This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2026.3713047

mmWave Radar-Based Unsupervised Person ReID via Multi-Level Mutual Signal Contrastive Learning

Qihua Feng, Chunhui Duan, Litian Zhang, Zhiquan Liu, *Senior Member, IEEE*, Feiran Huang, *Member, IEEE*, Jian Weng, *Senior Member, IEEE*, Philip S. Yu, *Life Fellow, IEEE*

***Abstract*—Person Re-IDentification (ReID) is vital for public-safety applications, yet camera-based ReID suffers in low-light environments and poses serious visual privacy risks. In contrast, millimeter-wave (mmWave) radar sensing has gained significant attention in ReID due to its ability to perceive human movement patterns by emitting electromagnetic waves, making ReID unaffected by lighting conditions and capable of privacy preserving. However, current radar-based ReID methods are primarily supervised, and radar data is challenging for human eyes to interpret, making annotation extremely difficult and costly. To this end, we propose an unsupervised radar ReID method based on multi-level cross-view mutual signal contrastive learning. Firstly, to drive unsupervised contrastive learning that depends on data augmentations, we utilize signal processing techniques to transform raw radar signals into frequency spectrum and point clouds, and these forms constitute different and complementary augmented versions of the same sample. Secondly, we construct a multi-level contrastive model between frequency spectrum signals and point clouds through global contrast, cross-view prediction contrast, clustering contrast, and cross-view cluster center contrast, effectively aligning different views and enhancing unsupervised robustness. Additionally, we employ a well-designed local and global graph attention network to learn more discriminative radar point cloud features. Extensive experimental results illustrate that our approach significantly outperforms existing radar-based unsupervised methods. Our code is available at https://github.com/onlinehuazai/mmReID.**

***Index Terms*—Person ReID, mmWave radar, wireless sensing, unsupervised Learning, contrastive learning**

I. Introduction

Person ReID aims to search for specific individuals across different times and locations according to target [1]–[3], whose workflow is illustrated in Fig. 1(a). The system extracts deep features from both query and gallery using the ReID model, then performs matching by computing similarity scores between feature representations to identify the same pedestrian in the gallery. ReID finds extensive applications in intelligent monitoring and security surveillance, such as searching for suspicious individuals. Traditional ReID relies on optical cameras, which face significant obstacles in low-light conditions and raise severe privacy concerns [4]–[8]. In contrast, Radio Frequency (RF) sensing can perceive human motion patterns through electromagnetic wave emissions, unaffected by lighting conditions, and is superior for privacy preservation [6], [7], [9], [10]. Therefore, ongoing research in RF sensing offers a feasible solution for ReID in camera-restricted scenarios. Compared to common low-frequency RF sensing like WiFi, mmWave radar operates at a higher frequency with finer sensing granularity [11], garnering widespread attention for mmWave radar-based ReID [5]–[8], [12]–[14].

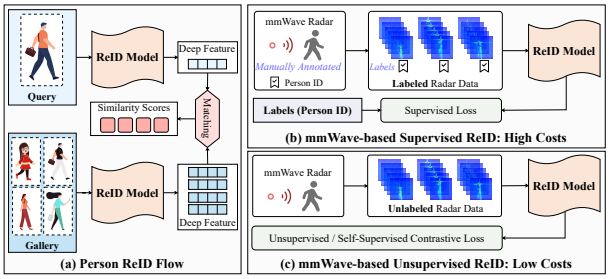


Fig. 1: (a): Person ReID workflow. (b): mmWave-based supervised ReID. (c): mmWave-based unsupervised ReID.

Current radar-based ReID approaches typically extract features from raw signals first and then feed them into supervised Deep Neural Networks (DNNs) to learn relationships between samples, as shown in Fig. 1(b). For instance, the works [5], [7], [15] extract frequency spectrum features from radar signals and learn deep representations of pedestrians through metric and classification losses. The works [6], [13] construct ReID models by extracting Point Clouds (PC) from radar signals and employing supervised point cloud networks. These excellent works [5]–[7], [13], [15] have demonstrated satisfactory supervised results even in scenarios with poor lighting. However, current methods are predominantly supervised, leading to significant labeling costs. Moreover, radar signals, being challenging for human interpretation, further complicate the labeling process.

To alleviate labeling costs, we explore unsupervised ReID using mmWave radar, as shown in Fig. 1(c). Unsupervised image-based ReID has achieved relative maturity [2], [16], [17], with typical approaches employing Self-Supervised Contrastive Learning (SSCL) [18]–[20] to construct unsupervised models, which perform contrastive and clustering losses of different augmented versions of images in high-dimensional feature space. Inspired by unsupervised image-based ReID, we employ SSCL to construct unsupervised radar-based ReID. However, SSCL relies on data augmentation driving [19], [21], whereas structured radar signals cannot undergo direct

Qihua Feng and Chunhui Duan are with Beijing Institute of Technology, Beijing, China.
Litian Zhang is with Beijing University of Posts and Telecommunications, Beijing, China.
Feiran Huang is with Beihang University, China.
Zhiquan Liu and Jian Weng are with Jinan University, Guangzhou, China.
Philip S. Yu is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607 USA.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2026.3713047

data augmentation akin to images [21], [22], such as rotation and color transformations. Complex radar signals contain a plethora of non-informative environmental information and lack rotational invariance and color information, making it easy for unsupervised models to learn useless shortcut information through image-based contrastive learning [21].

In the absence of explicit data augmentation to drive SSCL-based unsupervised radar-based ReID, we propose an implicit augmentation approach through different signal views, which is inspired by the classical unsupervised approach CMVC [23] in the image domain. CMVC [23] treats different modalities or transformations of the same scene as positive sample pairs (e.g., depth maps and RGB images), while considering views from different scenes as negative sample pairs. Through unsupervised contrastive learning, it brings different views of the same scene closer in the feature space while pushing apart views from different scenes, thereby learning view-invariant shared features. Therefore, we can similarly treat different radar data views as distinct data augmentations. Current radar-based ReID methods extract either frequency spectrum [5], [7], [15] or PC [6], [13] from raw data, indicating both representations can characterize pedestrians. We therefore treat the raw signal's frequency spectrum and PC representations as distinct augmented views to facilitate SSCL. Moreover, frequency spectrum and PC complement each other: frequency spectrum is rich in environmental noise but contains all original information; PC data, albeit with lesser noise post-signal processing, suffer from information loss with sparse PC due to radar characteristics [24]. Therefore, we aim to learn an unsupervised radar ReID via mutual signal contrasting between frequency spectrum and PC.

However, directly transferring image-domain methods to our mutual signal contrastive learning is inappropriate. Radar data differs fundamentally from image modalities in terms of data structure, temporal dependency, and semantic gap. Moreover, different radar signal views exhibit substantial disparities in data type, dimensionality, and representation space, making it difficult to perform fine-grained alignment of these signal views in the embedding space and fully exploit the potential of unsupervised learning.

To address the above challenge, we propose a mmWave-based unsupervised ReID using cross-view multi-level mutual signal contrastive learning. The frequency spectrum belongs to the frequency-domain view, containing all environmental and target information; the PC belongs to the spatial coordinate view, containing sparse structural information. Frequency spectrum and PC originate from the same set of raw reflection signals. They are synchronized in time and aligned in space, serving as different augmented versions of the data to drive contrastive learning. For the different signal views, we employ unshared encoders and shared layers to map them into embedding spaces for multi-level contrastive learning at varying granularities. Specifically, we first use global contrast to bring embeddings of different views from the same sample closer together. Second, since mmWave signals are sequential data, we design a cross-view prediction contrast to achieve local alignment between different views. Finally, we perform cluster contrast for each view's embeddings and introduce unsymmetrical cluster center contrast to enable cross-view alignment at the cluster level. Furthermore, leveraging the inherent graph structure of the human body, we design a graph neural network incorporating both local and global attention mechanisms to learn more discriminative PC features. To our knowledge, we are the first to explore mmWave radar-based unsupervised ReID. Experimental results demonstrate that our approach achieves about 60% mean Average Precision (mAP) on actual radar data, significantly exceeding 10% mAP over current unsupervised learning solutions.

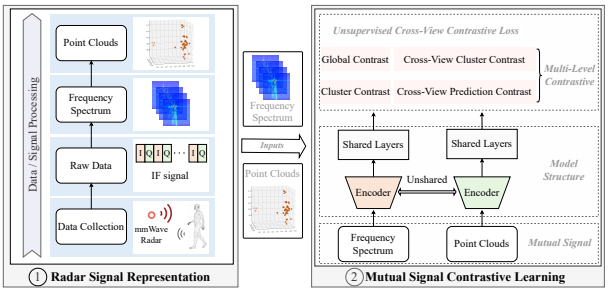


Fig. 2: System overview of our approach.

In summary, our contributions are as follows:

- We introduce a novel mmWave radar-based unsupervised ReID framework, uncovering the relationship between different signal views and unsupervised ReID.
- We treat frequency spectrum and PC as augmented views of the raw data and propose a cross-view multi-level mutual contrasting learning to align different views and enhance unsupervised model generalization.
- The experimental results demonstrate that our approach achieves about 60% mAP and 90% rank-10 accuracy on real radar data, significantly outperforming current unsupervised learning solutions.

The rest of the paper is organized as follows. Section II describes our system overview. Section III describes our approach in detail. Section IV demonstrates and analyzes the experimental results of our approach. Section VI provides the related work. Section V highlights potential future research directions. Finally, Section VII summarizes the paper.

II. System Overview

The system overview of our approach is illustrated in Fig. 2. We employ a commercial TI IWR1843BOOST radar [25] to collect pedestrian data. Our system consists of two cascaded modules: radar signal representation and mutual signal contrastive learning. Initially, the first module processes raw radar signals into two distinct views: the frequency spectrum and the PC data, which are then fed into the second module, and the spectrum and PC data from the same sample constitute the mutual signal pair. Subsequently, we construct an unsupervised ReID model through cross-view contrastive learning between these two signal representations. Our deep model architecture includes separate encoders for each view (non-shared) and shared layers. The unsupervised loss function comprises four contrastive loss components, corresponding to the multi-level contrastive learning strategy.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2026.3713047

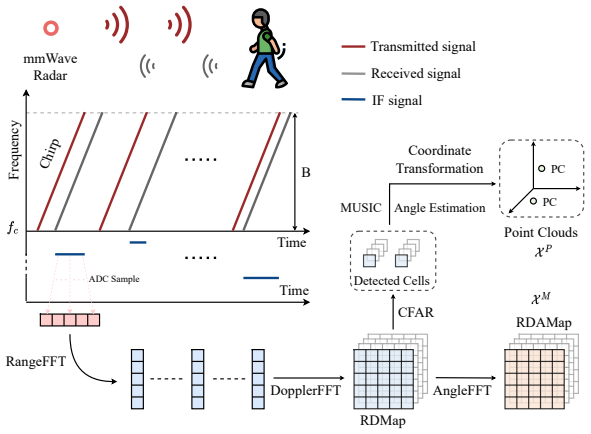


Fig. 3: Our two types of signal representations ($\mathcal{X}^M$ and $\mathcal{X}^P$) and its processes from FMCW mmWave radar.

Radar Signal Representation. Following radar data collection, we obtain raw Intermediate Frequency (IF) signal data. Through data processing and signal denoising techniques, we derive frequency spectrum and extract PC data. The frequency spectrum and PC collectively form two distinct signal representations.

Mutual Signal Contrastive Learning. We feed two signal representations into our ReID model, where they are processed through non-shared encoders and shared layers to obtain deep features in a high-dimensional space. Unsupervised loss is then achieved through cross-view contrastive learning. We propose a multi-level contrastive learning framework to align the two signal representations, consisting of global contrast, cross-view prediction contrast, cluster contrast, and cross-view cluster center contrast.

