17}$$

where Re and Im denote the real and imaginary parts, respectively.

3.2.9 Wall detection

Wall detection is performed using the YOLOv9 model, which operates on range-Doppler images derived from range-Doppler-angle data to identify various types of targets, including walls and humans. The YOLOv9 model analyzes these images to detect static, high-magnitude reflections with consistent spatial patterns, characteristic of walls, by leveraging its advanced object detection capabilities. Clusters of points with low Doppler shift and high range attenuation are identified as walls, enabling precise segmentation of the sensing area. Once walls and other targets are identified, a filtering mechanism is applied to isolate and exclude objects located on the opposite side of detected walls, ensuring that only relevant targets within the desired region are considered for further analysis.

This methodology ensures that mmWave FMCW radar data is transformed into a privacy-aware 3D point cloud, supporting the objectives of accurate and ethical RF-based perception discussed in Section 2.

3.3 Data Acquisition and Preprocessing

CSI data is extracted from distributed multi-antenna systems, such as WiFi or mmWave radar, to capture RF channel variations for motion and gesture detection. Each CSI sample comprises complex-valued amplitude and phase information across multiple subcarriers and antenna pairs. For mmWave-based CSI, we utilize a TI MMWCAS radar, with its configuration detailed in Table 1. The raw CSI data is represented as a 3D matrix of dimensions $N_c \times N_{chirp} \times N_{chan}$, corresponding to the number of channels, number of chirps, and number of antenna elements, respectively.

Table 1: Radar parameter configuration

Parameter Name	Symbol	Parameter Setting
Number of transmitting antennas	N_{Tx}	12
Number of receive antennas	N_{Rx}	16
Number of frames acquired	N_f	28
Number of chirps per frame	N_c	128
Number of samples per chirp	N_{adc}	128
FM bandwidth	B	5 GHz
FM slope	K	125 MHz/μs
ADC sampling rate	F_s	4000 Ksps
Frame period	T_f	72 ms

Holographic radar processing reconstructs spatial RF responses into 2D or 3D representations by computing wavefront summations from the antenna array. Specifically, we employ 2D-FFT processing, where the antenna array captures reflected signals across both spatial and temporal dimensions. The received data is transformed using a two-dimensional Fourier transform to generate spatial distribution images, referred to as range-Doppler images, which reveal the positions and movements of targets.

Examples of holographic images for 12 distinct gestures are illustrated in Figs. 3, 4, and 5. These holographic images, which evolve over time, can serve as input to a 2D Convolutional Neural Network (CNN) for classifying specific gestures. However, to reduce computational complexity and enhance privacy, we preprocess the data further to generate point clouds, which are then processed using GNNs for gesture classification, as discussed earlier.

To align with privacy-by-design principles, range-Doppler images are processed using the YOLOv9 model to identify and filter data outside the designated sensing area. Walls and structural boundaries are detected by analyzing range-Doppler images for static, high-magnitude reflections with consistent spatial patterns, indicative of high material density and geometric continuity. The YOLOv9 model leverages its advanced object detection capabilities to accurately distinguish walls from other targets, such as humans, within these images. Data corresponding to regions beyond detected walls are excluded, ensuring that the processed data only includes information from the intended area, such as a specific room or apartment. This step is crucial for preventing unintended data capture from neighboring spaces. Using the mmWave SDK, we apply 3D FFT to estimate range, angle, and Doppler for each target reflection. Detected peaks are converted to (x, y, z) coordinates using the radar's intrinsic parameters and projection matrices. The sequence of frames capturing a subject performing gestures forms a dynamic point cloud trajectory. Each gesture is segmented into fixed-length windows, which are then input into neural networks for classification.

The data undergoes normalization, noise reduction, and anonymization to remove personally identifiable information, ensuring compliance with privacy-by-design principles. These preprocessing steps enhance the robustness and efficiency of the gesture classification pipeline while safeguarding user privacy.

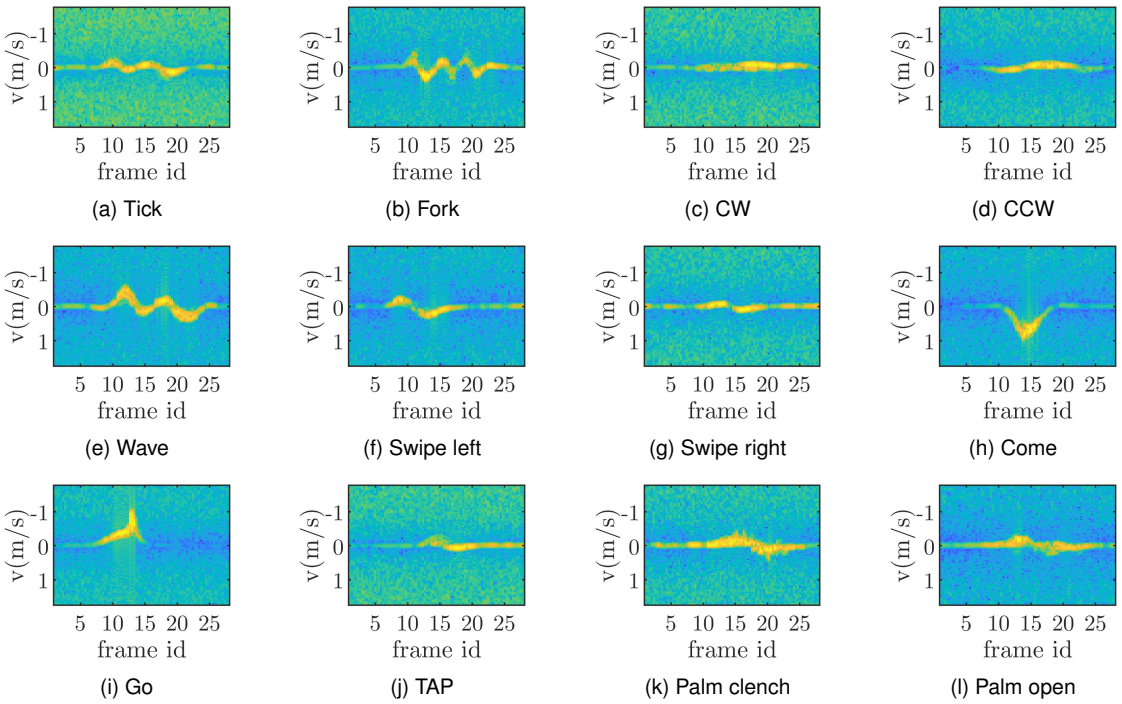


Figure 3: Velocity-time map examples of 12 gestures.

