Mission: mmWave Radar Person Identification with RGB Cameras

Ruofeng Liu[1*], Tianshuo Yao[2*]
Ruili Shi[2], Luoyu Mei[2,3], Shuai Wang[2†]
Zhimeng Yin[3], Wenchao Jiang[4], Shuai Wang[2]
[1]Michigan State University, East Lansing, MI, USA
[2]Southeast University, Nanjing, Jiangsu, China
[3]City University of Hong Kong, Hong Kong SAR, China
[4]Singapore University of Technology and Design, Singapore
liuruofe@msu.edu,{220215632,shiruili,lymei-,shuaiwang_iot}@seu.edu.cn
zhimeyin@cityu.edu.hk,wenchao_jiang@sutd.edu.sg,shuaiwang@seu.edu.cn

ABSTRACT

This paper presents Mission, the first-of-this-kind cross-modal reidentification (ReID) design for mmWave Radar and RGB cameras. Given a person of interest detected by Radar in camera-restricted scenarios, Mission can identify the image of the person from cameras that are ubiquitously deployed in camera-allowed areas. We envision that cross Vison-RF ReID can significantly enrich mmWave human sensing with a wide spectrum of applications in security surveillance, tracking, and personalized services. Technically, we introduce a novel method for cross-modal similarity estimation that exploits inherent synergies between fine-grained 2D images and coarse-grained 3D Radar point clouds to effectively overcome their modal discrepancy. Through extensive experiments, we demonstrated that our proposed system can achieve 85% top-1 accuracy and 90% top-5 accuracy among 58 volunteers.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing**; • **Computer systems organization → Embedded and cyber-physical systems**.

KEYWORDS

Millimeter Wave, Person Identification, Deep Learning

ACM Reference Format:
Ruofeng Liu, Tianshuo Yao, Ruili Shi, Luoyu Mei, Shuai Wang, Zhimeng Yin, Wenchao Jiang, Shuai Wang. 2024. Mission: mmWave Radar Person Identification with RGB Cameras. In *ACM Conference on Embedded Networked Sensor Systems (SenSys '24), November 4–7, 2024, Hangzhou, China.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3666025.3699340

1 INTRODUCTION

Millimeter-wave (mmWave) radar is becoming increasingly popular in human sensing tasks such as occupancy detection [29], trajectory tracking [42], action recognition, and pose estimation [38]. Discerning a person's identity is crucial for radar applications in contexts like security surveillance and personalized services. However, current mmWave identification methods require extensive data collection and training using prior instances of the person of interest in the same area [8, 39], which is not practical in dynamic real-world situations where individuals constantly change. For example, using mmWave radar to recognize an intruder becomes unfeasible when the individual was previously captured and labeled.

The practical limitation of radar motivates us to propose Mission (mmWave + vision), a *cross-modal* Re-identification (ReID) design: given a person detected by a mmWave sensor, the system can identify the same person in camera footage based on the consistency of gait characteristics across distinct sensors. By doing this, we exploit ubiquitously deployed cameras to extend the reach of radar identification towards previously unencountered individuals. This technology also presents the opportunity to extract valuable characteristics from associated images, such as age, gender, and clothing. This additional data can be harnessed to enhance the personalization of services in radar sensing applications while ensuring that fine-grained activities and sensitive behaviors remain confidential.

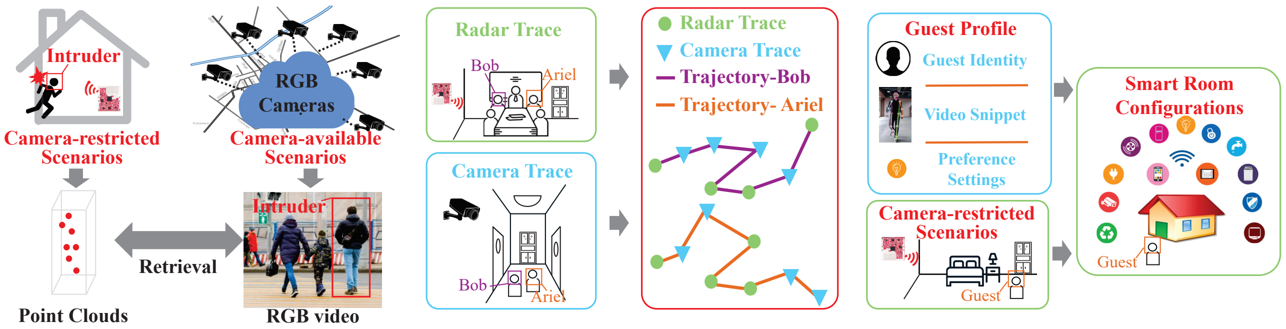
Despite the ReID problem being separately studied for the single modality (i.e., images [13, 24, 31] or mmWave [39, 42]), the cross-modal identification imposes new challenges stemming from the inherent modality discrepancy between images and radar in dimensionality and granularity. As Fig.1 depicts, the RGB cameras provide fine-grained details of individuals, but two-dimensional (2D) images lose critical depth information (e.g., height and stride length), while mmWave radar returns three-dimensional (3D) point clouds. However, point clouds of radar are very sparse and noisy due to the limited resolution of the low-cost radar. Consequently, the similarity between a person's images and radar point clouds cannot be evaluated directly.

To bridge the gap between camera and radar, we propose a novel cross-modal similarity estimation method that carefully leverages the inherent synergies between 2D images and 3D radar point clouds. Our key insight is that RGB images and radar point clouds

[*] Both Ruofeng Liu and Tianshuo Yao contributed equally to this work.
[†] Shuai Wang (shuaiwang_iot@seu.edu.cn) is the corresponding author.

SenSys '24, November 4–7, 2024, Hangzhou, China

ACM ISBN 979-8-4007-0697-4/24/11…$15.00
https://doi.org/10.1145/3666025.3699340



(a) Application 1: Surveillance.

(b) Application 2: Trajectory tracking.

(c) Application 3: Personalized experience.

Fig. 1. Appications of `Mission` .

of the same individual contain highly complementary features. For example, images excel at capturing intricate body shapes, whereas radar can precisely estimate absolute height. After exchanging these mutually complementary features, both modalities can be enhanced to accurately project 3D skeletons of an individual's gait. This explicitly brings them into a close alignment within the shared feature space. In contrast, amalgamating multi-modal data from distinct identities yields less accurate and potentially unreal skeleton estimation. This discrepancy of results arises from the intrinsic disparities in gait characteristics between different individuals, causing their features to disperse within the feature space.

To translate these insights into a practical solution, we present a novel end-to-end neural network consisting of several critical components. We start by extracting single modal gait features from radar point clouds and candidate images. A novel coordinated representation network is introduced to align features of different modalities into a coordinated feature space. This network incorporates a multi-modal non-local network (NLN) [34] with inter-modal mutual attention to exchange complementary feature among images and radar. Our design manages to integrate the most pertinent features from radar (e.g., depth) into RGB images and vice versa, while carefully handling their spatial-temporal alignment. With the coordinated features, a similarity estimator based on deep metric learning is adopted to find the identity of the most akin individual in the gallery of candidates.