III. Method

A. Radar Signal Representation

We process the radar signals and obtain two representations: Range-Doppler-Angle Map (RDAMap) and PC, and the sketch is shown in Fig. 3. The mmWave radar periodically transmits Frequency-Modulated Continuous Wave (FMCW) signals and receives reflected signals from surroundings [25]. The transmitted and received signals can be expressed as [26]:

$$S_{Tx}(t_f) = \exp(j(2\pi f_c t_f + \pi \frac{B}{T_f} t_f^2)), \tag{1}$$

$$S_{Rx}(t_f, t_s) = \sum_l \alpha_l S_{Tx}(t_f - \frac{2R_l(t_s)}{c}), \tag{2}$$

where $S_{Tx}(t_f)$ and $S_{Rx}(t_f, t_s)$ denotes transmitted and received FMCW signals, respectively; t_f and t_s represents fast time (Analog-to-Digital Converter (ADC) sample dimension) and slow time (chirp dimension), respectively. f_c, B, T_f, and c are the operating frequency, bandwidth, pulse width, and light speed, respectively; α_l is the reflection coefficient of l-th scatter and R_l is the corresponding radial range. The mixer in the FMCW radar system produces an IF signal by combining received chirps with transmitting chirps, and the IF signal, $s(t_f, t_s)$, is expressed as:

$$\begin{aligned} s(t_f, t_s) &= S_{Tx}(t_f) S^*_{Rx}(t_f, t_s) \\ &= \sum_l \alpha_l \exp(j(2\pi f_c t_f + \pi \frac{B}{T_f} t_f^2)) \cdot (-j(2\pi f_c \\ &\quad (t_f - \frac{2R_l(t_s)}{c}) + \pi \frac{B}{T_f}(t_f - \frac{2R_l(t_s)}{c}))) \\ &= \sum_l \alpha_l \exp(j4\pi \frac{R_l(t_s)}{c}(f_c + \frac{B}{T_f} t_f) - j4\pi \frac{B}{T_f} \frac{R_l^2(t_s)}{c^2}) \\ &\approx \sum_l \alpha_l \exp(j4\pi \frac{R_l(t_s)}{\lambda}) \exp(j4\pi \frac{B}{T_f} \frac{R_l(t_s)}{c} t_f), \end{aligned} \tag{3}$$

where $S^*_{Rx}(t_f, t_s)$ represents conjugate representation of $S_{Rx}(t_f, t_s)$; λ is wavelength of transmitted signal. Since B, λ, T_f are constant, the radial range can be calculated by frequency analysis of $s(t_f, t_s)$ across fast time t_f, such as Fast Fourier Transform (FFT). Applying FFT on $s(t_f, t_s)$, called as RangeFFT, it can be transformed into:

$$\begin{aligned} S(R, t_s) &= \mathcal{F}_{t_f}\{s(t_f, t_s)\} \\ &= \sum_l \alpha_l \delta(f - \frac{2B}{T_f} \frac{R_l(t_s)}{c}) \exp(j4\pi \frac{R_l(t_s)}{\lambda}) \\ &= \sum_l \alpha_l \delta(R - R_l(t_s)) \exp(j4\pi \frac{R_l(t_s)}{\lambda}), \end{aligned} \tag{4}$$

where $R = \frac{cT_f}{2B} f$ is resulting radial range information; $\mathcal{F}_{t_f}\{\cdot\}$ is FFT operator, f is the frequency of FFT, and $\delta(\cdot)$ denotes an impulse-like signal envelope. In addition to range information, the velocity of the target can also be estimated by the Doppler effect. Following Eq. (4), the phase change between the consecutive chirp signals can be represented as:

$$\Delta\phi(t_s) = \frac{4\pi\Delta R}{\lambda} = \frac{4\pi\hat{v}\Delta t_s}{\lambda}, \tag{5}$$

where $\hat{v}$ represents the relative velocity between the mmWave radar and target. Due to the Doppler shift $f_{Doppler} = \frac{2\hat{v}}{\lambda}$, applying FFT on slow time, called DopplerFFT, can reflect the velocity information. After employing RangeFFT and DopplerFFT, we can obtain Range-Doppler Map (RDMap) from FMCW radar. As shown in Fig. 4(a), we give an example of RDAMap. It can be seen that the regions related to human are small, and there are many uninformative backgrounds. Next, we further process the signals on RDMap, resulting in the final signal forms: RDAMap and PC.

RDAMap. mmWave Radar utilizes multiple transmit and receive antennas to form an antenna array, and angle information between the object and radar can be translated into the relationship between the object and antenna array [25]. Based on the different distances from the target to each receiving antenna, slight variations in distance lead to corresponding phase changes, and this phase difference can be used to measure angles. Therefore, by performing FFT on the phase in the antenna dimension, angle information can be obtained, and this operation is denoted as AngleFFT. Since there are both horizontal and vertical antennas, AngleFFT can be applied separately to obtain azimuth and elevation angle information. Performing AngleFFT on RDMap yields one of the final radar signal representations, RDAMap $\{\mathcal{X}_t^M\}_{t=1}^T$, where T is the number of frames. The dimensions of $\mathcal{X}_t^M$ are $2 \times K \times H \times W \times E$, where 2 represents the real and imaginary of the FFT coefficients, and K, H, W, and E are

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2026.3713047

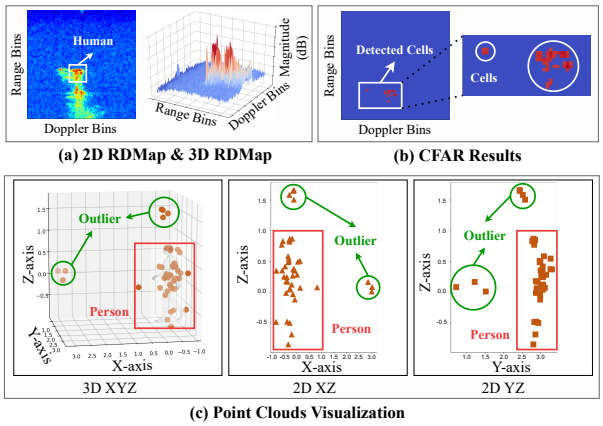


Fig. 4: (a): An example of RDMap. (b): CFAR detection on RDMap. (c): An example of PC visualization.

slow time, fast time, horizontal antenna, and vertical antenna dimensions, respectively.

PC. Unlike cameras, which can sense visible light at megapixel resolution, radar operates at a lower resolution and relies solely on antennas to sense the aggregate signal intensity reflected from the entire environment. Consequently, although RDAMap encapsulates all information, it contains numerous extraneous details unrelated to objects. To identify cells pertinent to the target, Constant False Alarm Rate (CFAR) [27] detection is typically employed on RDMap to retain essential target-related information. As illustrated in Fig. 4(b), we provide an example of CFAR detection on the RDMap, showcasing the retention of only a fraction of the information. Leveraging the distance information of these retained cells, coupled with angle estimation, enables the derivation of specific coordinates of the PC. By utilizing the established angle estimation MUSIC algorithm [28], the azimuth α_A and elevation α_E angles of cells can be estimated. Consequently, the coordinates of PC $\mathcal{C}^P \in \mathbb{R}^3$ can be represented as:

$$\mathcal{C}^P = R \cdot (\sin\alpha_A \cos\alpha_E, \cos\alpha_A \cos\alpha_E, \sin\alpha_E). \tag{6}$$

As depicted in Fig. 4(c), the target cells identified post-CFAR detection are transformed into three-dimensional PC data. However, due to multi-path effects, these signals often carry environmental noise unrelated to human targets. In PC data, such noise typically manifests as isolated outliers, as shown in Fig. 4(c). To address this issue, we employ Statistical Outlier Removal (SOR) [29] to detect and eliminate these anomalous points. To be specific, for each point, it performs a k-Nearest Neighbor (kNN) search and computes the average distance to its neighboring points. If the average distance of a point exceeds a predefined threshold τ^P, the point is classified as an outlier. The threshold τ^P is defined as:

$$\tau^P = \mu^P + \sigma^P, \tag{7}$$

where μ^P and σ^P are the mean and standard deviation of the corresponding average distances of all points, respectively.

We use hyperparameter $\hat{k}$ to control the number of neighbors considered in SOR, and we set $\hat{k} = 5$ in our paper. SOR, as a global statistical method, has limitations in distinguishing dense multi-path reflections that are spatially close to the target. However, denoising methods can only be designed for specific scenarios, and there is almost no single denoising technique capable of completely eliminating all possible types of noise. Furthermore, denoising may also inadvertently remove useful information [24]. Therefore, for PC denoising, we adopt the widely used SOR method. In addition to three-dimensional spatial coordinates of the PC, velocity v^P and energy e^P information corresponding to the PC are integrated into the PC representation. Consequently, the ultimate PC representation is denoted as $\{\mathcal{X}_t^P\}_{t=1}^T$, where $\mathcal{X}_t^P = (\mathcal{C}_t^P, v_t^P, e_t^P)$. The (v_t^P, e_t^P) are state values, denoted as $\mathcal{S}_t^P \in \mathbb{R}^2$, namely $\mathcal{X}_t^P = (\mathcal{C}_t^P, \mathcal{S}_t^P)$.

PC can be obtained from the range-angle domain, but this occurs after signal processing steps such as CFAR denoising [27] and MUSIC angle estimation [28]. These signal processing and denoising techniques, while capable of making the signal data more concise, inevitably lead to the loss of original information [24]. The core function of CFAR is to detect targets in an unknown and time-varying noise/clutter background while maintaining a constant false alarm rate. The Signal-to-Noise Ratio (SNR) is a fundamental metric for determining whether a target is visible against the background [27]. Consequently, only points with sufficiently high SNR are retained, where the energy significantly and consistently exceeds this high threshold. The vast majority of low-SNR responses from weak targets, fluctuating clutter, or noise are rigorously filtered out. Therefore, PC and frequency spectra offer complementary information precisely because PC provides highly refined, task-relevant semantic information, while frequency spectra retain rich, context-related details. By combining the two, our approach captures key focal points without losing the broader context.

B. Mutual Signal Contrastive Learning

We construct an unsupervised ReID model through cross-view mutual signal contrastive learning. The overall model, as shown in Fig. 5, involves two branches to learn different signal representations. Through multi-level contrasts, relationships between different signals are learned and aligned to establish an unsupervised model. Subsequently, we delve into detailed explanations of model structure and loss function separately.

1) Model Structure: Our model comprises two branches, tasked with learning $\mathcal{X}^M$ and $\mathcal{X}^P$ separately. Initially, $\mathcal{X}^M$ and $\mathcal{X}^P$ are fed into their respective encoders ME and PCE, followed by mapping them to the same embedding space using a shared-weight Multi-Layer Perceptrons (MLP) to obtain features z^M and z^P. Subsequently, considering the temporal nature of radar data, z^M and z^P are input into a Bidirectional Long Short-Term Memory [30], [31] (BiLSTM) with attention mechanisms. Finally, a projection head is employed to globally aggregate features and map them to a high-dimensional space. Therefore, the structure of the proposed model is described in detail as follows. First, we introduce encoders ME and PCE. Next, we present the shared modules, including the MLP, attention-based BiLSTM, and projection head.

RDAMap Encoder (ME). The RDAMap $\mathcal{X}_t^M$ is projected into $\hat{f}_t^M$ by using our ME, where $1 \leq t \leq T$. We first perform

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2026.3713047

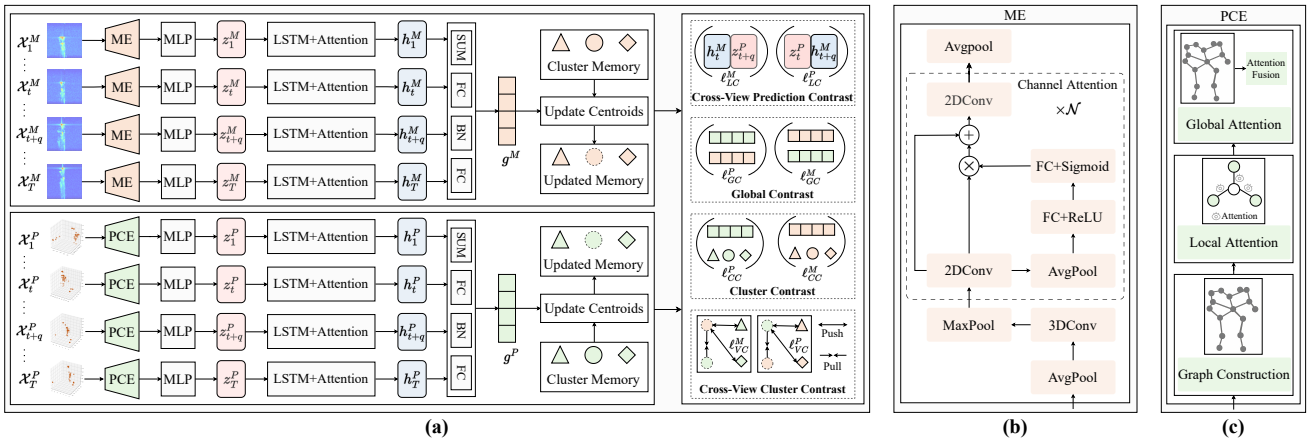


Fig. 5: (a): Overview of our cross-view mutual signal contrastive learning model. (b): ME structure. (c): PCE structure.

average pooling along the vertical antenna dimension, resulting in an intermediate result of dimensions $2 \times K \times H \times W$. Subsequently, in the slow time dimension, we apply a 3D convolution $\frac{K}{2} \times 1 \times 1$ followed by max pooling to obtain the feature map $f_{t,0}^M \in \mathbb{R}^{D \times H \times W}$, where D represents the number of channels. To focus on more crucial channels, we utilize channel attention [32] to learn the importance of different channels and dynamically allocate weights. Our ME has $\mathcal{N}$ channel attention layers, and the n-th channel attention can be expressed as:

$$f_{t,n}^M = \text{2DConv}(f_{t,n-1}^M), \quad n \in [1, \cdots \mathcal{N}], \tag{8}$$

$$\zeta_{t,n}^M = \sigma_2(\text{FC}(\sigma_1(\text{FC}(\text{AvgPool}(f_{t,n}^M))))), \quad n \in [1, \cdots \mathcal{N}], \tag{9}$$

$$\hat{f}_{t,n}^M = \text{2DConv}(f_{t,n}^M + \zeta_{t,n}^M \otimes f_{t,n}^M), \quad n \in [1, \cdots \mathcal{N}], \tag{10}$$

where 2DConv is 2D convolution, FC is fully-connected layer, σ_1 is ReLU [33], σ_2 is Sigmoid activation function, and $\otimes$ is element-wise product. $\zeta_{t,n}^M$ is the learned dynamical attention weights. The results of channel attentions are then subjected to average pooling and FC layer to obtain the output feature $\hat{f}_t^M$, which can be expressed as:

$$\hat{f}_t^M = \text{FC}(\text{AvgPool}(\hat{f}_{t,\mathcal{N}}^M)). \tag{11}$$

To enhance the model's robustness, we apply random masking to $\mathcal{X}^M$. We perform masking on both spatial and temporal dimensions. In the spatial domain, we randomly mask partial information in $\mathcal{X}_t^M$. Given that most of the information in $\mathcal{X}_t^M$ is environmental noise, randomly masking may fail to mask useful information [34]. Therefore, we propose a masking strategy based on intensity levels. Specifically, we select the top μ cells with stronger energy and then randomly mask out $1 \sim \eta_1$ of them, where $1 \leq \eta_1 \leq \mu$. Experimental results indicate that our simple yet effective intensity-based masking strategy outperforms random spatial masking. Additionally, we directly mask $1 \sim \eta_2$ frames in the temporal domain, including masking the consecutive frames and random frames, where $1 \leq \eta_2 \leq T$. The masking probability is set to 0.5 during the training process.