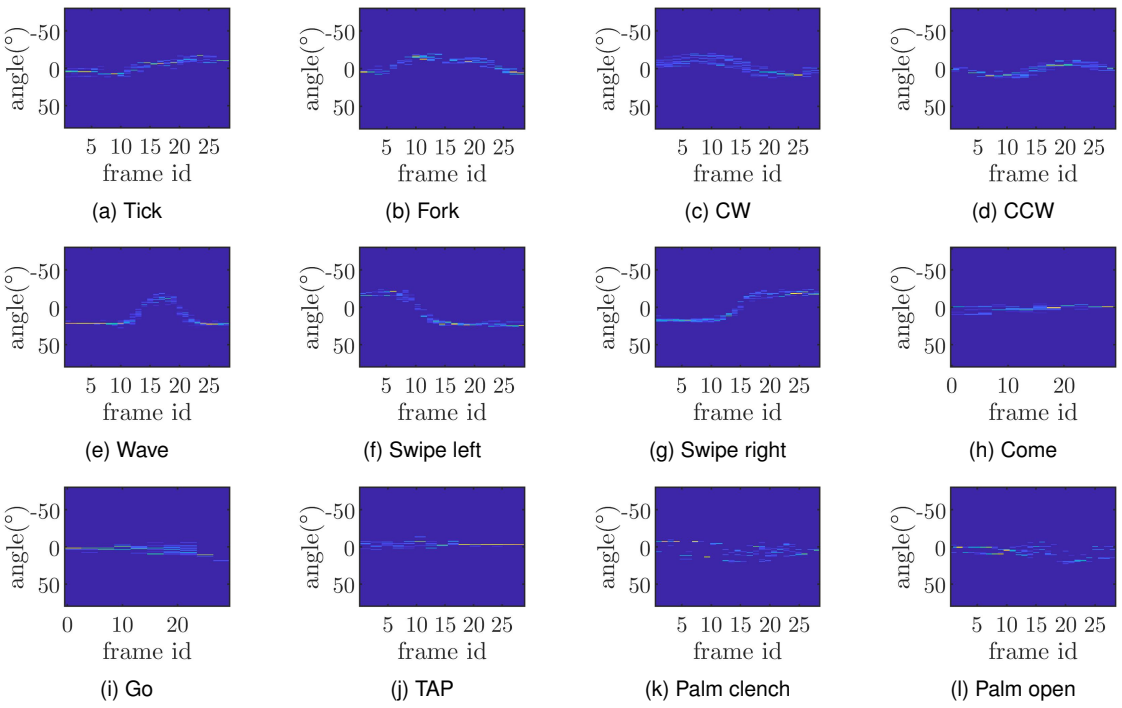


Figure 4: Azimuth angle-time map examples of 12 gestures.

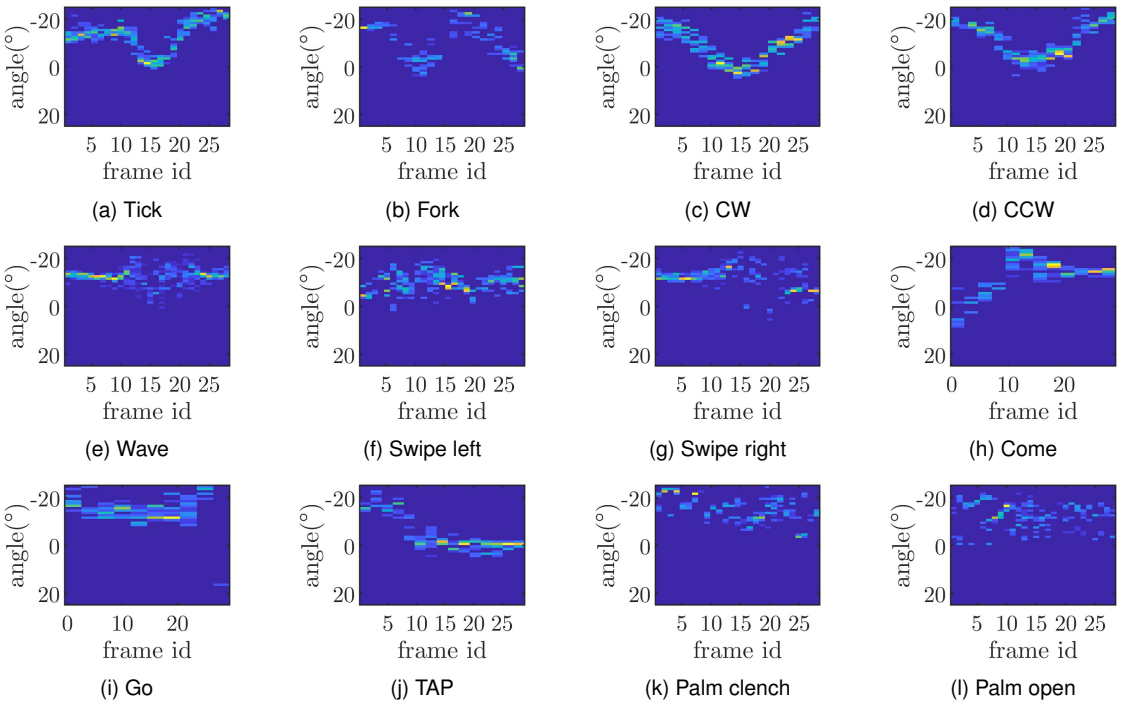


Figure 5: Elevation angle-time map examples of 12 gestures.