To summarize, our contributions are as follows:

- To the best of our knowledge, we present `Mission`, the first cross-modal ReID design among commercial mmWave radar and RGB cameras, enriching emerging radar applications with ubiquitous cameras.
- We propose a new cross-modal similarity estimation method that judiciously uses multi-modal coordinated representation to address model discrepancy and practical challenges (e.g., spatial-temporal misalignment).
- We collect a multi-modal gait dataset of 58 volunteers across various scenes and viewpoints. The evaluations show our design achieves a top-1 accuracy of 85% and a top-5 accuracy of 90%, which outperforms the traditional cross-modal ReID baselines by 30%.

2 MOTIVATION AND CHALLENGES

2.1 Benefits of Cross Vision-RF ReID

The emergence of low-cost mmWave radar presents a compelling alternative to cameras in various human sensing tasks, bringing several benefits. As a radio frequency (RF) sensor, radar is more robust against poor lighting conditions (e.g., darkness and glare) than cameras. Moreover, it can operate inconspicuously behind the wall, causing fewer privacy concerns than cameras in scenarios where cameras are not allowed such as homes, hospitals, and confidential areas within office spaces. For example, hospitals in Hong Kong and Israel have deployed mmwave radar for contactless patient monitoring [1]. Radar technology is also increasingly used in smart buildings to provide human presence detection that is robust to occlusion and particles in the air [3].

As shown in Fig.1, this work aims at associating the person detected by radar with video footage of cameras deployed in public areas, which can find a broad set of applications including (i) **Surveillance**: When an intruder broke into an area monitored by radar (as shown in Fig.1(a)), the police can use the ubiquitous cameras to identify the intruder's image. Similarly, when the footage of a criminal is available, the radar infrastructure can also be used to detect if the person is hiding inside the camera-restricted areas; (ii) **Tracking:** In Fig.1(b), `Mission` enables seamless tracking of people when they walking between camera-allowed area (e.g., lobby) and camera-restricted (e.g., meeting room), which is important for trajectory analysis and contact tracing; (iii) **Personalized experience**: In a smart hotel (Fig.1(c)), each guest can upload the preferences (e.g., temperature, lighting, and favorite music) and a video snippet of gait, by which the radar can identify who is in the room and adjust the smart room accordingly.

This work primarily focuses on the technical feasibility of cross-modal ReID. However, it is essential to emphasize that similar to other biometrics recognition techniques [32], the application of `Mission` must be carefully managed to prevent potential misuse and address ethical concerns. In practice, several ethical processes can be followed to ensure that the technology is developed and used responsibly. First, data collection and storage must obey the Data Protection Laws and Surveillance Laws of the country or region such as the General Data Protection Regulation (GDPR) in Europe. For example, the data could be restricted to be used by authorized

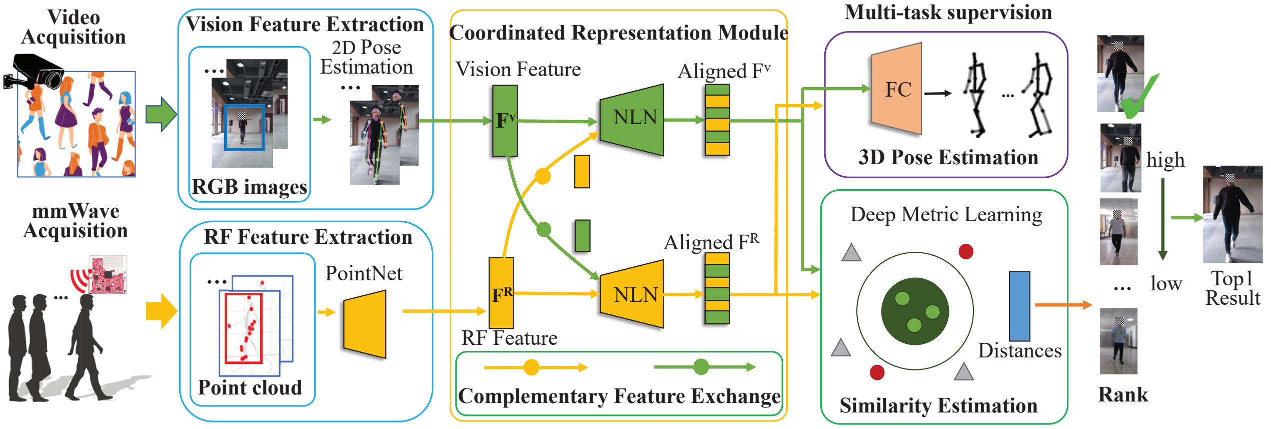


Fig. 2. System architecture.

personnel, such as law enforcement who are investigating an intrusion. Moreover, the application must undergo a review by an ethics board or Institutional Review Board (IRB) to make sure that the it aligns with ethical standards and minimizes harm to individuals. Finally, individuals (e.g., hotel guests) could be asked to provide informed consent before their data is collected.

2.2 Challenges

The crux of cross-modal ReID lies in addressing dramatically different characteristics between two modalities that make their similarity unable to be gauged directly. As Table. 1 summarizes, RGB images and radar point clouds suffer from modal discrepancy in dimension and granularity. The surveillance camera records RGB images containing rich vision information. Nonetheless, the human motions in the three-dimensional space are flattened by the camera into the two-dimensional image plane and therefore lose important depth information, leading to inherent ambiguities in 2D-to-3D mapping [17]. In contrast, mmWave radar uses frequency-modulated continuous waves (FMCW) that can estimate the range, velocity, and angles and generate the 3D point cloud of the target. However, the low-cost radar suffers from limited angular resolution (typically > 15° [2]), specular reflection of mmWave signal on the skin, and multi-path interference. This results in sparse and noisy 3D point clouds without fine-grained details of a person.

Sensor	Data	Dimen.	Granularity	Noise
Radar	Point cloud	3D	64 points	high
Camera	RGB image	2D	1280x720 pixels	low

Table 1: Modal discrepancy: Radar vs. Camera.

In addition, it is noteworthy that our cross-modal ReID is distinct from conventional multi-modal fusion problems [6, 29]. Specifically, video and radar point clouds are acquired at different times and disjointed locations. The query (radar) and candidates (RGB) are different instances of gaits, leading to several practical challenges. For instance, the camera and radar could capture the subject from distinct angles for different amounts of time. The subject could also exhibit a minor gait variation caused by the mood or environment [11]. As a result, the cross-modal similarity estimator has to robustly deal with temporal misalignment, diversity of viewpoints, and minor gait changes among different instances.

3 DESIGN

3.1 System Overview

Our system is designed to re-identify individuals detected by radar in videos captured by RGB cameras. We assume that the radar and camera are installed in non-overlapping areas and capture different instances of individuals' gait. As illustrated in Fig. 2, the system takes as input radar point clouds corresponding to the person of interest (denoted as a query) and a gallery of RGB video footage containing potential candidates. The RF feature extraction module denoises and clusters the sparse 3D point cloud of the person produced by radar. Gait features (denoted as RF features) are extracted from the pre-processed point cloud using PointNet [25]. Meanwhile, the vision feature extraction module detects and bounds the candidates present in the RGB videos. Then it uses a pre-trained TCMR [9] network to extract visual features from the bounded targets. These vision features represent the 2D pose information of the person in the video. The critical component is the coordinated representation module that aligns the heterogeneous RF and vision features in the coordinated feature space. Our coordinated representation design features 1) a complementary feature exchange mechanism that eliminates the modality discrepancy between vision and RF features and 2) an auxiliary 3D pose estimation task that guides the cross-modal representation learning procedure. Finally, deep metric learning estimates the similarity between the aligned RF feature of the query and the aligned vision feature of each candidate image. We sort the similarity and select the candidate with the highest similarity score as the final output.