PC Encoder (PCE). There are some advanced networks for PC data in deep learning, such as PointNet [35], PointMLP [36], and PointNet++ [37], and these network architectures are widely employed for learning radar PC data [6], [13], [38]. Although these works can extract features directly from an unordered set of points, they neglect the matching relationships between points [39]. On one hand, our radar PC representation reflects human motion patterns and the human body exhibits graph-like structural relationships [40], [41]. On the other hand, advanced Graph Neural Networks (GNNs) [42]–[44] have been specifically designed to handle graph-structured data like PC [39], [45], which updates vertex representations by aggregating information from neighboring vertices. Therefore, we design a GNN-based PCE encoder to learn more discriminative features from radar PC, incorporating local and global graph attention mechanisms. The structure of the proposed PCE is illustrated in Fig. 5(c), and we input $\mathcal{X}_t^P$ into PCE, converting it into deep features $\hat{f}_t^P$, where $1 \leq t \leq T$. *First*, a graph is constructed from the points. Given the PC set $\{\mathcal{X}_{t,m}^P\}_{m=1}^{\hat{\mathcal{M}}}$ for $\mathcal{X}_t^P$, we construct a graph $\mathcal{G} = (\mathcal{P}, \mathcal{E})$, where $\mathcal{P}$ represents the vertices and $\mathcal{E}$ denotes the edges. Each point $\mathcal{X}_{t,m}^P$ is treated as a vertex, and we employ the kNN algorithm to select neighboring vertices for each point, which are then connected to form edges. As illustrated in Fig. 6(a), an example is given to present graph construction. For each vertex, the nearest $k = 2$ points are selected as neighbors, and edges are established by connecting the vertex to its neighboring vertices. *Next*, we introduce the proposed local-global-based graph attention network to learn the PC graph. Specifically, we utilize local attention mechanisms to dynamically aggregate information from neighboring vertices along edges by learning the importance of each neighbor. Additionally, a global attention mechanism is applied to aggregate information from all vertices, enabling the learning of a comprehensive representation of the entire graph.

The structure of our proposed local-global-based graph attention network is shown in Fig. 6. *First*, we use a MLP block to project $\mathcal{S}_{t,m}^P$ into high-dimension state embedding $\hat{\mathcal{S}}_{t,m}^P$, which can be expressed as:

$$\hat{\mathcal{S}}_{t,m}^P = \text{MLP}(\mathcal{S}_{t,m}^P) = \text{FC}(\sigma_1(\text{FC}(\mathcal{S}_{t,m}^P))). \tag{12}$$

The MLP block contains two FC layers and one activation function. *Second*, we utilize a local graph attention network

to update vertices. Generally, GNN updates vertex state by aggregating neighboring vertex information [39], which can be expressed as:

$$\widetilde{\mathcal{S}}_{t,m}^{P} = \mathcal{F}_1^P\Big(\sum_{(m,y)\in\mathcal{E}} \mathcal{F}_2^P(\widetilde{e}_{m,y}), \hat{\mathcal{S}}_{t,m}^{P}\Big), \tag{13}$$

$$\widetilde{e}_{m,y} = \mathcal{F}_3^P(\hat{\mathcal{S}}_{t,m}^{P}, \hat{\mathcal{S}}_{t,y}^{P}), \tag{14}$$

where $\widetilde{e}_{m,y}$ is edge feature. Function $\mathcal{F}_3^P$ computes the edge feature between two vertices. $\mathcal{F}_2^P$ aggregates the edge features for each vertex. $\mathcal{F}_1^P$ employs the aggregated edges features to update the vertex features. Since PC is sensitive to translation within the neighborhood area, we use the auto-registration mechanism [39] to reduce the translation variance. The auto-registration mechanism aligns neighbors' coordinates by their structural features and uses the center vertex to predict an alignment offset. Therefore, our edge update process is expressed as:

$$\Delta\mathcal{C}_{t,m}^{P} = \mathcal{F}_4^P(\hat{\mathcal{S}}_{t,m}^{P}), \tag{15}$$

$$\widetilde{e}_{m,y} = \mathcal{F}_3^P(\text{concat}(\mathcal{C}_{t,m}^{P} - \mathcal{C}_{t,y}^{P} + \Delta\mathcal{C}_{t,m}^{P}, \hat{\mathcal{S}}_{t,m}^{P}, \hat{\mathcal{S}}_{t,y}^{P})), \tag{16}$$

where $\mathcal{F}_4^P$ calculates the offset using the center vertex state. Radar PC is a complex structure, and different neighborhoods may contribute unequally to the center vertex. Therefore, to dynamically assign the important score of each neighbor node when aggregating edge features, we introduce local attention to learn salient nodes:

$$\alpha_{m,y}^{L} = \frac{\exp(\mathcal{W}_L\widetilde{e}_{m,y})}{\sum_{(m,\cdot)\in\mathcal{E}}\exp(\mathcal{W}_L\widetilde{e}_{m,\cdot})}, \tag{17}$$

$$\widetilde{\mathcal{S}}_{t,m}^{P} = \hat{\mathcal{S}}_{t,m}^{P} + \sum_{(m,y)\in\mathcal{E}} \alpha_{m,y}^{L}\cdot\widetilde{e}_{m,y}, \tag{18}$$

where $\mathcal{W}_L$ is learnable parameters, and $\alpha_{m,y}^{L}$ is learned edge weight. We model function $\mathcal{F}_3^P$ and $\mathcal{F}_4^P$ by using MLP block. The function $\mathcal{F}_1^P$ is residual connection in Eq. (18) and $\mathcal{F}_2^P$ is local attention mechanism in Eq. (17). In Fig. 6(b), we give an example to present the local graph attention to update the 2-nd node. *Third*, we employ a global attention mechanism to fuse all nodes. Specifically, we project the coordinates of each node using positional encoding and combine this information with the node's state information. Subsequently, we learn dynamic weights for each node and fuse them to form the final graph representation. The structure of the global attention mechanism is illustrated in Fig. 6(c), which can be expressed as:

$$\widetilde{\mathcal{X}}_{t,m}^{P} = \text{MLP}(\text{MLP}(\mathcal{C}_{t,m}^{P}) + \widetilde{\mathcal{S}}_{t,m}^{P}), \tag{19}$$

$$\alpha_{m}^{G} = \frac{\exp(\mathcal{W}_G\widetilde{\mathcal{X}}_{t,m}^{P})}{\sum_{m'=1}^{\hat{\mathcal{M}}}\exp(\mathcal{W}_G\widetilde{\mathcal{X}}_{t,m'}^{P})}, \tag{20}$$

$$\hat{f}_t^P = \sum_{m=1}^{\hat{\mathcal{M}}} \alpha_{m}^{G}\cdot\widetilde{\mathcal{X}}_{t,m}^{P}, \tag{21}$$

where $\mathcal{W}_G$ is learnable parameters, and α_{m}^{G} is learned global node weight.

Shared Modules. Due to the disparate structures of $\mathcal{X}^M$ and $\mathcal{X}^P$, their resultant features inhabit distinct feature spaces, disrupting subsequent comparative learning. Consequently, we employ a shared MLP block to map them into a unified space. The MLP block can be mathematically articulated as:

$$z_t = \text{FC}(\sigma_3(\text{FC}(\text{LN}(\hat{f}_t)))), \quad 1 \le t \le T, \tag{22}$$

where LN is layer normalization [46], σ_3 is GeLU [47], $\hat{f}_t = \{\hat{f}_t^M, \hat{f}_t^P\}$, and $z_t = \{z_t^M, z_t^P\}$. Given that radar data comprises multiple frames with temporal dependencies and varying importance across frames, we utilize BiLSTM [30], [31] for temporal modeling and employ attention mechanisms to learn the significance of different frames. This can be expressed as:

$$\hat{h_t} = \text{BiLSTM}(\text{LN}(z_t)), \quad u_t = \tanh(\mathcal{W}_1\hat{h}_t), \tag{23}$$

$$\omega_t = \frac{\exp(u_t^T\mathcal{W}_2)}{\sum_t\exp(u_t^T\mathcal{W}_2)}, \quad h_t = \omega_t\hat{h_t}, \tag{24}$$

where $\mathcal{W}_1$ and $\mathcal{W}_2$ are learnable parameters, ω_t is the learned temporal weight, and $h_t = \{h_t^M, h_t^P\}$. Finally, we employ a projection head for global aggregation of the temporal sequence, which can be expressed as:

$$g = \text{FC}(\text{BN}(\text{FC}((\sum_t h_t)))), \tag{25}$$

where BN is batch normalization [48], and $g = \{g^M, g^P\}$. g^M and g^P are final outputs of the signal representations $\mathcal{X}^M$ and $\mathcal{X}^P$ through our unsupervised model, respectively.

2) Loss Function: Our unsupervised model aligns the two radar representations by multi-level contrastive learning and yields multiple part losses in our model, as shown in Fig. 5. Next, we detail each part of our loss function as follows.

Global Contrast. Given a batch of A radar samples, there are two signal representations for each sample. Therefore,

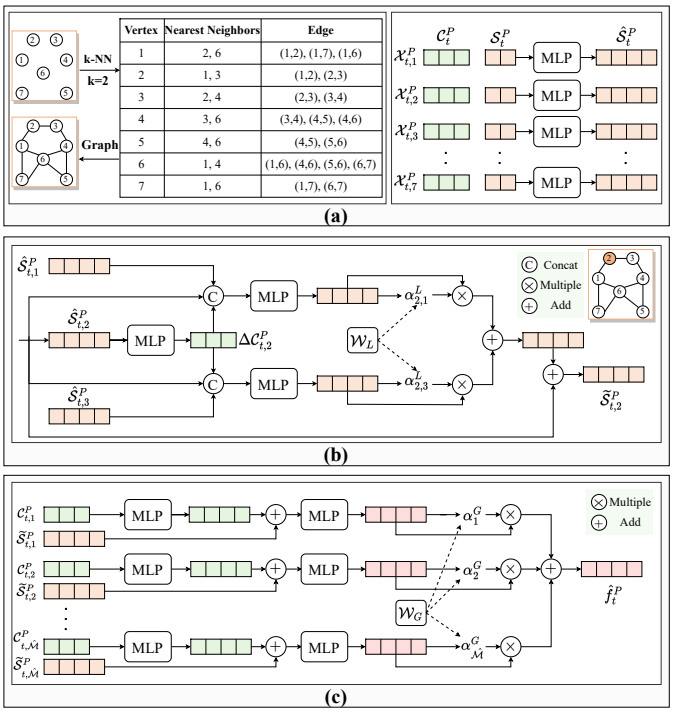


Fig. 6: Overview of our local-global-based graph attention network. (a): An example of graph construction. (b): Local graph attention. (c): Global graph attention structure.

we have $2A$ representations, and the final global features are $\{g_i^M\}_{i=1}^A$ and $\{g_i^P\}_{i=1}^A$. Global contrast aims to bring embedding of different signals from the same sample closer in a high-dimensional space [19]. For a context g_i^M, (g_i^M, g_i^P) are considered positive pairs, and the other $2A-2$ contexts are considered as negative pairs. Therefore, we can employ a global contrasting loss to maximize the similarity between the positive pair and minimize the similarity between negative pairs, which can be defined as:

$$\ell_{GC}^M = -\sum_{i=1}^{A} \log \frac{\exp(g_i^M \cdot g_i^P/\tau_1)}{\sum_{a=1,a\neq i}^{A} \exp(g_i^M \cdot g_a^P/\tau_1) + \exp(g_i^M \cdot g_a^M/\tau_1)}, \quad (26)$$

where τ_1 is the temperature parameter, and we set it to 0.2 in our study. Similarly, For a context g_i^P, symmetric global contrasting loss can be defined as:

$$\ell_{GC}^P = -\sum_{i=1}^{A} \log \frac{\exp(g_i^P \cdot g_i^M/\tau_1)}{\sum_{a=1,a\neq i}^{A} \exp(g_i^P \cdot g_a^M/\tau_1) + \exp(g_i^P \cdot g_a^P/\tau_1)}. \quad (27)$$

Hence, our global contrasting loss ℓ_{GC} can be expressed as:

$$\ell_{GC} = \ell_{GC}^M + \ell_{GC}^P. \quad (28)$$

Cross-View Prediction Contrast. Due to the strong temporal relationships in radar data, we employ cross-view local contrastive prediction [49] to capture and align the temporal features of the samples. Specifically, we use the vector h_t at time t to predict the future timestep vector z_{t+q}, where $1 \leq q \leq Q$, and Q is the hyper-parameter timestep. For instance, for i-th sample within a batch, its vector h_t is denoted as $h_{i,t}$ and the prediction for time $t+q$ is $\mathcal{W}_q h_{i,t}$, where $\mathcal{W}_q$ is learnable parameter. Subsequently, cross-view contrastive learning is employed to minimize the dot product between the predicted values and the true values. For example, $\mathcal{W}_q h_{i,t}^M$ and $z_{i,t+q}^P$ form a positive pair. Simultaneously, for other samples at time $t+q$ in the batch, $\mathcal{W}_q h_{i,t}^M$ forms negative pairs with them, requiring the maximization of their dot products. Accordingly, the prediction contrastive loss of two cross-views can be expressed as:

$$\ell_{LC}^M = -\sum_{i=1}^{A}\sum_{q=1}^{Q} \log \frac{\exp(\mathcal{W}_q h_{i,t}^M \cdot z_{i,t+q}^P)}{\sum_{a=1,a\neq i}^{A} \exp(\mathcal{W}_q h_{i,t}^M \cdot z_{a,t+q}^P)}, \quad (29)$$

$$\ell_{LC}^P = -\sum_{i=1}^{A}\sum_{q=1}^{Q} \log \frac{\exp(\mathcal{W}_q h_{i,t}^P \cdot z_{i,t+q}^M)}{\sum_{a=1,a\neq i}^{A} \exp(\mathcal{W}_q h_{i,t}^P \cdot z_{a,t+q}^M)}. \quad (30)$$