3.4 Target Detection Model Architecture

We leverage the YOLO family of algorithms, specifically YOLOv9, for target detection in range-Doppler images derived from mmWave radar or WiFi systems. Unlike traditional CFAR methods, which rely on statistical thresholding for target detection, YOLOv9 employs a deep neural network to simultaneously predict bounding boxes and class probabilities for targets, such as walls and humans, within range-Doppler images. Compared to other deep learning approaches like Region-based Convolutional Neural Networks (R-CNN) [**?**] or Deformable Parts Models (DPM) [12], which separate object localization and classification into distinct steps, YOLOv9 performs end-to-end detection. This is achieved by dividing the input range-Doppler image into an $S \times S$ grid, where each grid cell predicts bounding boxes and associated class probabilities for objects whose center lies within that cell. The output for each grid cell is defined as:

$$O_i = \{(x_j, y_j, w_j, h_j, c_j, p_j^1, \ldots, p_j^C)\}_{j=1}^{B}, \tag{18}$$

where (x_j, y_j) are the bounding box center coordinates, w_j and h_j are the width and height, c_j is the confidence score, p_j^k are the class probabilities for C classes (e.g., wall, human), and B is the number of anchor boxes per grid cell. The confidence score is computed as:

$$c_j = \Pr(\text{Object}) \cdot \text{IoU}_{\text{pred}}^{\text{truth}}, \tag{19}$$

where $\text{IoU}_{\text{pred}}^{\text{truth}}$ is the Intersection over Union between the predicted and ground-truth bounding boxes [13].

The YOLOv9 architecture, built upon YOLOv7 [14] and Dynamic YOLOv7, introduces two key innovations: Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN), as detailed in [**?**]. PGI addresses information loss in deep networks by generating reversible auxiliary supervision signals, allowing gradients to flow back to earlier layers without degradation. GELAN enhances model efficiency by optimizing feature aggregation across multiple scales. The network structure is illustrated in Fig. 6.

The backbone of YOLOv9 is responsible for extracting multi-scale features from the input range-Doppler image, which captures static reflections (e.g., walls with low Doppler shift) and dynamic targets (e.g., humans with varying Doppler shifts). The input image, typically of size $H \times W$, is processed through two initial convolutional (Conv) layers that downsample the input by factors of 2 and 4, respectively, producing feature maps of sizes $H/2 \times W/2$ and $H/4 \times W/4$. These layers apply convolution operations defined as:

$$F_{\text{out}} = \text{Conv}(F_{\text{in}}, W, b) = W * F_{\text{in}} + b, \tag{20}$$

where F_{in} is the input feature map, W is the convolutional kernel, and b is the bias term. The core module, RepN-CSP-ELAN, adopts the GELAN architecture, which combines Cross-Stage Partial (CSP) connections [15] with Efficient Layer Aggregation Networks (ELAN). This module aggregates local and global features through multiple convolutional operations and a branched structure, enhancing the model's ability to detect both static structures (walls) and dynamic targets (humans). The feature extraction process can be expressed as:

$$F_{\text{agg}} = \text{GELAN}(\{F_i\}_{i=1}^{N}) = \sum_{i=1}^{N} \alpha_i \cdot \text{Conv}_i(F_i), \tag{21}$$

where F_i are input feature maps from different layers, α_i are learnable weights, and Conv_i denotes convolutional operations. Subsequent average pooling and convolution layers further downsample the feature maps to capture coarse-grained features.

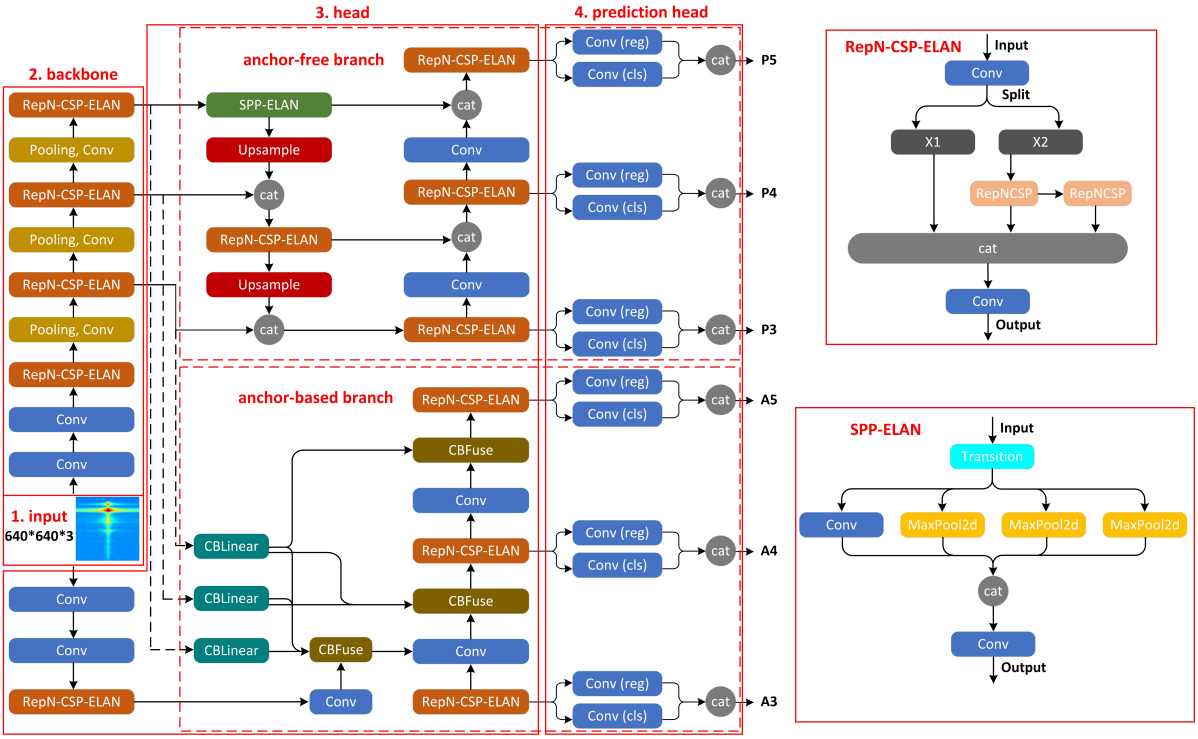


Figure 6: The overall network architecture of YOLOv9

The detection head fuses multi-scale features from the backbone to generate predictions at three scales: P_3 (small objects), P_4 (medium objects), and P_5 (large objects). The head consists of two branches: an anchor-based branch and an anchor-free branch. The anchor-free branch incorporates Spatial Pyramid Pooling (SPP) [16], which pools features at multiple scales to enhance robustness to varying object sizes:

$$F_{\text{SPP}} = \text{Concat}(\{\text{Pool}_k(F_{\text{in}})\}_{k\in\{k_1,k_2,\dots\}}), \tag{22}$$

where Pool_k denotes pooling operations with different kernel sizes. Upsampling and concatenation (Concat) operations merge high-level features with lower-level features, enabling the detection of objects across scales. The anchor-based branch leverages the PGI framework, which includes a multi-level reversible auxiliary branch. The CBLinear module generates auxiliary supervision signals for each layer:

$$F_{\text{aux}} = \text{CBLinear}(F_{\text{in}}) = W_{\text{aux}} \cdot F_{\text{in}} + b_{\text{aux}}, \tag{23}$$

while the CBFuse module integrates these signals back into the main feature maps:

$$F_{\text{fused}} = \text{CBFuse}(F_{\text{main}}, F_{\text{aux}}) = F_{\text{main}} + \beta \cdot F_{\text{aux}}, \tag{24}$$

where β is a learnable scaling factor. This reversible supervision mitigates information loss, improving detection accuracy for static objects like walls, which exhibit high-magnitude reflections in range-Doppler images.

For wall detection, YOLOv9 processes range-Doppler images to identify static, high-magnitude reflections with consistent spatial patterns, indicative of walls. Walls are characterized by low Doppler shift ($v \approx 0$) and high range attenuation, distinguishing them from dynamic targets like humans. The model

outputs bounding boxes and class probabilities, with the loss function combining localization, confidence, and classification losses:

$$L = \lambda_1 L_{\text{loc}} + \lambda_2 L_{\text{conf}} + \lambda_3 L_{\text{cls}}, \tag{25}$$

where L_{loc} is the bounding box regression loss (e.g., CIoU loss [17]), L_{conf} is the binary cross-entropy loss for objectness, and L_{cls} is the classification loss for target types (e.g., wall, human). Non-Maximum Suppression (NMS) is applied as a post-processing step to eliminate redundant bounding boxes, retaining only the most confident detections:

$$\text{NMS}(B) = \{b_i \mid \text{IoU}(b_i, b_j) < \theta, \forall j \neq i, c_i > \tau\}, \tag{26}$$

where θ is the IoU threshold and τ is the confidence threshold. Detected walls are used to segment the sensing area, filtering out data from regions beyond walls to ensure privacy-by-design compliance, preventing unintended data capture from neighboring spaces.

For gesture classification, the detected targets in range-Doppler images are further processed to generate 3D point clouds, as described in the previous section. These point clouds are fed into GNNs [**?**], which model the spatial relationships between points as a graph $G = (V, E)$. The GNN updates node features through message passing:

$$h_v^{(l+1)} = \text{UPDATE}(h_v^{(l)}, \text{AGGREGATE}(\{h_u^{(l)} \mid u \in \mathcal{N}(v)\})), \tag{27}$$

where $h_v^{(l)}$ is the feature vector of node v at layer l, and $\mathcal{N}(v)$ is the set of neighboring nodes. This approach reduces computational complexity compared to 2D CNNs on range-Doppler images and enhances privacy by focusing on anonymized point cloud representations.

3.5 Privacy and Ethical Compliance

To align with HOLDEN's privacy-centric goals, our system incorporates robust mechanisms to safeguard user data and ensure ethical compliance throughout the data processing and model training pipeline. These mechanisms are designed to protect sensitive information, prevent unauthorized data access, and mitigate biases while maintaining the system's performance for target detection and gesture classification using range-Doppler images and point clouds. The following strategies are employed:

3.5.1 Differential Privacy in Model Training:

To prevent data leakage and protect individual privacy, we apply Differential Privacy (DP) to the training of the YOLOv9 model and GNNs. Differential privacy ensures that the model's output is minimally affected by the inclusion or exclusion of any single data point, thereby reducing the risk of reconstructing sensitive information from the training data. Specifically, we add calibrated Gaussian noise to the gradients during training, following the DP-SGD algorithm [**?**]. The update rule for model parameters θ is:

$$\theta_{t+1} = \theta_t - \eta \cdot \text{Clip}(\nabla L(\theta_t, D) + \mathcal{N}(0, \sigma^2 \mathbf{I}), C), \tag{28}$$

where η is the learning rate, $\nabla L(\theta_t, D)$ is the gradient computed on the training data D, $\text{Clip}(\cdot, C)$ clips the gradient norm to a threshold C, and $\mathcal{N}(0, \sigma^2 \mathbf{I})$ is Gaussian noise with variance σ^2. The privacy budget is controlled by the noise scale σ and the number of training iterations, ensuring (ϵ, δ)-differential privacy, where ϵ quantifies the privacy loss and δ is the failure probability [18]. This approach protects sensitive features in range-Doppler images and point clouds, such as unique gesture patterns, from being inferred by adversaries.

3.5.2 Federated Learning for Distributed Data Processing

To minimize centralized data storage and enhance privacy, Federated Learning (FL) can also be employed for distributed processing of CSI data across multiple edge devices (e.g., mmWave radar systems). In this framework, local models are trained on individual devices using locally collected range-Doppler images and point clouds, and only model updates (e.g., weight gradients or parameter differences) are shared with a central server for aggregation. The federated averaging algorithm [19] is used to update the global model:

$$\theta_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_{t,k}, \tag{29}$$

where $\theta_{t,k}$ is the local model parameter for the k-th device, n_k is the number of samples on device k, $n = \sum_k n_k$, and K is the number of participating devices. Secure aggregation protocols, such as those proposed in [20], ensure that individual updates remain private during transmission. By keeping raw CSI data on the edge devices, federated learning reduces the risk of data breaches and complies with privacy regulations, such as General Data Protection Regulation (GDPR), while enabling scalable training across distributed environments.

In conclusion, the integration of differential privacy and federated learning ensures that our system adheres to stringent privacy standards while maintaining high performance in target detection and gesture classification. By applying DP-SGD, we protect sensitive information in range-Doppler images and point clouds, mitigating the risk of data leakage. Federated learning further enhances privacy by decentralizing data processing, keeping raw CSI data on edge devices and sharing only anonymized model updates. These measures collectively uphold HOLDEN's privacy-centric objectives, ensuring that user data remains secure and compliant with ethical and regulatory standards, such as GDPR, while enabling robust and scalable radar-based sensing applications.