The description of the design is organized as follows. We first introduce our high-level methodology and key insights in Section 3.2. The feature extraction of individual modality is given in Section 3.3. Section 3.4 explains the coordinated representation module in detail. Finally, 3D pose estimation and similarity estimation are discussed in Section 3.5.

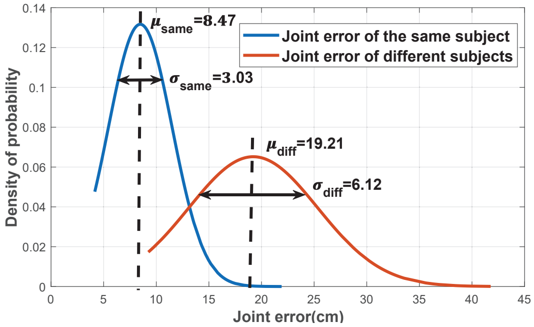


Fig. 3. Distributions of joint location errors for multi-modal 3D pose estimation.

3.2 Design Methodology

The most critical task of Mission is to find a multi-modal feature representation method that can effectively project RGB video and radar point cloud into the same embedding space for similarity estimation. As discussed in Section 2.2, RGB video and radar point cloud of a person's walking motion suffer from significant heterogeneity in terms of representation and physical meaning. This makes the conventional representation learning methods (e.g., metric learning) fail in our task. Our benchmark shows that directly training a model with deep metric learning results in an unacceptable ReID accuracy of 54% (more details in Section 5).

The key insight of our design is that we can explore the inherent synergy between RGB images and radar point clouds to address their modality discrepancy. More specifically, RGB image and radar point cloud of the *same person* contain highly *complementary* features of the person's gait. Radar point cloud reflects 3D characteristics of the target (e.g., height and stride length), which provide cues to eliminate the depth ambiguity in 2D RGB images. On the other hand, the fine-grained body shape obtained from the person's images can also enrich the sparse radar points and make mmWave features more robust to noise and sparsity. By exchanging the complementary information between two modalities, we can unify their feature representation for similarity estimation.

To this end, we introduce a novel coordinated representation design (details in Section 3.4) that extracts and integrates the beneficial radar features into original image features such that the augmented features of 2D images can accurately predict the 3D pose of the person. Simultaneously, it incorporates important image features into radar point clouds to enhance radar 3D pose estimation. The processes enrich and project both modalities into an identical embedding space that represents 3D walking poses. Our experiment shows feature exchange dramatically reduces joint location error by 50% compared to methods with single-modal information. This makes the features of identical subjects well-clustered in the embedding space. Moreover, integrating the 3D human model for gait recognition makes our design robust to the viewpoint variation between radar and camera. This is critical for Mission because radar and camera deployed in the disjointed locations could capture the person from different perspectives. To guide the coordinated representation network to extract features that represent 3D human models, we incorporate 3D pose estimation as an auxiliary task (details in Section 3.5).

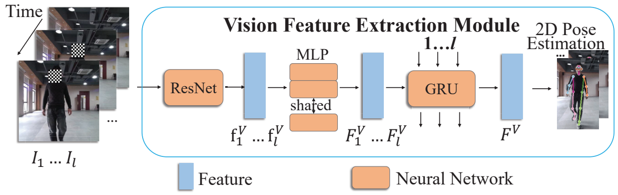


Fig. 4. Vision Feature Extraction Module.

In addition, we observe that the RGB and mmWave features of *different identities* are inherently *conflicted* rather than complementary, and performing the aforementioned multi-modal coordinated representation leads to very random results. To demonstrate this observation, Fig.3 compares the 3D joint location error distribution in both complementary and conflicted cases. Integrating RGB and radar features from different identities exhibits significantly larger average errors (17.9cm vs. 8.4cm) and error variances (5.2cm vs. 3.1cm). The amalgamated features cannot be clustered well in the embedding space. This unique observation further helps us to distinguish people with different identities.

3.3 Multi-Modal Feature Extraction

This section introduces the feature extraction from individual modalities (i.e., RGB images and Radar point cloud).

3.3.1 *Vision Feature Extraction Module.*

To extract pertinent two-dimensional gait such as body shape and gait cycle from the provided images, we employ a DNN illustrated in Fig. 4, which comprises two key components: ResNet and GRU. In particular, when presented with a sequence of l RGB frames denoted as $I_1, \dots, I_l$, we harness a ResNet-50 network [16] that has been pre-trained according to [19]. The objective was to derive static features from each frame, yielding representations $f_1^V, \dots, f_l^V$, where the superscript V signifies the visual modality and $f_*^V \in \mathbb{R}^{2048}$. Notably, the ResNet's weights were shared across all frames to maintain consistency. To align with the features obtained from the mmWave modality, we employ a shared-weight Multi-layer Perceptron (MLP) to condense the dimensionality of the derived features. This transformation resulted in $F_*^V = \text{MLP}(f_*^V)$, where $F_*^V \in \mathbb{R}^{256}$. Given the intuition that a frame could benefit from prior pose information, we adopt a similar strategy as presented in TCMR [9]. Once the static features of all input frames are computed, we use a time encoder composed of bidirectional Gated Recurrent Units (GRUs) that encode temporal features into the current frame. This approach acknowledges the potential influence of past pose information on subsequent frames, contributing to a comprehensive temporal understanding. The output of the GRUs (denoted as F^V) is the ultimate output of the vision feature extraction module. F^V represents the 2D gait characteristics. This feature can be used to estimate 2D pose in the images but cannot accurately estimate 3D pose due to loss of depth information (mentioned in Section 2.2). We will discuss in Section 3.4 that the coordinated representation module will augment F^V into 3D gait features with the complementary depth information extracted from radar point clouds.

3.3.2 *RF Feature Extraction Module.*

We adopt PointNet [38] to extract features from 3D Radar point cloud, which consists of a Base

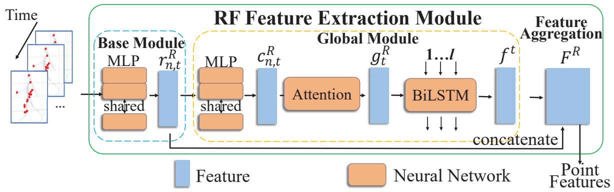


Fig. 5. RF Feature Extraction Module.

module and a Global module (depicted in Fig. 5). Base module operates on individual points, denoted as $p^R_{n,t} = x^R_{n,t}, y^R_{n,t}, z^R_{n,t}, i^R_{n,t}, v^R_{n,t}$ within the point set S_t of the t^{th} frame. Each point is independently processed using a shared-weight MLP. Here, x, y, z, i, and v respectively stand for the coordinates, intensity, and velocity of the point, while the subscript n signifies the point's index within the set and superscript R denotes Radar modality. The output of this encoding, referred to as the "point feature", is a high-level representation of each point and is denoted as $r^R_{n,t}$= MLP$(p^R_{n,t}; \theta_r)$, where θ_r is the parameter set of the MLP layers.