Therefore, our cross-views prediction contrastive loss ℓ_{LC} is expressed as:

$$\ell_{LC} = \ell_{LC}^M + \ell_{LC}^P. \quad (31)$$

Cluster Contrast. To enhance the performance of unsupervised ReID, existing methods often employ clustering to generate pseudo-labels and then enforce high similarity among samples with the same pseudo-label using contrastive learning [2], [16], [50]. Specifically, by utilizing DBSCAN [51] to cluster the embedding g, assuming DBSCAN clusters the samples into V classes, cluster centroids $\{\rho_v\}_{v=1}^V$ can be represented as:

$$\rho_v = \frac{1}{|\mathcal{H}_v|} \sum_{g_i \in \mathcal{H}_v} g_i, \quad (32)$$

where $\mathcal{H}_v$ is the v-th cluster set and $|\cdot|$ indicates the number of samples per cluster. Therefore, the cluster contrasting loss is defined as:

$$\ell'_{CC} = -\sum_{i=1}^{A} \log \frac{\exp(g_i \cdot \rho_+[g_i]/\tau_2)}{\sum_{v=1}^{V} \exp(g_i \cdot \rho_v/\tau_2)}, \quad (33)$$

where τ_2 is the temperature hyper-parameter, and we set it to 0.05 in our study. The $\rho_+[g_i]$ is positive cluster centroids which g_i belongs to. The cluster centroids are stored in a memory bank [2], [16], [50], which are iteratively updated via non-parametric momentum strategy [16], [50], [52]:

$$\rho_v \leftarrow \kappa\rho_v + (1-\kappa)g_{i,v}, \quad (34)$$

where κ is momentum factor and $g_{i,v}$ represents the g_i with the v-th cluster centroid. The DBSCAN runs in each epoch, so V is changing as model training and Eq. (32) is re-executed. We apply cluster contrastive learning to both two branches, $\mathcal{X}^M$ and $\mathcal{X}^P$. Based on Eq. (33), we can calculate the cluster contrastive loss for each branch:

$$\ell_{CC}^M = -\sum_{i=1}^{A} \log \frac{\exp(g_i^M \cdot \rho_+^M[g_i^M]/\tau_2)}{\sum_{v=1}^{V_M} \exp(g_i^M \cdot \rho_v^M/\tau_2)}, \quad (35)$$

$$\ell_{CC}^P = -\sum_{i=1}^{A} \log \frac{\exp(g_i^P \cdot \rho_+^P[g_i^P]/\tau_2)}{\sum_{v=1}^{V_P} \exp(g_i^P \cdot \rho_v^P/\tau_2)}. \quad (36)$$

Therefore, our cluster contrasting loss can be defined as:

$$\ell_{CC} = \ell_{CC}^M + \ell_{CC}^P. \quad (37)$$

We chose DBSCAN clustering rather than K-means, since DBSCAN does not require prior knowledge of the number of pedestrians, whereas the K-Means algorithm requires manually specifying the number of clustering categories [53]. In the scenario of unsupervised ReID, we have no knowledge of how many distinct pedestrian IDs exist in the training set. Arbitrarily setting a value for the number of clustering categories may lead to issues: if the number of clustering categories is too large, the same pedestrian may be split into multiple clusters; if the number of clustering categories is too small, multiple different pedestrians may be merged into a single cluster [53]. In contrast, DBSCAN does not require a predefined number of clusters. It automatically discovers clusters based on the density distribution of the data. Additionally, due to the complexity of the ReID feature space, different samples of the same pedestrian may not form a perfectly spherical cluster in the vector space [53]. K-Means partitions data based on distance to centroids and inherently tends to discover spherical clusters. For non-spherical data distributions, its performance is often suboptimal. DBSCAN, on the other hand, connects sample points through density reachability. As long as sample points are sufficiently close, they can be linked into a cohesive group, enabling the discovery of clusters of arbitrary shapes. This makes DBSCAN better suited to the complex distribution of pedestrian features in real-world scenarios. Therefore, in current unsupervised ReID methods [2], [16], [50], [53], clustering contrast commonly employs DBSCAN clustering.

Cross-View Cluster Center Contrast. Due to the distinct data types of frequency spectrum and PC, their respective cluster centers may exhibit significant divergence in the embedding

space. To address this, we employ cross-view cluster contrast to align cluster centers across different views. Specifically, for cluster centers $\rho_+^M[g_i^M]$ and $\rho_+^P[g_i^P]$, our objective is to minimize the distance between them while maximizing the distance between $\rho_+^M[g_i^M]$ and other cluster centers in $\{\rho_v^P\}_{v=1}^{V_P}$. For a context g_i^M, its corresponding loss is:

$$\ell_{VC}^M = -\sum_{i=1}^{A} \log \frac{\exp(\rho_+^M[g_i^M] \cdot \rho_+^P[g_i^P]/\tau_2)}{\sum_{v=1}^{V_P}(\rho_+^M[g_i^M] \cdot \rho_v^P/\tau_2)}. \tag{38}$$

Similarly, for a context g_i^P, symmetric loss can be defined as:

$$\ell_{VC}^P = -\sum_{i=1}^{A} \log \frac{\exp(\rho_+^P[g_i^P] \cdot \rho_+^M[g_i^M]/\tau_2)}{\sum_{v=1}^{V_M}(\rho_+^P[g_i^P] \cdot \rho_v^M/\tau_2)}. \tag{39}$$

Therefore, our cross-view cluster contrasting loss is:

$$\ell_{VC} = \ell_{VC}^M + \ell_{VC}^P. \tag{40}$$

We intentionally designed asymmetric loss functions l_{VC}^M and l_{VC}^P (with different denominators in the formulas) precisely to flexibly handle the case where $V_M \neq V_P$. Each loss function only requires comparing a cluster center from one view with all cluster centers from the other view, where one is positive and the rest are negative. This design inherently accommodates scenarios where the two views have different cluster set sizes and structures. Additionally, our cross-view cluster center contrast is performed at the sample level rather than the cluster level. For example, for a given sample g_i^M, we only focus on how the cluster center $\rho_+^M[g_i^M]$ it belongs to aligns with the cluster center $\rho_+^P[g_i^P]$ of the corresponding sample g_i^P in the other view. Although the denominator of the loss function includes all cluster centers $\{\rho_v^P\}_{v=1}^{V_P}$ from the other view, this is solely to construct a set of hard negative samples. The core objective remains to bring the distance between $\rho_+^M[g_i^M]$ and $\rho_+^P[g_i^P]$ closer. Since the summation term $\sum_{v=1}^{V_P}$ iterates over all centers in the PC view, regardless of the value of V_M, its denominator can be computed as a single numerical value. Therefore, whether the total number of clusters V_M and V_P in the two views is equal does not affect the construction of this positive sample pair and model training.

In summary, our total loss can be defined as:

$$\ell_{total} = \ell_{GC} + \ell_{CC} + \epsilon_1 \cdot \ell_{LC} + \epsilon_2 \cdot \ell_{VC}, \tag{41}$$

where ϵ_1 and ϵ_2 are hyper-parameters for loss weights.

The core idea of contrastive learning is to pull similar samples closer together and align them in the high-dimensional space while pushing dissimilar samples apart. In unsupervised settings, where labels are unavailable, contrastive learning typically uses data augmentations to generate similar positive samples. In our unsupervised radar ReID task, we design a series of contrastive learning strategies based on the characteristics of ReID and radar data, also following the principle of pulling similar positive samples closer and aligning them in the high-dimensional space. Global contrast treats different views of the same sample as positive pairs and globally pulls them closer in the high-dimensional space for alignment. Cross-view prediction contrast leverages the temporal sequence property of radar data to align features across views by predicting future representations, further pulling the views closer at a local level in the high-dimensional space. Clustering contrast addresses the identity assignment problem in unsupervised scenarios by using pseudo-labels generated from clustering to establish inter-class discrimination, pulling samples with the same pseudo-label closer in the high-dimensional space. Cross-view cluster center contrast enables the model to seek consistency between the cluster center representations of two different views, thereby achieving cross-view alignment. Overall, the four contrastive learning strategies aim to pull positive samples closer and align different view modalities.

The pseudo-code for the training and inference procedures of our approach is presented in Algorithm 1. During the

Algorithm 1: Training & Inference of Our Approach

input : Unlabeled radar training set $\mathcal{X}_{train}$, query $\mathcal{X}_{query}$, gallery $\mathcal{X}_{gallery}$, training epochs $epoch$, encoder ME Θ_m, encoder PCE Θ_p, shared module Θ_s

output: Samples similar to the query in the gallery

```
// ***Training Stage***
// Two Radar Signal Representations
```
$\mathcal{X}_{train}^M \leftarrow$ RDAMap($\mathcal{X}_{train}$);
$\mathcal{X}_{train}^P \leftarrow$ PC($\mathcal{X}_{train}$);
```
// Unsupervised Mutual Signal
   Contrastive Learning
```
for *ep* ← *1 to epoch* **do**
  $\hat{f}^M \leftarrow$ ME($\mathcal{X}_{train}^M$; Θ_m) ;
  $\hat{f}^P \leftarrow$ PCE($\mathcal{X}_{train}^P$; Θ_p) ;
  z^M, h^M, $g^M \leftarrow$ Shared Module($\hat{f}^M$; Θ_s) ;
  z^P, h^P, $g^P \leftarrow$ Shared Module($\hat{f}^P$; Θ_s) ;
  `// Calculate Loss Function`
  $\ell_{GC} \leftarrow$ Global Contrast(g^M, g^P) ;
  $\ell_{LC} \leftarrow$ Prediction Contrast(h^M, z^P; z^M, h^P);
  Update cluster centroids memory by Eq. (34);
  $\ell_{CC} \leftarrow$ Cluster Contrast(g^M; g^P) ;
  $\ell_{VC} \leftarrow$ Cluster Center Contrast(g^M, g^P) ;
  Calculate total loss ℓ_{total} by Eq. (41) ;
  Update model by gradient descent optimizer,
   Θ_m, Θ_p, $\Theta_s \leftarrow \nabla(\ell_{total}, \Theta_m, \Theta_p, \Theta_s)$;
end
```
// ***Inference Stage***
```
if *Using RDAMap* **then**
  $\mathcal{X}_{query}^M/\mathcal{X}_{gallery}^M \leftarrow$ RDAMap($\mathcal{X}_{query}/\mathcal{X}_{gallery}$);
  $g_{query}^M/g_{gallery}^M \leftarrow \Theta_s(\Theta_m(\mathcal{X}_{query}^M/\mathcal{X}_{gallery}^M))$;
  `// Calculate similarity scores`
  scores ← similarity(g_{query}^M, $g_{gallery}^M$) ;
end
if *Using PC* **then**
  $\mathcal{X}_{query}^P/\mathcal{X}_{gallery}^P$, ← PC($\mathcal{X}_{query}/\mathcal{X}_{gallery}$);
  $g_{query}^P/g_{gallery}^P \leftarrow \Theta_s(\Theta_p(\mathcal{X}_{query}^P/\mathcal{X}_{gallery}^P))$;
  `// Calculate similarity scores`
  scores ← similarity(g_{query}^P, $g_{gallery}^P$) ;
end
return Top-K($\mathcal{X}_{gallery}$, scores)

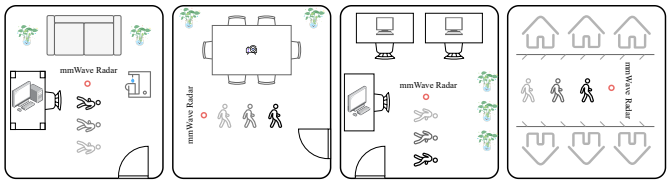


Fig. 7: Four experimental environments.

TABLE I: Our radar dataset description.

	In-Environment		Cross-Environment	
	No. Person	Environments	No. Person	Environment
Training	1∼30	Lab, Meeting Room	1∼30	Lab, Meeting Room
Gallery	31∼40	Lab, Meeting Room	31∼40	Office, Passage
Query	31∼40	Lab, Meeting Room	31∼40	Office, Passage

model training process, model parameters are updated using a gradient descent strategy. In the inference stage, the query is first transformed into an RDAMap or PC, then passed through the corresponding branch to obtain the final depth feature g_{query}. Finally, the cosine similarity scores between g_{query} and the depth features of samples in the gallery are calculated, and samples with the higher scores are returned.

IV. Experiments

A. Dataset and Setting

Dataset. We collected radar data in real-world scenarios. The dataset comprises data from 40 subjects aged between 18 and 50 years, and data collection is obtained with consent from all participants. Data is gathered in four distinct environments: office, meeting room, Lab, and passage, as depicted in Fig. 7. The total dataset exceeds 2TB in storage size, and each participant contributes approximately 100∼200 samples. Our dataset comprises approximately 500,000 frames over a collection period of roughly one month. Following the standard division of ReID datasets [54], we partitioned the dataset into training, gallery, and query sets. The gallery and query sets are used for testing the performance of the ReID model. Additionally, to investigate the model's cross-environment performance, we categorized the dataset into in-environment and cross-environment types, as outlined in Table I. For the in-environment dataset, both the training and test sets are collected from the lab and meeting room. In the cross-environment scenario, the training set is sourced from the lab and meeting room, while the test set is collected from the office and passage. The training set comprises samples from the first 30 individuals, while the gallery and query sets consist of data from the remaining 10 individuals.