4 Implementation

The implementation of our target detection and gesture classification system leverages advanced deep learning frameworks and high-performance hardware to process range-Doppler images derived from mmWave radar systems. This section outlines the development environment, hardware setup, and model training strategies employed to ensure robust and efficient performance. By utilizing the YOLOv9 model for wall and target detection and GNNs for gesture classification, the system achieves high accuracy while adhering to privacy-by-design principles through careful data filtering and preprocessing, as detailed in the following subsections.

4.1 Development Environment

The development and integration of the target detection and gesture classification system were carried out using Python with the PyTorch framework for model training and deployment. PyTorch was chosen for its flexibility in handling dynamic computational graphs and its robust support for deep learning operations, particularly for the YOLOv9 model and GNNs used in processing range-Doppler images and point clouds. This environment ensures efficient model development, testing, and deployment, leveraging PyTorch's optimized libraries for GPU-accelerated computations.

4.2 Hardware Setup

All experiments were conducted on a high-performance workstation running Ubuntu 24.04.2 LTS with Linux kernel version 6.8.0-60-generic. The system was equipped with four NVIDIA A40 GPUs, each providing 46 GB of memory, and utilized CUDA version 12.9 for accelerated deep learning computations. The workstation had approximately 878 GB of available disk storage and sufficient RAM to support large-scale training and evaluation of the YOLOv9 model and subsequent GNN processing. This hardware configuration ensured efficient handling of the computationally intensive tasks associated with processing range-Doppler images and generating point clouds for gesture classification.

4.3 Model Training

The YOLOv9 model was trained with a batch size of 2 over 100 epochs to ensure convergence while managing computational resources effectively. A composite loss function was adopted, combining classification loss, regression loss, and confidence loss to optimize the model's performance in detecting targets (e.g., walls and humans) within range-Doppler images. The classification loss, used to measure the error between predicted category probabilities and true category labels, is computed using binary cross-entropy loss:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\left[y_{i,c}\log(p_{i,c}) + (1-y_{i,c})\log(1-p_{i,c})\right], \tag{30}$$

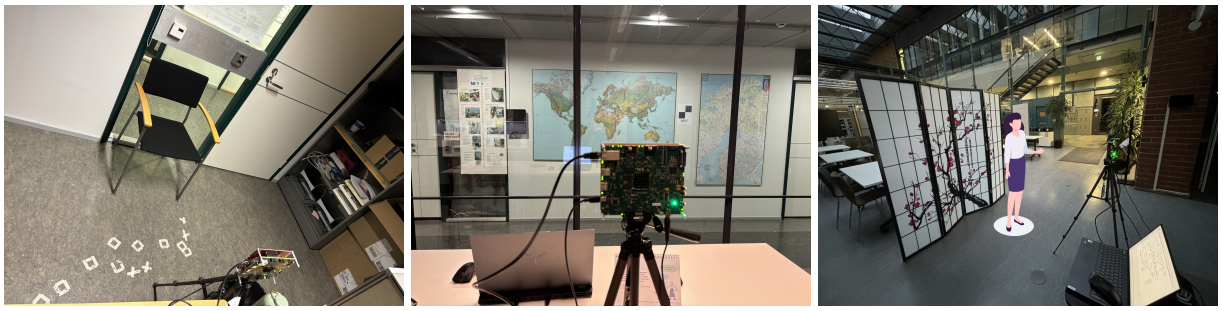
where N represents the total number of bounding boxes, C is the number of categories (including human hand, body, and wall in this work), $y_{i,c}$ is the true label indicating whether the i-th box belongs to category c, and $p_{i,c}$ is the predicted probability for category c. For high-accuracy bounding box prediction, Distribution Focal Loss (DFL) is employed, transforming the regression problem into a distribution learning problem:

$$\mathcal{L}_{DFL} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{D} q_j \log(p_j), \tag{31}$$

where q_j represents true distributional weights (e.g., discrete distribution), p_j denotes predicted distributional probabilities, and D is the number of classification intervals after discretization. DFL enhances prediction accuracy by discretizing the predicted values for each bounding box. Additionally, Complete Intersection over Union (CIoU) loss is used to improve bounding box regression by considering overlap area, centroid distance, and aspect ratio:

$$\mathcal{L}_{CIoU} = 1 - \text{IoU} + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v, \tag{32}$$

where $\text{IoU} = \frac{\text{Intersection Area}}{\text{Union Area}}$ is the intersection-over-union ratio, $\rho(\mathbf{b}, \mathbf{b}^{gt})$ is the Euclidean distance between the centers of the predicted and ground-truth boxes, c is the diagonal length of the bounding box, v is the aspect ratio consistency term, and α is a weighting factor. The CIoU loss enhances regression robustness by improving centroid distance and aspect ratio consistency. The final loss function is a weighted combination of these three losses, ensuring balanced optimization across classification, localization, and confidence prediction for accurate target detection in range-Doppler images.



(a) The office with a concrete wall (b) The coffee room surrounded by a glass wall (c) The open lobby with a Japanese paper screen inside

Figure 7: Different data collection environments.

5 Evaluation

To assess the performance of our gesture recognition system, we conducted a comprehensive evaluation using data collected with Texas Instruments' MMWCAS mmWave radar. The YOLOv9 model, employed for detecting walls and gestures in range-Doppler images, was trained on a curated dataset and evaluated on unseen test data to validate its generalization and robustness across diverse scenarios.