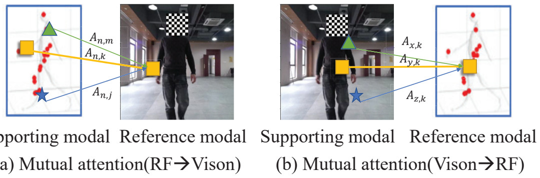
Global module aggregates the point features from each frame into a singular frame feature, capturing comprehensive characteristics of the gait (e.g., height and center of mass). Given the point feature $r^R_{n,t}$ produced by the Base module for a specific frame, we employ an MLP to transform it into a more refined representation, denoted as $c^R_{n,t}$ = MLP$(r^R_{n,t}; \theta_c)$ and compile these point-level representations into a holistic frame feature using attention. Let $A()$ denote the attention function that computes scores for each point. The frame feature g^R_t can be expressed as follows:

$$g^R_t = \sum_{n=1}^{N} A(c^R_{n,t}; \theta_a) \times c^R_{n,t} \tag{1}$$

where N is the number of points in t^{th} frame and θ_a is the parameter of attention function. Note that the original PointNet uses a max-pooling layer to aggregate frame features, which causes severe information loss due to the sparsity of mmwave point clouds and thus is replaced by attention. We feed g^R_t the multi-layer Bidirectional Long Short-Term Memory (BiLSTM) to further incorporate the temporal relationship between consecutive frames. The final global feature is denoted as f^R_t=BiLSTM$(f^R_{t-1}; g^R_t; f^R_{t+1}; \theta_f)$, where f^R_{t-1}, f^R_{t+1} is the global representation of the adjacent frame and θ_f is trainable parameters to in BiLSTM. Finally, we concatenate the global feature f^R_t to each point feature $r^R_{n,t}$ to obtain augmented point feature $F^R = concatenate(r^R_{n,t}, f^R_t)$. F^R is referred to as "point feature" for short in the following sections, which will be the input of coordinated representation in Section 3.4.

3.4 Coordinated Representation Module

The multi-modal feature extraction modules obtain two sets of feature vectors for RGB and radar respectively. As mentioned in Section 3.2, a multi-modal feature coordinated representation holds the potential to mitigate modal discrepancies by exchanging complementary features between different modalities. This section delineates the high-level feature exchange mechanism followed by details of coordinated representation network design.



(a) Mutual attention(RF→Vison) (b) Mutual attention(Vison→RF)

Fig. 6. Pixel-point feature exchange, where $A(n, m)$ represents correlation between the n^{th} point in a radar frame and the m^{th} pixel in an image frame.

3.4.1 *Feature Exchange Mechanism.*

As discussed in Section 3.2, images and radar encompass complementary gait features. To harness this complementarity effectively, we design an inter-modal mutual attention mechanism that enables fine-grained feature exchange between two modalities. The feature exchange process is visually illustrated in Fig. 6.

For RF⟶Vision exchange, each image pixel engages in inter-modal attention to compute its correlation with every radar point. The outcome of this attention process is the integration of the most pertinent point features into the pixel's representation. Correspondingly, in the Vision⟶RF exchange direction, the inter-modal attention mechanism focuses on the most relevant image pixels to extract specific features (such as the shape of the specific body part) and incorporate them into the corresponding radar point. Recall that a unique challenge in our design (compared to conventional multi-modal sensor fusion) is that radar and camera capture different instances of the gait. Due to the lack of strict temporal synchronization among radar and image frames, the most relevant information of inter-modal attention might be distributed across various frames of the other modality. To address this challenge, we adopt the Non-local neural network (NLN) [34] structure during feature exchange. This structure effectively handles long-distance dependencies, enabling each pixel to gather valuable insights from any radar frame and vice versa. We integrate NLN with inter-modal mutual attention, which overcomes temporal asynchrony and ensures that the most relevant cross-modal information is effectively harnessed in coordinated representation.

3.4.2 *Coordinated Representation Network.*

As depicted in the middle of Fig. 7, in order to achieve cross-modal coordinated representation, the vision and RF features undergo inter/intra-modal attention and two parallel Non-local networks (NLN) for Vision⟶RF feature exchange and RF⟶Vision feature exchange respectively. Take the RF⟶vision exchange process for example. The RF features F^R are fed into the vision NLN (the green arrow) with its spatiotemporal correlation with the vision features F^V calculated using inter-modal attention (red dot). The details of inter-modal attention in an NLN block are depicted on the right of Fig.7. We linearly transform a vision feature F^V and an RF feature F^R into a query and key in an embedding space using matrices $W^{R\rightarrow V}_q$ and $W^{R\rightarrow V}_k$. The query and key are then dot-product and normalized by a non-linear function σ(e.g., softmax) to obtain the correlation between the vision and RF feature. This inter-modal attention produces a matrix $a^{R\rightarrow V}$ that estimates the correlation between each pair of radar points and image pixel, enabling feature exchange at a fine granularity. Furthermore, to accommodate the long-distance dependency of

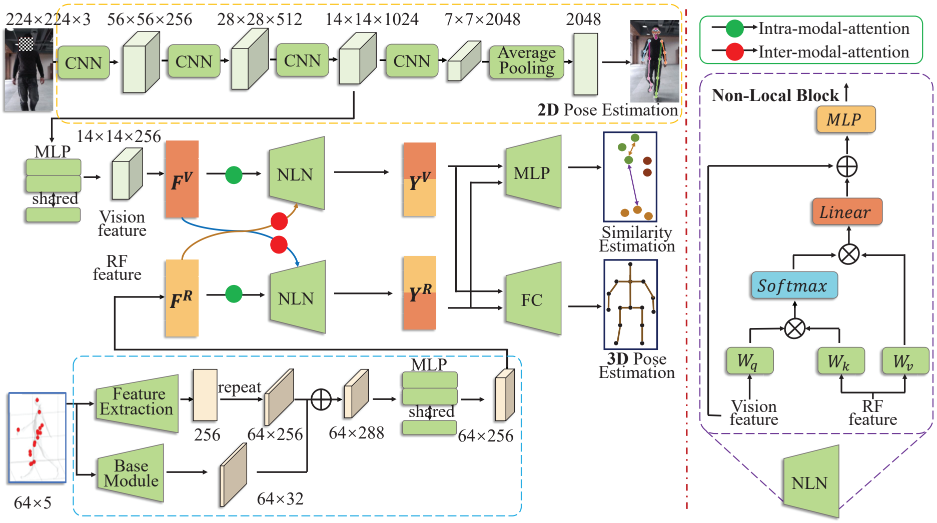


Fig. 7. Cross Vision-RF Coordinated Representation Network.

two modalities, the correlations are calculated for radar points and image pixels across different frames. The inter-modal attention matrix for the Vision→RF feature (denoted as $a^{V\rightarrow R}$) is derived similarly. The formal definition of correlation matrices is given in equation 2.