Experimental Setting. We use the TI IWR1843BOOST radar whose parameters are configured as outlined in Table II. The radar operates within a frequency range of 77GHz to 81GHz with periodic variations, featuring three transmit antennas and four receive antennas. The frequency slope is set to 60.012, with an ADC sampling dimension of 256, a chirp dimension of 128, and an acquisition of 100 frames per data collection. We develop a deep learning model using the PyTorch framework [55], leveraging a V100 machine. The batch size is set to 128, with training conducted for 256 epochs. We utilize the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9. The learning rate is set to 0.01 with a cosine warm-up strategy [20], [56], [57].

TABLE II: mmWave Radar parameters in our experiments.

Parameters	Transmitters	Receivers	Chirp Duration	Chirps	ADC Samples	Frames
Values	3	4	72 us	128	256	100

TABLE III: Overall performance comparison with baselines.

Scheme	In-Environment				Cross-Environment			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Unsupervised Learning								
TGUL [21]	36.2	48.3	60.2	71.4	33.1	43.7	56.7	67.5
PRISM [34]	42.5	53.1	61.6	75.7	38.2	50.9	58.9	70.6
RF-URL [58]	48.9	59.7	65.7	78.5	44.6	55.1	63.4	73.8
Ours	**59.3**	**71.8**	**80.4**	**90.6**	**55.6**	**69.2**	**78.5**	**85.7**
Supervised Learning								
MCGait [15]	68.7	80.6	90.6	96.3	66.2	81.8	89.4	95.2
mARPC [12]	70.5	83.1	91.7	97.2	68.3	82.9	90.2	95.9
mPIAN [13]	69.3	81.2	89.5	96.5	67.1	82.2	89.9	95.8
mmReID [7]	71.3	83.8	92.6	97.8	69.7	82.5	90.6	96.2
RF-ReID [5]	72.2	85.1	93.2	98.5	70.6	83.8	91.1	96.7
Ours	**75.7**	**87.6**	**94.3**	**98.9**	**73.1**	**84.9**	**92.2**	**97.3**

B. Baselines

We compare with existing mmWave radar-based ReID works. Since unsupervised ReID methods remain unexplored in mmWave sensing, we employ several unsupervised learning approaches from wireless sensing as comparative baselines, such as unsupervised pre-training methods TGUL [21], RF-URL [58], and PRISM [34]. Furthermore, we can fine-tune our unsupervised models with labels to form supervised learning, which employs the advanced supervised ReID scheme [59] on our datasets. Therefore, we include supervised radar-based ReID methods as baselines, such as mmReID [7], mARPC [12], MCGait [15], mPIAN [13], and RF-ReID [5]. Several supervised ReID methods [6], [14] utilize multi-modal training strategies, such as incorporating additional image samples. However, since our dataset contains only unimodal radar data, we compare exclusively with unimodal supervised approaches to ensure fair evaluation.

In evaluation, we use the common mAP and Cumulative Matching Characteristic (CMC) at Rank-1, 5, and 10 to evaluate unsupervised ReID performance. The greater these values, the better the performance. All schemes are evaluated in the same testing set.

C. Overall Performance

The comparison results with baselines are shown in Table III. It is evident from Table III that our approach outperforms both unsupervised and supervised approaches in terms of performance, validating the efficacy of our method.

In-Environment Performance. Compared to current radar-based unsupervised learning methods [21], [34], [58], our performance significantly surpasses them, such as exceeding

TGRU [21] by about 23% and RF-URL [58] by about 11% in mAP. Our method achieves 71.8% Rank-1 accuracy, 80.4% Rank-5 accuracy, and 90.6% Rank-10 accuracy, significantly outperforming existing unsupervised approaches. Presently, wireless-based methods have not specialized in unsupervised ReID and primarily focus on building sensing models through unsupervised representation learning. For instance, PRISM [34] adopts a masked autoencoder approach for unsupervised pre-training. However, our method leverages multi-level cross-view contrastive learning and a well-designed radar ReID model to comprehensively explore relationships between samples, resulting in superior unsupervised person ReID performance. For example, while RF-URL [58] employs global contrastive learning for unsupervised representation, our approach not only incorporates global contrastive learning but also designs clustering-based contrastive learning tailored for unsupervised ReID. Clustering contrast facilitates pseudo-label learning, which is highly suitable for unsupervised ReID, enabling the model to learn more discriminative information. Similarly, TGRU [21] and PRISM [34] leverage unsupervised autoencoder architectures to learn pre-trained models, yet they do not fully exploit sample information specifically for unsupervised ReID. In contrast, our multi-level contrastive design fundamentally aims to pull representations of the same pedestrian closer in high-dimensional space, enhancing the generalization performance of unsupervised ReID through our multi-level contrastive learning.

Cross-Environment Performance. As shown in Table III, our model achieves superior performance compared to other baselines under cross-environment conditions. For instance, our method achieves 55.6% unsupervised mAP and 73.1% supervised mAP, surpassing unsupervised baselines by over 10% mAP and supervised baselines by more than 2% mAP. Due to the inherent characteristics of radar electromagnetic signals, mmWave sensing demonstrates significant environmental sensitivity. Consequently, our cross-environment ReID performance exhibits a 2%$\sim$4% mAP degradation relative to in-environment scenarios, though this remains within acceptable limits. Performance degradation in cross-domain scenarios is common in mmWave sensing, and our unsupervised learning paradigm mitigates environmental interference by learning category-agnostic general representations [58], maintaining effectiveness under cross-environment conditions. In future work, we will explore additional techniques to enhance cross-domain performance for mmWave-based unsupervised ReID and reduce environmental impacts. Compared to current radar-based supervised ReID methods [5], [7], [12], [13], [15], our approach respectively achieves 75.7% mAP and 73.1% mAP under the in-environment and cross-environment settings, significantly outperforming them by about $3 \sim 7\%$ mAP. These supervised approaches [5], [7], [12], [13], [15] train models from scratch without utilizing pre-trained models. In contrast, our supervised scheme uses a trained unsupervised model as a pre-trained model. Pre-trained models can learn more discriminative representations by capturing richer and more informative unsupervised features [21], [34], [58], thereby benefiting supervised learning and enhancing our model generalization performance.

TABLE IV: Performance comparisons with vision methods.

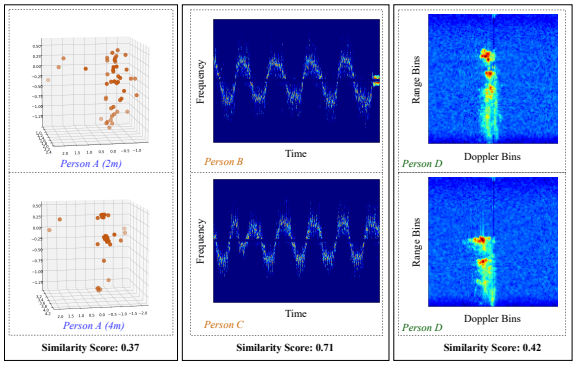
	SimCLR [19]	MoCo [52]	HCM [16]	HCM [16]+Multi-View	Ours
mAP	25.9	26.3	36.7	53.8	59.3

Comparison with Vision Methods. We compare popular unsupervised learning methods from the vision domain, such as SimCLR [19] and MoCo [52]. Additionally, we include the popular unsupervised ReID method HCM [16] from the vision field. Among these, SimCLR [19] and MoCo [52] are contrastive learning frameworks driven by image-based data augmentation. HCM constructs an unsupervised ReID model by applying traditional data augmentations to images and then performing contrastive learning with clustering contrast. Since these methods rely on single-view direct image-style augmentation, we use RDAMap and apply rotation, cropping, and other similar augmentations as one would do for images. The in-environment test performance is shown in Table IV. RDAMap is not suitable for image-style data augmentation, which causes the unsupervised model to fall into shortcut learning [21]. Hence, the unsupervised ReID performance of SimCLR and MoCo is quite poor, reaching mAP less than 30%. Similarly, for HCM [16], applying single-view image-style augmentation directly to RDAMap also leads to shortcut learning, which hinders the model from learning discriminative representations. However, when our cross-view mutual signal contrastive learning is applied to HCM [16], its mAP improves from 36.7% to 53.8%, demonstrating the effectiveness of our multi-view strategy. Compared to the global instance contrast and clustering contrast designed in HCM [16], our study additionally introduces radar-specific local temporal contrast and cluster-center contrast. Therefore, our multi-level mutual signal contrastive learning approach achieves better generalization performance.

Dynamic Performance of Single-View Inference. Since our model employs mutual contrastive learning with different radar signals, we can use any type of signal representation during the inference stage. The inference performance of each modality is dynamically monitored throughout the training process, as shown in Table V. The following observations can be made: (a) In the early stages of training (e.g., the first 32 epochs), the inference performance using the PC modality is better than that of the RDAMap modality. This may be because PC data contains less noise unrelated to human body information, allowing useful features to be learned more quickly. (b) During the mid-training phase (e.g., epochs 64$\sim$128), the inference performance of the two modalities becomes comparable, indicating that the two views begin to achieve balance and complementarity through mutual contrastive learning. (c) In the later stages of training, it can be observed that the performance of these two input types is comparable, demonstrating that our model is robust to the choice of input modality and effectively learns representations from both signals. Using RDAMap as input achieves slightly better performance than using PC. This minor difference may be attributed to the sparsity of radar PC, which inherently contain less information compared to RDAMaps.

TABLE V: The inference performance of each modality during the training epochs under the in-environment setting.

Epoch	8	16	32	64	128	256
mAP (RDAMap)	13.9	35.1	46.8	53.2	57.1	59.3
mAP (PC)	19.2	40.6	49.5	53.4	57.6	58.7



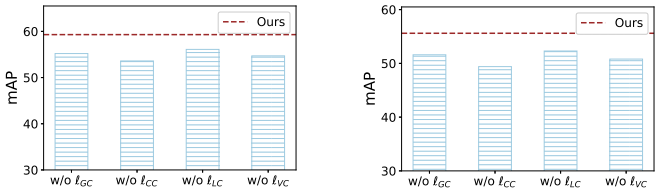
(a) (b) (c)

Fig. 8: Error case analysis. (a): Point clouds of the same person at different distances from the radar; (b) Micro-Doppler spectrograms of different individuals; (c) RDMap of the same person in two different environments.

Error Case Analysis. Given a query pedestrian sample, our unsupervised model searches for similar pedestrians from the gallery set. A higher similarity between samples indicates a greater likelihood that they belong to the same individual. By examining misidentified cases in the search results, we summarize the main causes of recognition errors as follows: (1) Limitations of unsupervised learning. As our method is unsupervised and lacks ground-truth labels, its generalization capability is inherently weaker than that of supervised learning, leading to certain recognition errors. This is a common challenge in unsupervised learning [6], [16], [60]. (2) Variation in pedestrian-radar distance. As shown in Fig. 8(a), when a pedestrian is 4m away from the radar, the PC becomes significantly sparser compared to the 2m case. Increasing the distance reduces the SNR because the reflected energy decays with range. Consequently, PC at longer distances are sparser, which is an unavoidable issue in radar sensing [11]. The calculated similarity between the same person at 2m and 4m is below 0.4, resulting in recognition errors. (3) Similar gaits across different individuals. Fig. 8(b) visualizes the micro-Doppler spectra of two different pedestrians to show gait information. Their spectra appear quite similar, and without label supervision, the model struggles to separate them effectively in the feature space, yielding a high similarity score of 0.71 between two distinct individuals. (4) Environmental variations for the same person. Wireless signals are sensitive to the environment due to multi-path interference, causing the distribution of the same person's features to differ across different surroundings. As illustrated in Fig. 8(c), the RDMaps of the same person in two distinct environments show noticeable differences, leading to a low computed similarity score of 0.42 and thus a misclassification as different individuals.

TABLE VI: Comparison of training time across different unsupervised approaches.

Scheme	TGUL [21]	PRISM [34]	RF-URL [58]	Ours
Batch Time	3.2 ms	3.9 ms	5.1 ms	6.2 ms



(a) In-Environment (b) Cross-Environment

Fig. 9: Loss function ablations.

Training Time Cost. The training time of our approach averages approximately 6.2 ms per batch. We compared the training time overhead with mainstream unsupervised learning methods in the wireless sensing field, and the results are presented in Table VI. Since our design incorporates specific unsupervised ReID modules, such as DBSCAN clustering and GNN models, leading to slightly higher training time costs. However, the performance of our approach is superior. In future work, we plan to explore lightweight unsupervised ReID models to reduce training time overhead.

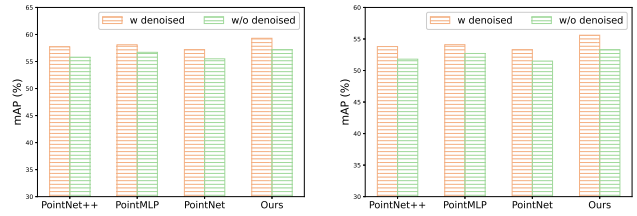
D. Ablation Study

Loss Function Ablation. Our approach incorporates four types of contrastive learning: global contrast, cross-view prediction contrast, clustering contrast, and cross-view cluster center contrast, corresponding to the loss functions ℓ_{GC}, ℓ_{LC}, ℓ_{CC}, and ℓ_{VC}, respectively. Here, we remove one of loss functions and observe the impact on performance, and the results are shown in Fig. 9. It is apparent that each loss function contributes to enhancing model generalization across in-environment and cross-environment settings. For instance, deleting ℓ_{LC} and ℓ_{CC} can decrease mAP by approximately 3% and 6%, respectively. Therefore, our cross-view multi-level contrastive learning enhances the robustness of our unsupervised model.