5.1 Datasets

We evaluated the system using a dataset comprising five distinct gestures: "tap," "pinch in," "pinch out," "swipe clockwise (cw)," and "swipe counter-clockwise (ccw)." The "tap" gesture involves a rapid finger tap in the air, "pinch in" entails moving two fingers inward from a separated position to meet, "pinch out" involves spreading two fingers outward from a pinched position, "swipe cw" denotes a clockwise finger motion, and "swipe ccw" indicates a counter-clockwise finger motion. Data collection occurred in three distinct environments to ensure robustness across varied settings, as illustrated in Fig. 7. In the primary setting, an office with a concrete wall (Fig. 7a), we collected 30 samples per gesture, with the human participant positioned 1.5 meters in front of the radar and the wall located behind them. To broaden the applicability, we gathered 10 additional samples per gesture in a coffee room with a glass wall (Fig. 7b) and an open lobby with a paper wall (Fig. 7c). Eight participants (four female, four male, aged 20–30) contributed to the dataset, ensuring diversity in gesture execution. All raw data were processed into range-Doppler images, which were then split into training, validation, and test sets in a 6:3:1 ratio, resulting in 1500 training samples, 750 test samples, and 250 validation samples. The validation set was used to monitor model performance during training and fine-tune hyperparameters, while the test set, unseen during training, evaluated the model's generalization. In all scenarios, a human was present behind the wall, either walking or standing, to simulate realistic conditions and test the system's ability to filter out irrelevant targets using YOLOv9's wall detection capabilities.

5.2 Target Detection Results

After training the YOLOv9 model for 100 epochs, it achieved a precision of **0.953**, a recall of **0.936**, a mean Average Precision at IoU threshold 0.5 (mAP50) of **0.966**, and a mAP50:95 of **0.626** on the test dataset. The mAP50 metric represents the average precision at an Intersection over Union (IoU) threshold of 0.5, while mAP50:95 is computed as the average precision across multiple IoU thresholds from 0.5 to 0.95 in steps of 0.05, providing a comprehensive measure of detection performance. The overall test performance for detecting human hand, human body, and wall targets is detailed in Table 2. Detection

performance across the three distinct scenarios—office (concrete wall), coffee room (glass wall), and open lobby (paper wall)—is presented in the same table in different sections. The results demonstrate consistently high wall detection rates across all scenarios, attributed to the YOLOv9 model's ability to identify static, high-magnitude reflections in range-Doppler images. However, the lobby scenario, with its paper wall, exhibited slightly lower performance, likely due to the material's weaker reflective properties compared to concrete or glass walls.

Table 2: Detection performance of the YOLOv9 model on range-Doppler images across different environments.

Environment	Class	Precision (P)	Recall (R)	mAP50	mAP50-95
Overall	All	0.953	0.936	0.966	0.626
	Human hand	0.953	0.931	0.964	0.665
	Human body	0.992	0.989	0.995	0.636
	Wall	0.913	0.888	0.940	0.577
Office (Concrete Wall)	All	0.958	0.928	0.966	0.628
	Human hand	0.946	0.918	0.958	0.658
	Human body	0.995	0.993	0.995	0.630
	Wall	0.934	0.872	0.945	0.597
Coffee Room (Glass Wall)	All	0.953	0.932	0.963	0.610
	Human hand	0.963	0.921	0.959	0.663
	Human body	0.986	0.988	0.993	0.606
	Wall	0.911	0.887	0.939	0.560
Lobby (Paper Wall)	All	0.949	0.957	0.968	0.628
	Human hand	0.968	0.981	0.988	0.699
	Human body	0.986	0.979	0.988	0.690
	Wall	0.892	0.912	0.927	0.496

Table 3: Detection performance of the YOLOv9 model for humans outside the wall in range-Doppler images across different environments.

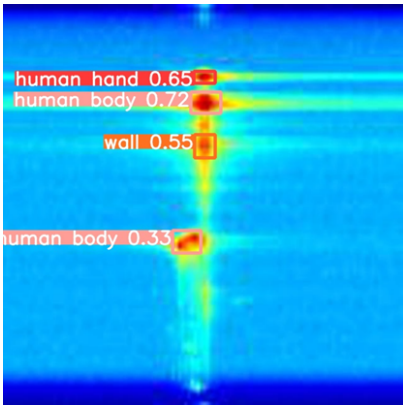
Environment	Precision (P)	Recall (R)	mAP50	mAP50-95
Overall	0.848	0.769	0.843	0.477
Office (Concrete Wall)	0.789	0.644	0.723	0.367
Coffee Room (Glass Wall)	0.782	0.737	0.793	0.420
Lobby (Paper Wall)	0.901	0.909	0.945	0.585

The YOLOv9 model was further trained to detect humans located beyond walls in range-Doppler images, ensuring robust filtering to uphold privacy-by-design principles. The overall detection performance on the test dataset is presented in Table 3. Notably, performance varies significantly across the three evaluated scenarios—office (concrete wall), coffee room (glass wall), and lobby (paper wall)—primarily due to differences in wall thickness and material properties. The paper wall in the lobby, being significantly thinner and less attenuating, results in higher detection rates for humans outside the wall (e.g., precision of 0.901 and recall of 0.909) compared to the thicker concrete and glass walls in the office and coffee room, which exhibit stronger signal attenuation and scattering. This variability underscores the critical role of accurate wall detection using YOLOv9, which leverages static, high-magnitude reflections

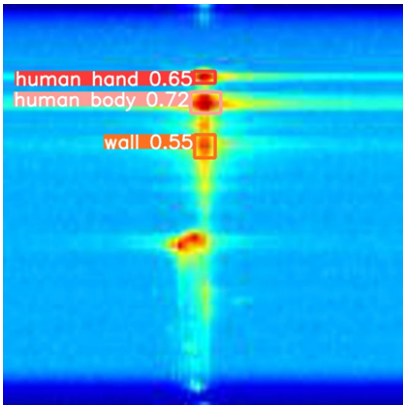
in range-Doppler images to identify walls and filter out irrelevant targets. The high precision and recall for detecting humans outside the wall highlight the necessity of robust wall detection to prevent unintended data capture from neighboring spaces, thereby reinforcing the system's commitment to privacy protection and compliance with regulations such as GDPR.