$$\begin{aligned} a^{R\rightarrow V} &= \sigma(W_k^{R\rightarrow V} F^R (W_q^{R\rightarrow V} F^R)^T) \\ a^{V\rightarrow R} &= \sigma(W_k^{V\rightarrow R} F^V (W_q^{V\rightarrow R} F^R)^T) \end{aligned} \tag{2}$$

In addition to inter-modal attention, we also introduce intra-modal self attention to calculate the correlation between features within single modal. Take the representation process of Vision as an example, the vision NLN takes a sequence of vision features as the input and performs intra-modal self attention (green dot in Fig.7) to aggregate features between different frames within a gait cycle. The advantage of doing so is that it generates a global context for better expressing the gait features of the target. As Eq.3 illustrates, F^V is linearly transformed into embedding space $W_k^V F^V$ and $W_v^V F^V$, which represents the key and value of F^V respectively. Then, these feature vectors are dot-producted and normalized by a non-linearly function σ (e.g. softmax) to get the intra-modal attention matrix a^V. a^V estimates the spatio-temporal correlation between frames throughout the gait cycle. A similar process is also implemented in RF modal, which outputs the intra-modal attention matrix a^R.

$$\begin{aligned} a^V &= \sigma[W_k^V F^V (W_q^V F^V)^T] \\ a^R &= \sigma[W_k^R F^R (W_q^R F^R)^T] \end{aligned} \tag{3}$$

With the inter-modal and intra-modal attention matrices, we extract and exchange the complementary information among vision features and RF features such that they are aligned in the same feature space for similarity estimation. To incorporate complementary vision feature into the RF feature, we linearly transform the vision feature F^V embedding space with a matrix $W^{V\rightarrow R}$ and then multiple it by the inter-modal attention matrix $a^{V\rightarrow R}$. This process essentially select and aggregate the most relevant vision features based on their correlations with a specific Radar point that are estimated by inter-modal attention. In addition, the RF feature F^R is multiplied by with intra-modal attention matrix a^R that aggregate features across various Radar frames. The aggregated inter-modal and intra-modal features are concatenated and the result is linearly transformed by W_y^R and combined with the original feature F^R by element-wise addition, which finally produces the augmented RF feature (denoted as Y^R). Corresponding, the vision feature F^V is augmented with the complementary RF features following a similar process, which augmented vision feature Y^V. This cross-modal feature coordinated representation is formally given in equation 4.

$$\begin{aligned} Y^V &= W_y^V [a^V W_v^V F^V ; a^{R\rightarrow V} W_v^{R\rightarrow V} F^R] + F^V \\ Y^R &= W_y^R [a^R W_v^R F^R ; a^{V\rightarrow R} W_v^{V\rightarrow R} F^V] + F^R \end{aligned} \tag{4}$$

3.5 Multi-task Supervision

As discussed in Section 3.4, our coordinated representation structure facilitates the exchange of complementary gait features between two modalities. During model training, it is crucial to guide the network to efficiently learn and transfer these complementary features. Additionally, since our primary objective is person reidentification (ReID), the model must not only capture multi-modal gait features but also distill the most discriminative elements from these features for accurate identification. This section introduces a multi-task learning framework designed to achieve both of these goals simultaneously.

3.5.1 Pose Estimation Network. We discussed in Section 3.2 that by exchanging the complementary information between RGB images and radar point cloud, we can obtain a unified feature representation that accurately represents 3D human gait. To assist the neural network in learning this unified feature representation during model training, we incorporate the 3D pose estimation task. It uses the coordinated representation (obtained in Section 3.4) to predict a sequence of 3D skeletons described by the 3D coordinates of key joints of the person during walking. We choose 3D skeleton estimation as an auxiliary task to supervise feature extraction and coordinated representation based on the domain knowledge about gait. First, 3D

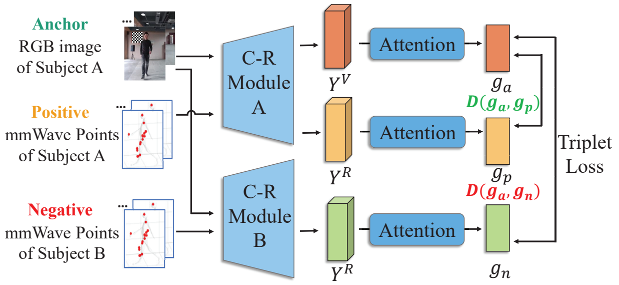


Fig. 8. Similarity Estimation via Deep Metric Learning. C-R Module refers to the coordinated representation module, which is responsible for the coordinated representation of both positive and negative samples with the anchor sample.

pose is widely used as a highly effective representation of human gaits for person identification [26]. In addition, 3D pose is invariant to viewpoint of sensors. By projecting both RGB images and radar point clouds into the 3D pose representation, the model can handle the viewpoint difference between radar and camera that are deployed in the non-overlapping areas (a critical challenge discussed in Section 2.2). More specifically, for both augmented vision and RF features (i.e., Y^V and Y^R), we use a fully-connected network (FC) to predict human skeleton point location $S = \{(x_i, y_i, z_i)|i = 1, ..., M\}$, where x_i, y_i, z_i represent the coordinate of the i^{th} point and M is the number of body skeleton points. Note that pose estimation and ground truth of pose are only required during the training stage. For the inference, Y^V and Y^R are only used by deep metric learning networks for ReID. The Loss function of pose estimation is to minimize the error between predicted and ground truth positions of skeleton joints. Given that the human skeleton contains M joints, we minimize the Mean Squared Error (MSE) loss:

$$L_p = \frac{1}{M} \sum_{m=1}^{M} ||\hat{p}_m - p_m|| \tag{5}$$

where $\hat{p}_m, p_m$ are the predicted position and corresponding ground truth of m^{th} joint and $|| * ||$ denotes the L2-norm.

3.5.2 Similarity Estimation via Deep Metric Learning. Due to the heterogeneity of the two modalities, it is difficult for traditional metric learning to project data from different modalities into the same embedding space and cluster them closely. However, our coordinated representation module effectively eliminates the modal discrepancy, so we perform deep metric learning on the augmented vision and RF features rather than directly performing on the extracted features of individual modality. As shown in Fig.8, we feed Y^V and Y^R (the output of coordinated representation module) into a deep metric learning network. Deep metric learning focuses on extracting high-level gait representations that are robust to temporal misalignment between camera and radar data, variations in frame lengths, and minor differences in gait instances (as discussed in Section 2.2). Note that Y^V and Y^R contain the features of the entire gait cycle. To highlight the most discriminative features for ReID, we use an attention module to calculate the importance of each frame and aggregate the features in the frames among the entire cycle by a weighted summation.