Channel Attention Ablation. We incorporate a channel attention mechanism into the ME. Here, we evaluate its impact on model performance by replacing the channel attention module with the same number of convolution layers with residual structures. The results are illustrated in Table VII, it can be seen that the channel attention significantly improves model accuracy. For instance, without the channel attention, the model experiences a performance degradation of 1.5% and 1.7% mAP under the in-environment and cross-environment

TABLE VII: Channel attention ablation results.

	In-Environment				Cross-Environment			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
w/o	57.8	70.5	79.5	89.4	53.9	68.1	77.1	84.3
w	59.3	71.8	80.4	90.6	55.6	69.2	78.5	85.7



(a) In-Environment (b) Cross-Environment

Fig. 10: Performance with different PC networks and PC denoising ablation results.

TABLE VIII: The performance comparisons of different signal views for in-environment unsupervised radar ReID.

Different Views	PC & DFS	AoA & RDAMap	AoA & DFS	Ours (PC & RDAMap)
mAP	56.1	52.7	50.9	59.3

settings, respectively. Therefore, our channel attention enables the model to learn more discriminative representations and enhance its generalization capability.

Different PC Networks. To learn PC representations, we employ a local-global-based graph attention network. Here, we compare the performance of model with different PC networks, such as advanced PointNet [35], PointNet++ [37], and PointMLP [36]. The results are illustrated in Fig. 10, and it can be seen that our well-designed local-global-based graph attention network outperforms the current PC networks by large margins. These PC networks, PointNet [35], PointNet++ [37], and PointMLP [36], regard PC as disordered sets. In contrast, given the inherent graph-like properties of a human PC, we construct a graph from the PC and employ a well-designed graph model accordingly. The results in Fig. 10 demonstrate that our proposed approach, by aggregating information from neighboring nodes to update graph representations, can learn more generalized PC representations. Additionally, we apply SOR for PC denoising. Here, we remove PC denoising and observe the performance impacts of our model. The results are shown in Fig. 10, showing that omitting denoising significantly degrades performance. For instance, denoising improves our mAP by approximately 2% under the in-environment and cross-environment scenes, confirming its effectiveness for model generalization. Furthermore, Fig. 10 shows that the PC denoising approach remains equally applicable to PointNet [35], PointNet++ [37], and PointMLP [36], demonstrating its generalizability to enhance the robustness of PC networks.

Performances under Different Signal Views. Our study relies on two signal representations: PC and frequency spectrum RDAMap. Other signal forms, such as AoA (Angle of Arrival) and DFS (Doppler Frequency Shift), are inherently contained within these two representations. For instance, AoA angle information is reflected in the spatial position of PC, while DFS frequency information exists in the Doppler dimension of RDAMaps. However, PC provides richer information than AoA alone, as they include not only angle but also range and velocity information. Similarly, RDAMaps contain more comprehensive information than standalone DFS data. Therefore, to achieve a more holistic representation of target objects,

TABLE IX: Model performance impacts with Local and global attention modules in our PCE.

Local Attention	Global Attention	mAP (In / Cross-Environment)
		58.2 / 54.3
✓		58.9 / 55.2
	✓	58.6 / 55.1
✓	✓	**59.3 / 55.6**

TABLE X: Masking strategy ablation results. RSM: Random Spatial Masking; TM: Temporal Masking; IBM: Intensity-Based Masking.

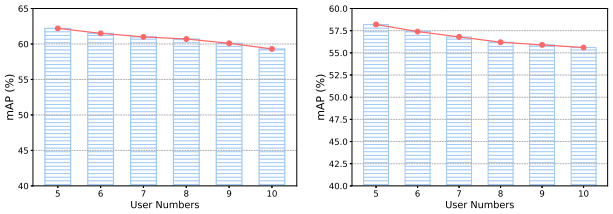
w/o Mask	RSM	IBM	TM	mAP (In / Cross-Environment)
✓				57.1 / 53.5
	✓			57.8 / 54.0
		✓		58.8 / 54.9
			✓	58.5 / 54.5
		✓	✓	**59.3 / 55.6**

current radar sensing approaches generally employ integrated PC or frequency spectrum representations. Here, we also compare the performance of different signal view combinations for in-environment unsupervised radar ReID. We evaluate the following view pairs: PC with DFS, AoA with RDAMap, and AoA with DFS. The experimental results are summarized in Table VIII, which clearly shows that our approach using PC and RDAMap views achieves the best performance. When the selected views contain more comprehensive information, our multi-level mutual signal contrastive learning enables the model to learn more discriminative representations.

Local and Global Attention in PCE. Our PCE adopts a graph neural network architecture, incorporating both local and global attention mechanisms. We conduct ablation studies about the local and global attentions of the PCE, and the results are presented in Table IX. When the local and global attention mechanisms are removed, the model achieves 58.2% and 54.3% mAP in in-environment and cross-environment settings, respectively, which are 1.1% and 1.3% lower than our proposed approach. The model performance improves in both in-environment and cross-environment settings when the local and global attention mechanisms are incorporated individually. When both local and global attention mechanisms are sequentially incorporated, the model achieves the best unsupervised performance. The results in Table IX validate the effectiveness of the local and global attentions in our PCE.

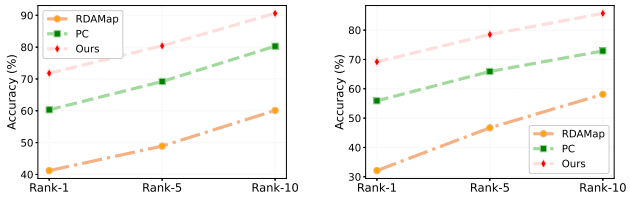
Masking Strategy in RDAMap. To enhance model generalization, our approach integrates temporal and spatial masking strategies in the ME encoder. As shown in Table X, we conduct ablation experiments on different masking strategies. When no masking strategy is applied, the model respectively achieves mAPs of 57.1% and 53.5% on in-environment and cross-environment, which are 2.2% and 2.1% lower than ours. This demonstrates that the proposed masking strategy effectively enhances the model's generalization capability. During spatial masking, due to the limited informative content in RDAMap, we adopt an intensity-based masking strategy instead of random masking. As illustrated in Table X, random spatial masking improves mAP from 57.1% to 57.8% under

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2026.3713047



(a) In-Environment (b) Cross-Environment

Fig. 11: mAP with different testing user numbers.



(a) In-Environment (b) Cross-Environment

Fig. 12: CMC curves of solely signal unsupervised training.

TABLE XI: Performance comparison of different sequential architectures.

Architectures	In-Environment		Cross-Environment	
	LSTM	Transformer	LSTM	Transformer
mAP	59.3	45.1	55.6	41.8

the in-environment scene. In contrast, intensity-based masking achieves a more significant improvement, increasing the mAP from 57.1% to 58.8%, outperforming random spatial masking. Additionally, temporal masking also enhances the model's generalization performance, boosting the mAP from 57.1% to 58.5% under the in-environment scene. When combining the proposed intensity-based spatial masking and temporal masking strategies, the model achieves optimal performance.

Different Sequential Model Architectures. Since wireless signals are sequential, many related works employ deep learning models that account for temporal dependencies [11], [61]–[63], such as LSTM [30] and Transformer [64]. In fields like Natural Language Processing (NLP) and time-series modeling, the Transformer architecture is generally considered more advanced and often yields superior performance compared to LSTM. However, the Transformer architecture is highly sensitive to the amount of data [65], [66]. Relevant studies indicate that with insufficient data, the Transformer struggles to learn discriminative representations [61], [62], [67]–[69], leading to limited generalization ability. Therefore, works in text or image domains that adopt the Transformer typically rely on tens of millions to hundreds of millions of samples. In wireless sensing, however, the available data volume is far below that of text or image domains, often comprising only a few thousand samples. Moreover, some studies [61], [62] in wireless sensing have also found that when evaluated on public datasets, models based on the Transformer architecture underperform compared to LSTM, and may even generalize worse than CNN-based models. These works further demonstrate that the Transformer architecture is generally less suitable for scenarios with limited data in wireless sensing. Additionally, the Transformer has higher computational complexity than LSTM. In our experiments, we also test the Transformer, but it perform worse than LSTM. As shown in Table XI, replacing LSTM with Transformer leads to a drop in mAP of approximately 15%. This finding aligns with the results reported in prior works [61], [62], confirming that LSTM is better suited for our wireless sensing setting where data volume is relatively small.

Different User Numbers. To evaluate the performance of our unsupervised model with varying numbers of users, we test the trained model with 5 to 10 users, with results shown in Fig. 11. Our findings indicate that while test performance shows a slight degradation as the number of users increases. This decline is not statistically significant, demonstrating the robustness of our model across different user population sizes. The increased number of users introduces greater data complexity, which may distract the model and consequently increase learning difficulty. In future work, we will investigate the impact of user population size and determine the boundary of user numbers for unsupervised mmWave-based ReID.

Single Signal Unsupervised Learning. We adopt a signal mutual contrast learning approach, involving treating RDAMap and PC as different augmented versions of a sample. In conventional contrastive learning, a single data representation is typically used, and different versions of the data are generated through augmentation to compute contrastive loss. Here, we also employ single-signal data augmentation to create contrastive versions for training our unsupervised model, and the experimental results are illustrated in Fig. 12. For RDAMap, there is a lack of tailored data augmentation techniques, so we adopt image-like transformations such as rotation and cropping to augment RDAMap and apply our multi-level contrastive learning on the augmented versions. However, contrastive learning using only RDAMap leads to a significant performance drop, achieving only around 40% Rank-1 accuracy. Radar signals are structured data, and RDAMap contains a large amount of non-informative content, with bright regions representing only a small fraction. Directly applying image-based transformations and contrastive learning to RDAMap may be inappropriate, which may cause the unsupervised model to fall into shortcut learning [21]. Thus, training our unsupervised model using only RDAMap fails to achieve satisfactory results. Similarly, constructing contrastive learning solely based on PC results in lower performance than ours, as shown in Fig. 12. Despite PC data having less noise, the sparsity of mmWave radar PC leads to limited information, resulting in unsatisfactory unsupervised performance. In contrast, our approach treats RDAMap and PC as different augmented versions of the same sample, addressing the challenge of augmenting RDAMap alone. Furthermore, RDAMap contains richer information, which complements the sparse PC data. As a result, our approach achieves a more generalized unsupervised model, capable of learning more comprehensive radar signal representations.

Different Training-Testing Person Split Ratios. Table III

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2026.3713047

TABLE XII: Performance comparison under different numbers of identities in training and testing sets.

Training : Testing	30:10	20:10	20:20
mAP	59.3	56.7	52.4

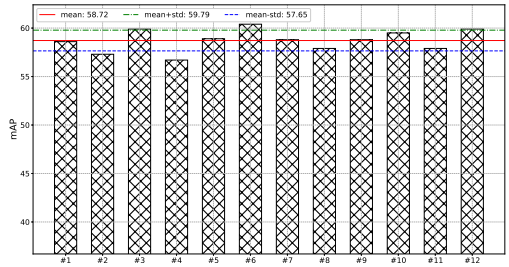


Fig. 13: In-environment unsupervised ReID performance under 12 randomized gallery sets.

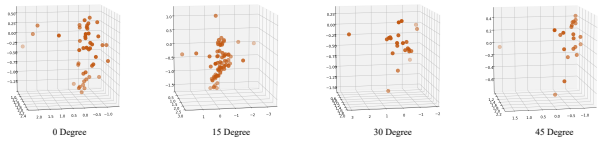
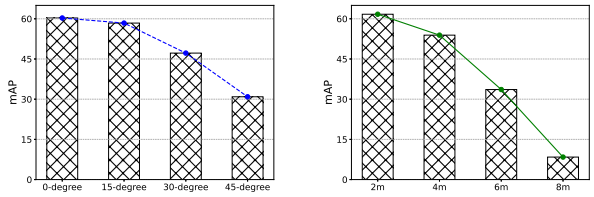


Fig. 14: Point clouds at different angles.



(a) Different Angles (b) Different Distances

Fig. 15: Performance with different angles and distances.

shows that our model achieves an mAP of 59.3% with 30 individuals for training and 10 new users for testing under the in-environment setting. Here, we test results about 20 individuals for training and 10 for testing, as well as 20 individuals for training and 20 for testing under the in-environment setting. The results are shown in Table XII. It can be observed that when training with 20 individuals and testing with 10, the mAP is approximately 3% lower than that achieved with 30 training individuals, indicating that a smaller training set leads to reduced model robustness. When the test set is expanded to 20 individuals, the model's performance drops to an mAP of 52.4%. Similar to most deep learning-based ReID methods [1], [59], as the search space of the test set increases, the retrieval performance slightly decreases, which is a normal phenomenon. Therefore, while the size of the training set and the search space of the test set influence model performance, they do not cause drastic fluctuations.

Performance with Randomized Identity. We conduct 12 randomized experiments by selecting 10 different identity combinations from the total population to form distinct gallery sets and evaluate their in-environment performance. As shown in Fig. 13, the ReID performance across different gallery sets remains stable, ranging between 56% and 60% mAP. This demonstrates that our model generalizes well across various test sets. The average result from multiple experiments is 58.7%, and the performance remains highly stable, with a standard deviation (std) of mAP of only 1.06%. This indicates that our model is not overfit to specific identity patterns but has learned generalized radar feature representations. Additionally, unsupervised learning does not rely on specific labels, and its training process inherently aims to learn universal signal representations. Therefore, theoretically, unsupervised learning is less susceptible to constraints from particular identities.