5.3 Data Filtering (Privacy Preserve) Results

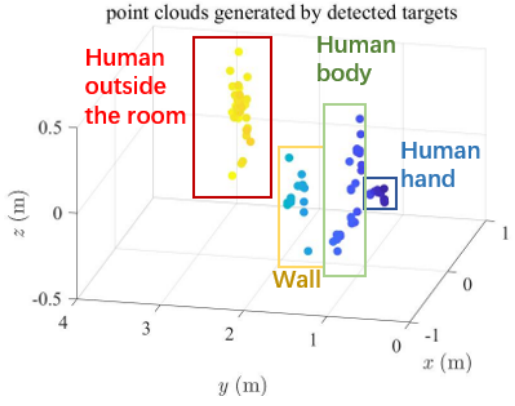
To ensure privacy preservation, we implemented a distance-based filtering mechanism to exclude targets located beyond detected walls, leveraging the YOLOv9 model's capability to identify walls in Range-Doppler Maps (RDMs) based on static, high-magnitude reflections. The detection results for an RDM sample, featuring a person walking outside a wall, are illustrated before and after applying the distance filter in Fig. 8a and Fig. 8b, respectively. The corresponding point clouds, labeled with detected targets, are shown in Fig. 8c (pre-filtering) and Fig. 8d (post-filtering). These figures demonstrate that the human target on the opposite side of the wall is effectively filtered out, ensuring that only data within the intended sensing area is retained. The post-filtered point clouds are dense and accurately represent the structure of in-boundary targets, such as gestures or objects within the designated space, providing high reliability for subsequent tasks like gesture classification using GNNs. This filtering process underscores the system's commitment to privacy-by-design principles, preventing unintended data capture from neighboring spaces and ensuring compliance with privacy regulations such as GDPR.



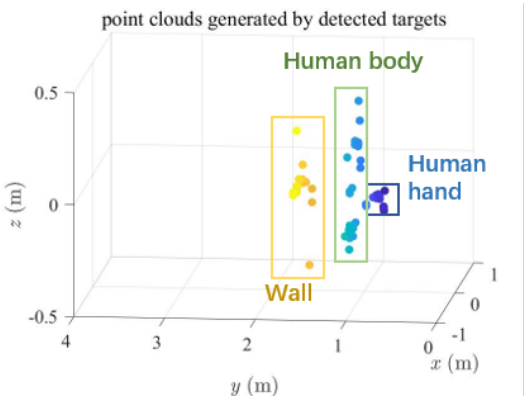
(a) Detection result before distance filter



(b) Detection result after distance filter



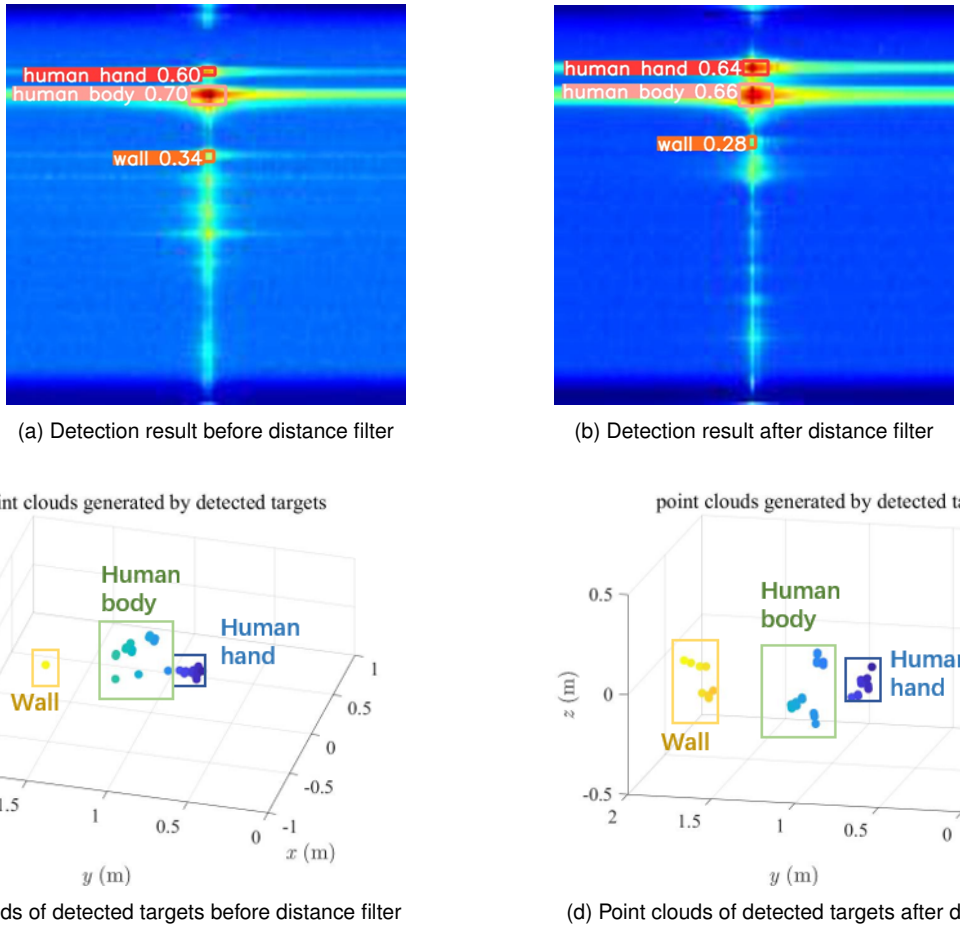
(c) Point clouds of detected targets before distance filter



(d) Point clouds of detected targets after distance filter

Figure 8: An example of detection results in the office.

The detection results for walls in the coffee room and lobby scenarios, using range-Doppler images processed by the YOLOv9 model, are presented in Fig. 9. The results demonstrate precise detection of the glass wall in the coffee room and the paper wall in the lobby, highlighting their viability as effective obstructions in radar-based sensing applications. These walls, despite their differing material properties, successfully block signals from individuals outside the designated sensing area, thereby protecting privacy by preventing unintended data capture from unrelated persons. This capability underscores the potential of using glass and paper walls as alternative barriers in privacy-preserving radar systems, ensuring robust segmentation of the sensing environment while maintaining compliance with privacy regulations.



(a) Detection result before distance filter

(b) Detection result after distance filter

(c) Point clouds of detected targets before distance filter

(d) Point clouds of detected targets after distance filter

Figure 9: Detection results in other scenarios.

6 Applications and Use Cases

The radar-based target detection system, utilizing the YOLOv9 model on range-Doppler images, prioritizes privacy preservation by filtering out objects beyond detected walls, ensuring that only data within the intended sensing area is processed. This privacy-centric approach, combined with the application of GNNs for tasks such as gesture recognition, enables a wide range of applications in secure and ethical environments. Below, we outline key use cases, emphasizing the system's ability to maintain privacy while leveraging GNNs for versatile task performance across diverse settings with concrete, glass, and paper walls.