Loss of similarity estimation. The deep metric learning minimizes the distance between the features from the same identity while maximizing the distance between features from different subjects. As shown in Fig.8, we employ triplet loss [27] as the loss function, where each input of the triplet loss function is a triple consisting of anchor, positive, negative instances which are denoted as $< g_a, g_p, g_n >$. The objective function is designed as follows:

$$L_s = max(D(g_a, g_p) + margin - D(g_a, g_n), 0) \tag{6}$$

where $D(g_a, g_p)$ is the euclidean distance between embedding of anchor and positive samples. $D(g_a, g_n)$ is the distance between embedding of anchor and negative samples. *margin* is a hyper-parameter, and by adjusting the value of *margin*, the distance between the anchor and the positive samples can be ultimately reduced while the distance between the anchor and the negative samples is increased. The total training objective is:

$$L = \alpha L_p + \beta L_s \tag{7}$$

where α and β are hyper parameters that adjust the contribution of different tasks. The whole framework is trained end-to-end.

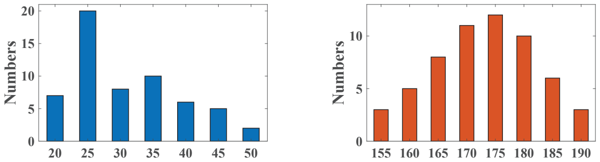
4 IMPLEMENTATION

This section presents the implementation of our design including the experiment devices and data acquisition process. We released `Mission` at https://github.com/EverRaynor/Mission.

4.1 Experiment Platform

4.1.1 mmWave Radar Platform. We utilized the commercial radar IWR6843-BOOST [2] for data acquisition. It operates within the frequency range of 60 GHz to 64 GHz, corresponding to a wavelength of 4mm. The device consists of three transmitting antennas and four receiving antennas, collectively providing a 60-degree field of view (FoV) in both azimuth and elevation, with an angular resolution of approximately 15 degrees. We use the standard frequency-modulated continuous wave (FMCW) processing chain provided by TI to generate a 3D point cloud. We include the detailed configuration parameters of the radar device below for reproducibility. The radar transmits 10 frames per second, each comprising 32 chirps. Each chirp begins at 60.065 GHz, with a bandwidth of 3194.88 MHz. The frequency slope is fixed at 12.5 MHz/microsecond.

4.1.2 Camera Platform. RGB image data of gaits are collected with Azure Kinect DK, equipped with a 12-megapixel full HD camera. We set the camera frame rate to 15 FPS and the image resolution to 720p, which simulates the common configuration in re-identification scenarios. In addition, we use the depth sensor of Kinect to obtain the ground truth joint positions for 3D pose estimation. Body Tracking SDK [4] can track multiple humans with Azure Kinect DK, returning 32 joint skeletons for each human in the field of view. It is noteworthy that the depth camera in our design is only used in the training stage and is not required during the operation (inference). In our evaluation, depth features are removed from the data. The trained model can estimate the similarity between 2D RGB images (without depth information) and mmWave point clouds.



(a) Age groups.

(b) Height groups.

Fig. 9. The cohort stats regarding the age and height.

4.2 Data Acquisition

We recruited 58 participants (36 males, 22 females) aged 19 to 51 years, with heights ranging from 152 cm to 187 cm, for data collection. Figure 9 presents the cohort statistics for age and height, both of which significantly influence gait patterns. Our experiments received approval from the Institutional Review Board (IRB).

Mission assumes that radar and camera systems are installed in separate scenes, with some areas designated as camera-restricted and others as camera-allowed. To study the re-identification across different scenes, we conducted data collection in three different scenarios: classroom, corridor, and hall, as shown in Figure 10. Our data collection equipment was positioned at one end of the test field, with the mmWave radar and camera placed at heights of 0.85 meters and 0.86 meters, respectively. Volunteers walked towards the equipment from the opposite end, with each participant walking at least 6 meters in each experiment.

To ensure that the data contain a variety of viewpoints for camera and radar, participants were instructed to follow various walking routes (indicated by yellow dashed lines in Fig.10(a) and 10(b)). This results in varying angles for sensors to view the subject (denoted as view angles) which range from 0° to 60°. This approach enabled us to capture instances of a person from varying perspectives, simulating real-world sensing scenarios. Due to natural variations in walking speed, the duration of radar capture varies between 4 to 5 seconds for each individual. To account for the variation in the data collection, each participant repeated a walking task 20 times. We integrated the multi-modal data collection into the Robot Operating System (ROS).

5 EVALUATION

In this section, we discuss the performance evaluation of our system. We begin with evaluation methodology, which covers the training and testing procedures, evaluation metrics, and competing approaches (Section 5.1). The overall performance results are presented in Section 5.2, followed by detailed evaluations of each critical design component in Section 5.3. Finally, we analyze the impact of various factors on system performance in Section 5.4.

5.1 Evaluation methodology

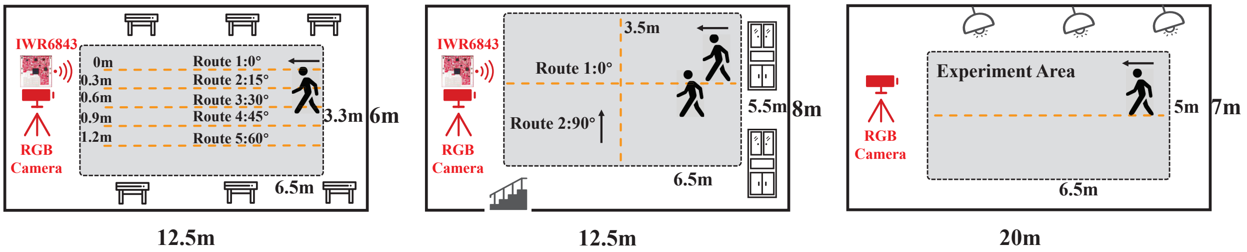
5.1.1 Overview. We utilized the configuration detailed in Section 4 to collect a comprehensive dataset consisting of 2000 records of radar and RGB data from 58 different subjects. For a systematic assessment of identification performance, we constructed two separate datasets. The first dataset comprises over 1800 records, each corresponding to a single individual present in the field of view at a given time. To explore scenarios with multiple subjects, we also assembled a multi-person dataset, consisting of 200 records captured when two individuals are simultaneously within the FoV.

During the evaluation, these records will be carefully separated to comprehensively produce various cross-modal re-identification scenarios. For example, we separated training and testing samples according to their identity to examine if Mission can identify unseen subjects during the model training (Section 5.4.2). Moreover, to test the feasibility of serving cameras and radar installed in disjointed areas, we evaluated situations where the query and candidates were acquired from different scenes (Section 5.4.3). Finally, we studied the impact of varying viewpoints of both radar and camera (Section 5.4.1), the size of the gallery (Section 5.4.6), the number of records of each subject (Section 5.4.4), and the duration of the snippet (Section 5.4.5) on the ReID accuracy.

5.1.2 Model Setting and Model Training/Testing. The details of the model, training, and testing procedure are as follows. For RF feature extraction, we implement MLPs in PointNet, the layer sizes of which are (5,12,24,48,64). For vision feature extraction, we resize the image from 1080 × 720 to 224 × 224 for training. We implement the NLN in coordinated representation using NONLocalBlock [34] with dot-product. In alignment with the prevalent approach adopted by various cross-modal ReID studies in computer vision [5, 15, 19], we use 75% of the data records collected from each subject for model training, and the remaining 25% serve as the testing set. In addition, we evaluated the effectiveness of the design in unseen subjects by dividing training and testing data according to subjects' identities. The learning rate is set to 0.0002 and the batch size is 16. The number of training epochs is 50000. The hyper-parameter *margin* assigned to the loss function in Equation 6 is set to 0.3, and α, β mentioned in Equation 7 is set to 1 and 2 respectively. We implement our deep learning model in PyTorch and train the model with NVIDIA RTX 3090.