Performances under Different Positions. We test radar positions relative to the person at 0-degree, 15-degree, 30-degree, and 45-degree. We visualize the PC at these different angles, as shown in Fig. 14. It can be observed that as the angle between the person and the radar increases, the PC becomes sparser, particularly at 30-degree and 45-degree positions. We group and test these different angle positions under the in-environment setting, with the results illustrated in Fig. 15(a). At the 0-degree position, the model achieves the best performance, while at 45 degrees, the mAP drops to approximately 30%. The 0-degree angle indicates that the radar is directly facing the person, yielding optimal performance since the radar signal energy is highest in this frontal orientation. As the angle increases, the energy of the radar beam decreases, leading to performance degradation. Therefore, when the person is at a larger angle relative to the radar, the PC becomes sparser, and ReID performance declines accordingly, which is a relatively common phenomenon in radar-based sensing [11].

Performances under Different Distances. We test distances of 2, 4, 6, and 8 meters, grouping the test sets based on the distance from the pedestrian to the radar. We visualize pedestrians at varying distances, and the results are shown in Fig. 16. It can be observed that as the distance increases, the PC becomes increasingly sparse. Particularly at 8 meters, there are only one or two PC remaining. Fig. 15(b) illustrates the mAP under the in-environment setting at different distances. It can be seen that as the distance increases, performance gradually declines. At a distance of 8 meters, the mAP drops to only about 10%, indicating that unsupervised ReID is almost ineffective at this range. For radar signals, increasing distance leads to a decrease in SNR, as the reflected energy attenuates with distance. Additionally, due to the low angular resolution of mmWave radar, greater distances result in fewer and sparser PC. Therefore, in current radar-based sensing tasks, it is necessary to control the distance to avoid excessive range [11].

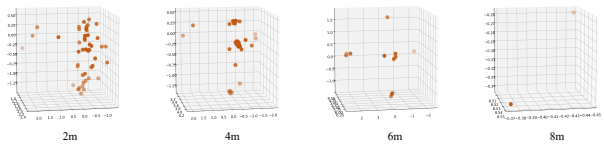
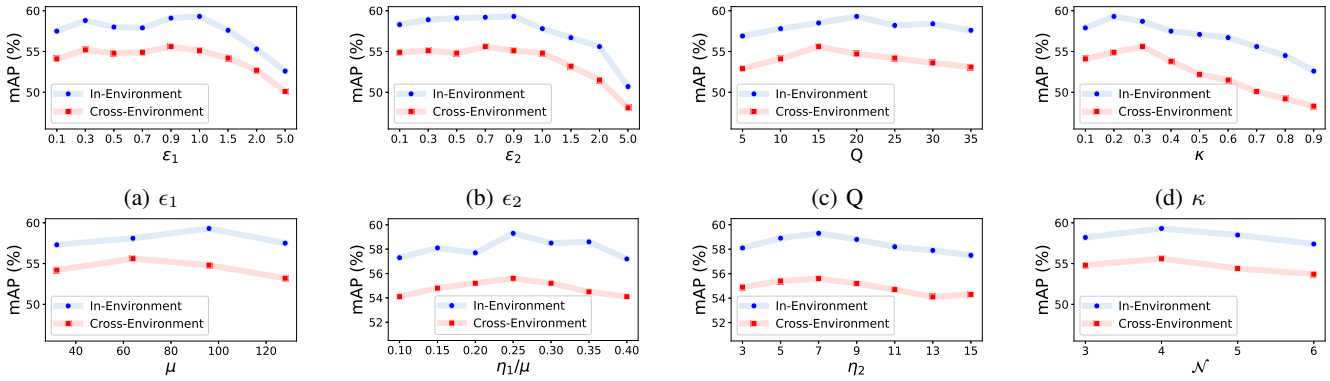


Fig. 16: Point clouds at different distances.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2026.3713047



(a) ϵ_1 (b) ϵ_2 (c) Q (d) κ

(e) μ (f) η_1 (g) η_2 (h) $\mathcal{N}$

Fig. 17: Hyper-parameter sensitivity analysis experiments.

E. Parameter Study

We investigate the impact of different hyper-parameters on model performance, including loss weights ϵ_1 and ϵ_2, prediction timesteps Q, momentum factor κ, masking ratios μ, η_1, and η_2, and channel attention layers $\mathcal{N}$. The results are shown in Fig. 17.

Fig. 17(a) and 17(b) show the results of varying ϵ_1 and ϵ_2. We observe that as $\epsilon_1, \epsilon_2 < 1.5$, model is less sensitive to their values. The model achieves best with $\epsilon_1 = 1$ and $\epsilon_2 = 0.9$. Fig. 17(c) shows that the model achieves the best with $Q = 15$ and $Q = 20$ under in-environment and cross-environment, respectively. Increasing predicted timesteps can improve model performance, but larger timesteps harm performance, which may lead to inaccurate predictions. Fig. 17(d) shows that the model achieves best with $\kappa = 0.2$ and $\kappa = 0.3$ under in-environment and cross-environment, respectively. The momentum factor controls the update speed of cluster centroids. The larger values κ, the slower the update results in trivial cluster centroids. Fig. 17(e), Fig. 17(f), and Fig. 17(g) demonstrate the results of varying μ, η_1, and η_2, and increasing masking ratios enhances model performance. However, too large values of μ, η_1, and η_2 may lead to much information loss, which harms model performances. Fig. 17(h) illustrates the impact of different channel attention layers on model performance. It can be observed that the model achieves optimal performance when $\mathcal{N} = 4$. Further increasing the number of channel attention layers results in a deeper model, which may lead to overfitting.

Our study sets $\hat{k} = 5$ in SOR denoising. If $\hat{k}$ is too small, the algorithm becomes extremely sensitive to noise; if $\hat{k}$ is too large, edge points of the human body may be misidentified as outliers and removed. For example, when $\hat{k} = 1$, SOR only considers the nearest neighbor. Even an isolated noise point may be retained if there happens to be another randomly generated noise point nearby, as the distance between them can be very small. SOR may mistakenly assume that this noise point lies in a high-density region and thus treat it as normal data. Conversely, when $\hat{k}$ is set to a large value (e.g., $\hat{k} = 20$), for a point at the end of an arm, the algorithm not only calculates its distance to neighboring points on the arm but is also forced to compute distances to more distant points on the torso or even the legs. This can lead to a significantly inflated average distance for that point. As a result, SOR may deem such points too far from others and classify them as outliers. As shown in Table XIII, we test $\hat{k} = \{4, 5, 6\}$ and find that the model exhibits stable generalization performance across these values, with optimal performance achieved at $\hat{k} = 5$.

TABLE XIII: Sensitivity analysis of parameter $\hat{k}$ in SOR.

	In-Environment			Cross-Environment		
$\hat{k}$	4	5	6	4	5	6
mAP	59.0	59.3	58.7	55.4	55.6	55.1

F. Performance on Public Dataset

Currently, there are few open-source radar datasets available on the market [7], [61], and most of them do not meet our requirements. This is because these open-source datasets do not provide raw data but only release processed data. For instance, when such processed data is converted into RGB thermal images, it loses fundamental amplitude and phase information, making it impossible for us to restore or further process the data. Since our method requires two different radar signal views, it is challenging to test on datasets where raw data is not available. Fortunately, although the mmGait dataset [70], [71] does not provide raw data, it does offer paired point cloud and Doppler spectrogram data. Therefore, we treat these two types of data as distinct views and train an unsupervised mmWave radar ReID model using our method. Here, we replace the RDAMap data in our approach with Doppler spectrograms. Since the Doppler spectrogram at timestep t is one-dimensional, the proposed ME module in our paper is replaced with a fully connected network, while the rest of the structure of our proposed model remains unchanged.

The mmGait dataset contains 314 valid paired samples from 10 users. We randomly selected 5 users for the training set, with the remaining 5 users assigned to the query and gallery sets. Table XIV presents the unsupervised performance tested on the mmGait dataset, showing that our method achieves better generalization compared to existing unsupervised approaches. For example, our method attains an mAP of 65.2%

TABLE XIV: Unsupervised ReID performance comparison with baselines on public mmGait dataset.

Scheme	mAP	Rank-1	Rank-5	Rank-10
TGUL [21]	41.2	51.9	63.2	75.6
PRISM [34]	46.4	58.4	65.7	79.8
RF-URL [58]	50.6	62.3	70.1	85.3
Ours	65.2	74.6	84.9	93.1

TABLE XV: Model training computational costs.

Model size	Model parameters	Flops	Batch time
43.8MB	11.4M	7.9G	6.2ms

on the mmGait dataset, which is approximately 15% higher than that of RF-URL [58]. Currently, wireless-based unsupervised methods [21], [34], [58] are not specifically designed for ReID and primarily focus on learning unsupervised representations. In contrast, our approach includes modules tailored for unsupervised ReID (e.g., multi-level contrastive loss), leading to superior unsupervised performance. The evaluation results on additional public datasets further demonstrate the strong generalization capability of our model.

V. Discussion

Our work still has room for improvement, and here we outline two potential directions for future research.

Lightweight Model. Since unsupervised learning operates without labels, it typically requires more complex model designs compared to supervised approaches, resulting in greater computational overhead. Additionally, our framework incorporates attention mechanisms and multi-level contrastive learning, which introduce extra parameters and further increase training computation. As shown in Table XV, our model requires 43.8MB of memory, contains 11.4M parameters, performs 7.9G FLOPs, and processes each batch in 6.2ms on average. Current sensing models primarily focus on accuracy improvement [72], which inevitably increases model complexity. However, in practical sensing scenarios, models are often deployed on mobile devices, which require lightweight architectures. Therefore, in future work, we plan to explore lightweight unsupervised ReID models by leveraging advanced techniques such as model compression and pruning.

Multi-Modal Unsupervised ReID. We achieve unsupervised ReID using a single modality, mmWave radar. However, in practical scenarios, multiple devices (e.g., cameras, WiFi) can be deployed collaboratively to enhance ReID performance. For instance, Wi-Fi and radar can be utilized for ReID in low-light conditions, while advanced camera-based ReID techniques can be employed in well-lit environments. Multi-modal fusion has the potential to surpass the performance of single-modality approaches and adaptively align with real-world conditions [6]. In future work, we aim to explore unsupervised multi-modal ReID, addressing challenges such as coordinating multiple devices and bridging domain distribution gaps across different modalities.

Imbalanced Radar ReID. Currently, our approach does not address the issue of data imbalance. Our dataset is balanced, with the training set, query set, and gallery set all evenly partitioned. Imbalanced data may influence model bias, which is likely to distort the overall feature space and impair the generalization capability of trained models [73], [74]. We will incorporate this consideration into our future work to explore unsupervised radar ReID under imbalanced data conditions.

Push Sensing Limits of Angle and Distance. Our experiments show that ReID performance degrades when the angle or distance between a pedestrian and the radar becomes too large. This phenomenon is related to the inherent characteristics of mmWave radar and is common and unavoidable [7], [10], [11]. However, in real-world applications, scenarios with excessively large angles and distances do exist. Therefore, in future research, we will focus on addressing the impact of distance and angle on ReID, for example, by employing cascaded multiple radars to achieve broader coverage.

VI. Related Work

Unsupervised ReID. Person ReID aims to search for a person of interest within target databases. With the advancement of deep learning, this technology has gained significant traction in the field of computer vision, leading to the development of numerous advanced ReID methods [1], [59], [75]–[78]. The ReID task is often formulated as a clustering problem, where representations of the same individual are pulled closer in the high-dimensional space, while those of different individuals are pushed apart [59]. For example, Luo *et al.* [59] propose a strong baseline of supervised person ReID, employing a model based on Convolutional Neural Networks (CNN) optimized with batch normalization neck, identity loss, and triplet loss [79]. However, supervised learning incurs substantial annotation costs in practical applications. To mitigate the substantial expense of data annotation, unsupervised person ReID is progressively gaining popularity [2], [50], [60]. Presently, unsupervised ReID primarily falls into two categories: unsupervised domain adaptation and purely unsupervised methods. Unsupervised domain adaptation necessitates a labeled source domain, which is then transferred to an unlabeled target domain. Nevertheless, unsupervised domain adaptation based on requiring a labeled source domain, along with the constraint that the distributions of the target and source domains should not differ significantly, imposes substantial limitations on practical applications [2], [16], [50]. Hence, pure unsupervised ReID has garnered widespread attention, which mainly explores sample relations by unsupervised contrastive learning. For instance, CCL [50] introduces a cluster-level memory dictionary and conducts contrastive loss computation at cluster level. HCM [16] proposes hybrid contrastive learning for unsupervised ReID at the identity level and the image level. In our study, due to the absence of labeled radar ReID source domains, we opt for pure unsupervised ReID.

mmWave Sensing. Radar captures variations in human movement patterns by emitting electromagnetic waves. In contrast to vision perception, radar can operate in dark environments and possesses a certain degree of penetrative ability,

such as through clothing and masks. Consequently, mmWave sensing has garnered significant attention in applications like gesture recognition [80], pose estimation [40], respiration monitoring [81], and human silhouette generation [58]. The existing mmWave-based human sensing works select suitable radar signal forms according to the sensing task, and there are generally three signal forms: frequency spectrum, phase waveform, and PC [11]. Specifically, frequency spectrum and PC are commonly used for motion recognition, since they indicate the environment reflection, including the position and posture of the human. Whereas in biometric measurements like respiration and heartbeat detections, phase waveforms are chosen due to their unique ability to characterize the micro-motions. For example, mmMesh [82] reconstructs the human mesh from mmWave signals by extracting PC and using an attention-based deep learning framework. HuPR [40] proposes a benchmark for pose estimation by extracting frequency spectrum features from mmWave signals and employing a graph-based deep learning model. In contrast, the works [83], [84] extract phase waveform from radar signals to sense vital signs and employ learning-based waveform extraction or signal decomposition techniques.

mmWave-Based ReID. Current studies extract frequency spectrogram features or PC from radar signals and then feed it into DNNs to construct ReID models. For instance, the works [6], [12], [13], [71] delve into person ReID by using PC extracted from radar signals and feeding it into DNNs, albeit in a supervised fashion. The works [5], [7], [15], [85], [86] utilize frequency spectrogram features for radar sensing and learn deep representations of pedestrians through supervised learning. Current mmWave-based ReID systems predominantly rely on supervised learning, which incurs significant labeling costs. Moreover, radar signals are not interpretable by the human eye, making the labeling process even more challenging. Therefore, our study aims to explore unsupervised person ReID using commercial mmWave radar, reducing labeling costs.