6.1 Privacy-Preserving Human-Computer Interaction

The system's privacy-preserving mechanism, which uses YOLOv9 to detect walls in range-Doppler images and filter out targets beyond them (e.g., humans outside the sensing area, as shown in Fig. 8b), enables secure human-computer interaction. By applying GNNs to process filtered point clouds, the system supports tasks like gesture recognition, accurately identifying gestures such as "tap," "pinch in," and "swipe clockwise" (Table 2). This is particularly valuable in sensitive environments like medical facilities, where contactless gesture control of devices (e.g., medical imaging systems) ensures hygiene and privacy. The high precision (0.953) and recall (0.936) of YOLOv9, coupled with GNN-based gesture classification, ensure reliable performance while excluding data from neighboring spaces, aligning with privacy regulations such as GDPR.

6.2 Secure Smart Environments

In smart homes and offices, the system leverages its wall detection and filtering capabilities to ensure that only activities within the designated area are processed, protecting user privacy. For instance, in the office scenario with a concrete wall (Table 2), YOLOv9 accurately detects walls (mAP50 of 0.945) and filters out external targets, enabling secure gesture-based control of lighting or appliances. GNNs, applied to the resulting point clouds, support a wide range of tasks, with gesture recognition as a key example for controlling smart devices. The ability to filter out humans beyond glass or paper walls in coffee rooms and lobbies (Table 3) ensures that the system operates ethically, preventing unintended data capture and supporting compliance with privacy standards.

6.3 Privacy-Focused Security Applications

The system's privacy-preserving filtering mechanism is critical for security applications in privacy-sensitive settings, such as retail stores or public lobbies. By using YOLOv9 to identify walls in range-Doppler images and exclude data from beyond them, as demonstrated in Fig. 8d, the system ensures that only authorized activities within the sensing area are monitored. GNNs enhance this capability by processing filtered point clouds for tasks like gesture recognition, enabling secure monitoring of user interactions without compromising external privacy. The high detection accuracy for walls across different materials (e.g., mAP50 of 0.939 for glass walls) ensures robust segmentation, making the system suitable for secure deployments in diverse environments.

6.4 Privacy-Enhanced Assistive Technologies

The system supports assistive technologies by providing a privacy-preserving, contactless interface for individuals with mobility or dexterity impairments. GNNs, applied to filtered point clouds, enable tasks like gesture recognition to control devices such as communication tools or wheelchairs, enhancing accessibility in private residences or care facilities. The system's ability to filter out targets beyond walls, as shown in the lobby scenario with a paper wall (Table 2), ensures that only the user's data is processed, safeguarding personal information. This privacy-centric approach, combined with the versatility of GNNs, makes the system ideal for ethical assistive applications.

In summary, the proposed system excels in privacy-preserving applications by using YOLOv9 to detect walls in range-Doppler images and filter out external targets, ensuring compliance with privacy regulations like GDPR. The integration of Graph Convolutional Networks enables versatile task performance, with gesture recognition serving as a prominent example for human-computer interaction, smart environments, security, and assistive technologies. The system's robust performance across diverse

wall materials and its ability to exclude irrelevant data make it a powerful solution for secure, ethical, and privacy-focused radar-based sensing applications.

7 Conclusion

This deliverable presents a unified deep learning model for processing CSI, holographic imaging data, and 3D point clouds, achieving robust target detection and gesture recognition while prioritizing privacy preservation. By leveraging the YOLOv9 model to detect walls and targets in range-Doppler images and GCNs to classify gestures from filtered point clouds, the system demonstrates high accuracy across diverse environments, as evidenced by a precision of 0.953, recall of 0.936, and mAP50 of 0.966 on the test dataset (Table 2). The privacy-by-design approach, which effectively filters out data beyond detected walls (Fig. 8b), ensures that only authorized data within the designated sensing area is processed, aligning with ethical standards and regulations such as GDPR. The system's ability to handle varied wall materials—concrete, glass, and paper—further underscores its versatility, with robust wall detection performance (e.g., mAP50 of 0.945 in the office scenario) enabling secure applications in smart homes, medical facilities, security, and assistive technologies.

The integration of GCNs enhances the system's capability to process sparse and high-dimensional point cloud data, making it adaptable to a wide range of tasks, with gesture recognition serving as a prominent example. The privacy-preserving mechanisms, including differential privacy in model training and the potential for federated learning, mitigate risks of data leakage and ensure ethical compliance. However, challenges remain, such as the lower detection performance for humans beyond paper walls (Table 3), indicating the need for further optimization in handling less attenuating materials. Future work will focus on enhancing the model's robustness to diverse wall properties, integrating additional RF modalities (e.g., WiFi-based CSI), and deploying the system in real-world settings to validate scalability and user acceptance. By advancing privacy-centric RF-based sensing, this work contributes significantly to the HOLDEN project's vision of ethical, high-fidelity ubiquitous perception, paving the way for transformative applications in human-computer interaction and beyond.

References

[1] 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.

[2] 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.

[3] 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.

[4] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019.

[5] 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.

[6] Junming Chen, Jingjing Meng, Xinchao Wang, and Junsong Yuan. Dynamic graph cnn for event-camera based gesture recognition. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2020.

[7] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[8] Vishnu V Ratnam, Hao Chen, Hao-Hsuan Chang, Abhishek Sehgal, and Jianzhong Zhang. Optimal preprocessing of wifi csi for sensing applications. *IEEE Transactions on Wireless Communications*, 23(9):10820–10833, 2024.

[9] Pooja Gupta and SP Kar. Music and improved music algorithm to estimate direction of arrival. In *2015 International Conference on Communications and Signal Processing (ICCSP)*, pages 0757–0761. IEEE, 2015.

[10] Cesar Iovescu and Sandeep Rao. The fundamentals of millimeter wave radar sensors. *Texas Instruments*, pages 1–7, 2020.

[11] LL Scharf. Detection, estimation and time series analysis. *Statistical Signal Processing*, 1991.

[12] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008.

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[14] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.

[15] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[17] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.

[18] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014.

[19] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[20] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.