5.1.3 Evaluation Metrics. Our evaluation adopts the cumulative matching curve (CMC), a widely adopted metric in ReID studies [5, 22]. Specifically, for each radar query, we calculate the similarity between the query and every candidate RGB record. We report the top-N accuracy, which is defined as the percentage of test cases where the RGB record of the target person is ranked among the top N positions among all the RGB records in the test. N varies from 1 to 10 in our evaluation given the number of our volunteers. In addition, we jointly train the ReID network with a pose estimation network to extract complementary gait features. To evaluate the effectiveness of the strategy, we examine the accuracy of pose estimation measured by Average Joint Localization Error (AJE) which is the average Euclidean distance between the predicted skelection key points (i.e., joint locations) and their ground truths.

5.1.4 Baseline. Mission is the first cross-modal ReID framework designed for RGB camera and radar modalities. To evaluate it, we introduce several baseline comparisons. First, we adopt a recent cross-modal ReID approach originally developed for RGB-D and radar data and apply it to our RGB-radar dataset. Additionally, to further highlight the effectiveness of our approach in tackling cross-modal challenges, we implement four representative (Re)ID methods from recent literature, which were originally designed for either mmWave or RGB. These baselines individually extract



(a) Scene1: Classroom.

(b) Scene2: Corridor.

(c) Scene3: Hall.

Fig. 10. Implementation scenarios.

features from RGB and radar using various network models and assess similarity using classical deep metric learning techniques.

HMR + Cross Radar/RGBD (HRD). We implemented [5] as a cross-modal baseline in our evaluation. This work was originally designed for cross-modal ReID between mmWave and RGB-D images. To fill the gap between RGB and RGB-D. We first predict the depth images from RGB images using the human mesh estimation method (HMR [17]) and then the predicted depth image and radar point cloud are processed by the end-to-end pipeline of [5] to estimate the similarity.

PointNet + GaitPart + LSTM(PGL) [5, 8, 13]. We use PointNet [25] and GaitPart [13] to extract features from radar points and RGB images respectively. Both RF and vision features are fed into LSTM to further exploit gait temporal patterns. We perform deep metric learning with a triplet loss to project features of different modalities into the same embedding space and estimate their similarity.

Voxelization + 3DCNN + GaitPart + LSTM (VCGL) [42]. We replace PointNet in PGL with voxelization + 3DCNN [42], a different feature extraction model for radar points. LSTM and average pooling are then utilized to obtain the final embedding. We also use deep metric learning with a triplet loss to achieve similarity estimation.

DGCNN + GaitPart + LSTM(DGL) [35]. We implement another radar feature extraction method based on DGCNN, a graph CNN that extracts features from radar points. It consumes the point cloud directly and applies the proposed EdgeConv which takes k adjacent points as graph structure to extract local features. The rest part is the same as VCGL.

PointNet + TCMR + LSTM(PTL) [9]. We employ PointNet and TCMR in Section 3.3 for RF and vision feature extraction. Distinct from our cross-modal feature coordinated representation, the baseline directly performs metric learning on RF and vision features with a triplet loss, attempting to align the heterogeneous features into the same embedding space and estimate their similarity.

5.2 Overall Performance

This section presents the overall ReID accuracy of Mission. We report the single-person accuracy (i.e. only one subject appears in the radar FoV) and multi-person accuracy (i.e. multiple subjects appear in the radar FoV).

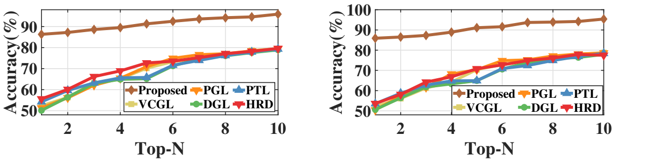
5.2.1 *Single-person ReID accuracy.*

As Fig.11(a) shows, our system achieves 85.42% top-1 accuracy, 87.65% top-3 accuracy, and 90.31% top-5 accuracy out of 58 volunteers in single-person scenarios. Our method significantly outperforms the cross-modal ReID baseline (**HRD**) with a Top-1 accuracy of 55.6%. This is mainly because the baseline directly estimates the depth mesh of the subject from RGB images. Suffering from depth information loss, the estimation introduces non-trivial errors. In contrast, our method effectively exploits the complementary features from mmWave radar to enhance the accuracy of 3D gait estimation. In addition, our approach achieves at least 32% higher top-1 and 14% higher top-5 accuracy than the baselines transplanted from single-modal ReID. The results indicate that our design effectively mitigates the modal discrepancy between 2D RGB and radar point cloud.

Recall that the coordinated representation learning is guided by 3D pose estimation. To further illustrate the effectiveness of the strategy, we also analyze the result of the auxiliary task. We compare the average joint localization errors (AJE) with classic pose estimation methods using single modality (mmMesh [38] for radar and TCMR [9] for camera). As Table.2 shows, the original TCMR causes a 14.02cm error due to the depth ambiguity of RGB image while our method brings it down to 4.62cm, which is attributed to the inter-modal attention design that incorporates 3D cues into the vision feature. It also improves the robustness of radar pose estimation, limiting the error within 3.9cm. The findings indicate that Mission adeptly facilitates the exchange of complementary features between 2D images and 3D radar point clouds. This process effectively aligns heterogeneous features from different modalities into the same embedding space.

5.2.2 *Multi-person ReID accuracy.*

As Fig.11 shows, Mission retains robustness in multi-person ReID with a top-1 accuracy of 84.77%, top-3 accuracy of 87.01%, and top-5 accuracy of 89.64%. This outcome substantiates that our method can effectively manage real-world scenarios involving multiple individuals within radar Fields of View (FOVs). It further underscores the unique advantage of mmWave ReID over the WiFi-based solution [21], particularly in its ability to recognize multiple individuals simultaneously in the same scene.



(a) Single subject.

(b) Multiple subjects.

Fig. 11. Overall performance (Cumulative Matching Curve).

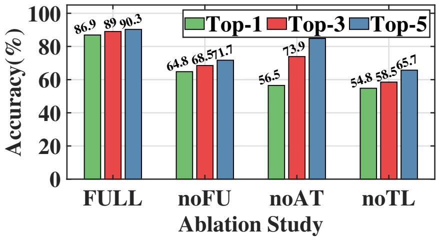


Fig. 12. Effectiveness of design components.

5.3 Effectiveness of Design

5.3.1 Ablation study. In Section 3, several critical designs were introduced. To demonstrate their effectiveness, we measure the performance of the system when specific components are removed. We conduct experiments in the following four settings and ReID accuracy is shown in Fig.12.