VII. Conclusion

This paper proposes a mmWave radar-based unsupervised person ReID method. We employ a multi-level mutual contrast learning mode using different signal representations to drive SSCL and create complementarity between RDAMap and PC, including global contrast, cross-view prediction contrast, cluster contrast, and cross-view cluster center contrast. Experimental results show that our approach achieves excellent unsupervised ReID performance on a real-world dataset, outperforming the existing methods by large margins.

References

[1] M. Ye *et al.*, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, 2022.

[2] M. Li *et al.*, "Cluster-guided asymmetric contrastive learning for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 3606–3617, 2022.

[3] W. Xu *et al.*, "Adversarial feature disentanglement for long-term person re-identification," in *Proceedings of IJCAI*, 2021, pp. 1201–1207.

[4] H. Yao *et al.*, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, 2019.

[5] L. Fan *et al.*, "Learning longterm representations for person re-identification using radio signals," in *Proceedings of CVPR*. IEEE, 2020, pp. 10 696–10 706.

[6] D. Cao *et al.*, "Cross vision-rf gait re-identification with low-cost RGB-D cameras and mmwave radars," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 3, pp. 102:1–102:25, 2022.

[7] C. Han *et al.*, "mmreid: Person reidentification based on commodity millimeter-wave radar," *IEEE Internet Things J.*, vol. 12, no. 13, pp. 24 057–24 070, 2025.

[8] Y. Ren *et al.*, "Person re-identification in 3d space: A wifi vision-based approach," in *Proceedings of USENIX Security*. USENIX Association, 2023, pp. 5217–5234.

[9] J. Liu *et al.*, "Wireless sensing for human activity: A survey," *IEEE Commun. Surv. Tutorials*, vol. 22, no. 3, pp. 1629–1645, 2020.

[10] Q. Feng *et al.*, "mmwave radar-based unsupervised gesture recognition via image-aligned heterogeneous domain transfer," *IEEE Trans. Mob. Comput.*, pp. 1–17, 2025.

[11] J. Zhang *et al.*, "A survey of mmwave-based human sensing: Technology, platforms and applications," *IEEE Commun. Surv. Tutorials*, vol. 25, no. 4, pp. 2052–2087, 2023.

[12] Y. Cheng *et al.*, "Person reidentification based on automotive radar point clouds," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–13, 2022.

[13] Y. Xiang *et al.*, "Person identification method based on pointnet++ and adversarial network for mmwave radar," *IEEE Internet Things J.*, vol. 11, no. 6, pp. 10 104–10 114, 2024.

[14] T. Wang *et al.*, "Open-set occluded person identification with mmwave radar," *IEEE Trans. Mob. Comput.*, vol. 24, no. 6, pp. 5229–5244, 2025.

[15] J. Li *et al.*, "Passive multiuser gait identification through micro-doppler calibration using mmwave radar," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 6868–6877, 2024.

[16] T. Si *et al.*, "Hybrid contrastive learning for unsupervised person re-identification," *IEEE Trans. Multim.*, vol. 25, pp. 4323–4334, 2023.

[17] H. Chen *et al.*, "ICE: inter-instance contrastive encoding for unsupervised person re-identification," in *Proceedings of ICCV*. IEEE, 2021, pp. 14 940–14 949.

[18] X. Liu *et al.*, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, 2023.

[19] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *Proceedings of ICML*, vol. 119, 2020, pp. 1597–1607.

[20] Q. Feng *et al.*, "Privacy-preserving image retrieval in cloud computing via adaptive secret keys and self-supervised block-augmented pretraining," *IEEE Trans. Serv. Comput.*, vol. 18, no. 4, pp. 2310–2325, 2025.

[21] T. Li *et al.*, "Unsupervised learning for human sensing using radio signals," in *Proceedings of WACV*. IEEE, 2022, pp. 1091–1100.

[22] W. Hou *et al.*, "Rfboost: Understanding and boosting deep wifi sensing via physical data augmentation," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 2, pp. 58:1–58:26, 2024.

[23] Y. Tian *et al.*, "Contrastive multiview coding," in *Proceedings of ECCV*, ser. Lecture Notes in Computer Science, vol. 12356. Springer, 2020, pp. 776–794.

[24] Y. Liu *et al.*, "Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion," in *Proceedings of NeurIPS*, 2023.

[25] S. Rao, "Introduction to mmwave sensing: Fmcw radars," *Texas Instruments (TI) mmWave Training Series*, pp. 1–11, 2017.

[26] J. Li *et al.*, *MIMO radar signal processing*. John Wiley & Sons, 2008.

[27] A. Farina *et al.*, "A review of cfar detection techniques in radar systems," *Microwave Journal*, vol. 29, p. 115, 1986.

[28] R. D. DeGroat *et al.*, "The constrained MUSIC problem," *IEEE Trans. Signal Process.*, vol. 41, no. 3, pp. 1445–1449, 1993.

[29] R. B. Rusu *et al.*, "3d is here: Point cloud library (PCL)," in *Proceedings of ICRA*. IEEE, 2011, pp. 1–4.

[30] M. Schuster *et al.*, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.

[31] G. Liu *et al.*, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.

[32] J. Hu *et al.*, "Squeeze-and-excitation networks," in *Proceedings of CVPR*. IEEE, 2018, pp. 7132–7141.

[33] A. Bordes *et al.*, "A semantic matching energy function for learning with multi-relational data - application to word-sense disambiguation," *Mach. Learn.*, vol. 94, no. 2, pp. 233–259, 2014.

[34] L. Fang *et al.*, "PRISM: pre-training RF signals in sparsity-aware masked autoencoders," in *Proceedings of INFOCOM*. IEEE, 2024, pp. 2109–2118.

[35] C. R. Qi *et al.*, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of CVPR*. IEEE, 2017, pp. 77–85.

[36] X. Ma *et al.*, "Rethinking network design and local geometry in point cloud: A simple residual MLP framework," in *Proceedings of ICLR*, 2022.

[37] C. R. Qi *et al.*, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proceedings of NeurIPS*, 2017, pp. 5099–5108.

[38] S. Palipana *et al.*, "Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 1, pp. 27:1–27:27, 2021.

[39] W. Shi *et al.*, "Point-gnn: Graph neural network for 3d object detection in a point cloud," in *Proceedings of CVPR*. IEEE, 2020, pp. 1708–1716.

[40] S. Lee *et al.*, "Hupr: A benchmark for human pose estimation using millimeter wave radar," in *Proceedings of WACV*. IEEE, 2023, pp. 5704–5713.

[41] H. Xue *et al.*, "M^{4}esh: mmwave-based 3d human mesh construction for multiple subjects," in *Proceedings of SenSys*. ACM, 2022, pp. 391–406.

[42] P. Velickovic *et al.*, "Graph attention networks," in *Proceedings of ICLR*. OpenReview.net, 2018.

[43] T. N. Kipf *et al.*, "Semi-supervised classification with graph convolutional networks," in *Proceedings of ICLR*. OpenReview.net, 2017.

[44] Z. Wu *et al.*, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 1, pp. 4–24, 2021.

[45] Y. Wang *et al.*, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 146:1–146:12, 2019.

[46] L. J. Ba *et al.*, "Layer normalization," *CoRR*, vol. abs/1607.06450, 2016.

[47] D. Hendrycks *et al.*, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," *CoRR*, vol. abs/1606.08415, 2016.

[48] S. Ioffe *et al.*, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of ICML*, vol. 37, 2015, pp. 448–456.

[49] E. Eldele *et al.*, "Time-series representation learning via temporal and contextual contrasting," in *Proceedings of IJCAI*, 2021, pp. 2352–2359.

[50] Z. Dai *et al.*, "Cluster contrast for unsupervised person re-identification," in *Proceedings of ACCV*, vol. 13846. Springer, 2022, pp. 319–337.

[51] M. Ester *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of KDD*, 1996, pp. 226–231.

[52] K. He *et al.*, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of CVPR*. IEEE, 2020, pp. 9726–9735.

[53] H. Hou *et al.*, "Unsupervised cross-domain person re-identification with self-attention and joint-flexible optimization," *Image Vis. Comput.*, vol. 111, p. 104191, 2021.

[54] L. Zheng *et al.*, "Scalable person re-identification: A benchmark," in *Proceedings of ICCV*. IEEE, 2015, pp. 1116–1124.

[55] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of NeurIPS*, 2019, pp. 8024–8035.

[56] Q. Feng *et al.*, "Evit: Privacy-preserving image retrieval via encrypted vision transformer in cloud computing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7467–7483, 2024.

[57] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of EMNLP*, 2020, pp. 38–45.

[58] R. Song *et al.*, "RF-URL: unsupervised representation learning for RF sensing," in *Proceedings of MobiCom*. ACM, 2022, pp. 282–295.

[59] H. Luo *et al.*, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multim.*, vol. 22, no. 10, pp. 2597–2609, 2020.

[60] X. Lin *et al.*, "Unsupervised person re-identification: A systematic survey of challenges and solutions," *CoRR*, vol. abs/2109.06057, 2021.

[61] F. Wang *et al.*, "XRF55: A radio frequency dataset for human indoor action analysis," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 1, pp. 21:1–21:34, 2024.

[62] J. Yang *et al.*, "Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing," *Patterns*, vol. 4, no. 3, p. 100703, 2023.

[63] Q. Feng *et al.*, "Imbalanced semi-supervised learning for wifi gesture recognition via dynamic threshold-based spatio-temporal attention networks," *IEEE Trans. Mob. Comput.*, vol. 25, no. 1, pp. 483–499, 2026.

[64] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of NeurIPS*, 2017, pp. 5998–6008.

[65] R. Durall *et al.*, "Combining transformer generators with convolutional discriminators," in *Proceedings of KI*, vol. 12873. Springer, 2021, pp. 67–79.

[66] W. Huang *et al.*, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proceedings of Interspeech*. ISCA, 2020, pp. 4676–4680.

[67] Q. Feng *et al.*, "DHAN: encrypted JPEG image retrieval via DCT histograms-based attention networks," *Appl. Soft Comput.*, vol. 133, p. 109935, 2023.

[68] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of ICLR*. OpenReview.net, 2021.

[69] Q. Feng *et al.*, "End-to-end privacy-preserving image retrieval in cloud computing via anti-perturbation attentive token-aware vision transformer," *Inf. Fusion*, vol. 121, p. 103153, 2025.

[70] R. Mazzieri *et al.*, "mmgait10 - open-set gait recognition from sparse mmwave radar point clouds (version 2)," 2025.

[71] ——, "Open-set gait recognition from sparse mmwave radar point clouds," *IEEE Sensors Journal*, vol. 25, no. 17, pp. 33 051–33 063, 2025.

[72] Y. Liu *et al.*, "Unifi: A unified framework for generalizable gesture recognition with wi-fi signals using consistency-guided multi-view networks," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 4, pp. 168:1–168:29, 2023.

[73] X. Cheng *et al.*, "Dual-path imbalanced feature compensation network for visible-infrared person re-identification," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 21, no. 1, pp. 20:1–20:24, 2025.

[74] P. Wang *et al.*, "Ltreid: Factorizable feature generation with independent components for long-tailed person re-identification," *IEEE Trans. Multim.*, vol. 25, pp. 4610–4622, 2023.

[75] L. Zheng *et al.*, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, 2019.

[76] S. Li *et al.*, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proceedings of CVPR*. IEEE, 2018, pp. 369–378.

[77] W. Li *et al.*, "Harmonious attention network for person re-identification," in *Proceedings of CVPR*. IEEE, 2018, pp. 2285–2294.

[78] J. Si *et al.*, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proceedings of CVPR*. IEEE, 2018, pp. 5363–5372.

[79] F. Schroff *et al.*, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of CVPR*. IEEE, 2015, pp. 815–823.

[80] Y. Li *et al.*, "Towards domain-independent and real-time gesture recognition using mmwave signal," *IEEE Trans. Mob. Comput.*, vol. 22, no. 12, pp. 7355–7369, 2023.

[81] J. Choi *et al.*, "Remote respiration monitoring of moving person using radio signals," in *Proceedings of ECCV*, vol. 13697. Springer, 2022, pp. 253–270.

[82] H. Xue *et al.*, "mmmesh: towards 3d real-time dynamic human mesh construction using millimeter-wave," in *Proceedings of MobiSys*. ACM, 2021, pp. 269–282.

[83] S. Zhang *et al.*, "Can we obtain fine-grained heartbeat waveform via contact-free rf-sensing?" in *Proceedings of INFOCOM*. IEEE, 2022, pp. 1759–1768.

[84] X. Xu *et al.*, "mmecg: Monitoring human cardiac cycle in driving environments leveraging millimeter wave," in *Proceedings of INFOCOM*. IEEE, 2022, pp. 90–99.

[85] Y. Yang *et al.*, "Multiscenario open-set gait recognition based on radar micro-doppler signatures," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[86] Z. Ni *et al.*, "Gait-based person identification and intruder detection using mm-wave sensing in multi-person scenario," *IEEE Sensors Journal*, vol. 22, no. 10, pp. 9713–9723, 2022.