Full version (Full): All the designs in Section 3 are enabled.

w/o coordinated representation module (noFU): The similarity is calculated without feature exachange using the coordinated representation module (Section 3.4). This setting aims to emphasize the significance of inter-modal attention and NLN in eliminating modality discrepancy between radar and RGB.

w/o auxiliary task(noAT): This setting removes the supervision of auxiliary task (i.e., pose estimation).This setting shows the effectiveness of using 3D pose prediction to assist the network in learning how to extract and exchange meaningful gait features.

w/o triplet loss(noTL): The training loss is replaced by contrastive loss with a two-tuple input. Specifically, we train the network with a pair consisting of one sample from each modality (denoted as <r,v>). The loss function used in contrastive loss is defined as:

$$L_c = yd^2 + (1 - y)max(margin_c - d, 0)^2 \tag{8}$$

where $margin_c$ is a hyper-parameter, d represents the distance of feature embedding of r and v. y is the label indicating whether r and v belong to the same identity($y = 1$) or not($y = 0$). This setting aims to evaluate the importance of using triplet loss during the training.

Fig.12 presents the results. As anticipated, the comprehensive version of Mission delivers the best performance among all configurations, thereby validating the effectiveness of various designs in enhancing overall performance. In the absence of a coordinated representation module, the top-1 accuracy of **noFU** declines to 64.8%, underscoring the importance of mutual attention across modalities. Moreover, the most significant reduction in accuracy is observed in **noAT** (a loss of 30.4% in top-1 accuracy). Therefore, we surmise that 3D pose estimation plays a pivotal role in our design, facilitating complementary feature extraction and feature exchange between RF and vision. Lastly, the decrease in accuracy in **noTL** indicates that the triplet loss is more effective than the contrastive loss.

Models	mmMesh	TCMR	Mission
AJE(cm)	4.432	14.021	**3.902(RF)/4.624(Vision)**

Table 2: Overall performance of pose estimation.

5.4 Sensitivity Analysis

5.4.1 Impact of view angles. Mission provides ReID for radar and camera installed in the non-overlapping areas. Two sensors might capture people from different view angles, causing significant challenges for cross-view 2D image ReID in conventional camera solutions [7]. Mission harnesses the complementarity of radar to augment 2D vision into the enhanced features that represent 3D skeletons, which are more resilient to the impact of view angles. We evaluate the performance of our design from various view angles.

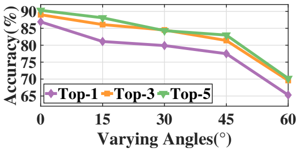


Fig. 13. Various view angles of RGB.

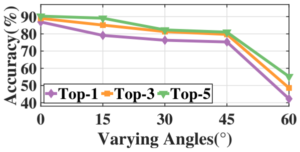


Fig. 14. Impact of the number of gait frames.

Fig. 15. Evaluation of various view angles.

View angles of RGB cameras. As depicted in Fig.10(a), to generate RGB images from various viewpoints, volunteers are instructed to traverse along a set of pathways with offsets ranging from 0 to 1.2 meters from the camera-aligned line, referred to as the mid-line. This procedure modifies the angle between the volunteers and the device from 0° to 60° at intervals of 15°. Conversely, volunteers proceed directly toward the mmWave radar along the mid-line. Fig.13 illustrates that our methodology maintains a 75% top-1 and 81% top-5 accuracy at a substantial angle of 45°. This result shows that our model effectively extracts gait features that are invariant to the viewpoint of the camera. The accuracy descends to 65% (top-1) when the angle escalates to 60°. This is because when volunteers walk with a large offset from the mid-line, they are no longer entirely within the camera's Field of View (FoV).

View angles of mmWave radars. We further assess the influence of radar viewing angles. In this experiment, volunteers are directed to approach the radar along a variety of trajectories with offsets ranging from 0 to 1.2 meters from the mid-line. It should be noted that in this experiment, volunteers walk directly toward the RGB camera. Fig.14 demonstrates sturdy performance across a range of view angles from 0° to 45°. This again proves that our coordinated representation and the supervision using 3D pose estimation effectively extract 3D gait features that can deal with viewpoint variation. However, when the angle exceeds 60°, the radar experiences self-occlusion of the subject in addition to a significant decline in angular resolution and signal strength, resulting in a performance drop. This outcome suggests that the deployment of multiple radar sensors to capture observations from assorted view angles could potentially enhance overall accuracy.

5.4.2 Unseen-subject performance. Mission doesn't require collecting and labeling mmWave data from the individuals to be recognized in advance. For example, it recognizes intruders that don't appear during model training. This benefits from our designs. We adopt deep metric learning to learn a metric such that the subjects with similar gaits are close in the feature space while different gaits are distant. Therefore, the model does not memorize the gaits of the people in training data but uses the training data to learn the optimal metric to estimate similarity. Moreover, we use 3D pose estimation as the auxiliary task which makes the network more generalizable to the varying environments and sensor setup. To test the performance of unseen subjects during training, we divide the dataset into two parts based on volunteers’ identities for training and testing respectively, and the volunteers’ identities are not duplicated to simulate real-world scenarios. Specifically, during the training phase, we selected all records from 40 out of 58 volunteers as the training set, while the remaining 18 out of 58 volunteers were designated as the test set. This approach ensures that Mission encounters previously unidentified volunteers during the testing phase. As shown in Fig.18, the top-1 accuracy in unseen-subject scenarios maintains 84.9%, indicating robust generalization of Mission to novel subjects.

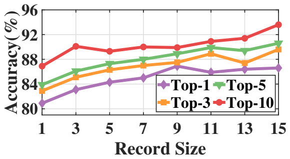


Fig. 16. Impact of RGB records per subject.

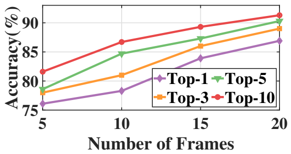


Fig. 17. Impact of the number of gait frames.

5.4.3 Cross-scene performance. Mission exploits geographically dispersed cameras, which potentially capture candidates in a variety of environments. To assess the model’s generalization capability to diverse environments, we juxtapose the results with RGB videos captured in three distinct scenes as depicted in Fig.10. Specifically, during the training phase, we utilize 70% of the data collected from Scene 1 as the training set. In the testing phase, we select the remaining 30% of the collected radar records (excluding those in the training set) from Scene 1 as the query set, while the gallery set comprises RGB records from across three different scenes. We randomly select a radar record from the query set as the query and build different RGB galleries for ReID based on the scene. For Scene 1, the remaining 30% of the collected RGB records were used as the gallery, representing intra-scene re-identification. Scenes 2 and 3 employed all RGB records collected from Scene 2 and Scene 3, respectively, representing cross-scene re-identification.

As illustrated in Fig.19, alterations in scenes exert a negligible impact on the outcomes, with a mere 4% deviation in top-1 accuracy. Given that Mission primarily focuses on the human body’s motion across various scene backgrounds, it consistently maintains high performance in cross-scene experiments. This evaluation implies that our design maintains robustness amidst environmental changes and our feature exchange mechanism effectively aggregates individual features that remain unaltered by the environment.

5.4.4 Impact of the number of records. In real-world scenarios, the public camera network may capture an individual an indefinite number of times. We adjust the number of records for each individual in the RGB database and scrutinize the impact on accuracy. Fig.16 illustrates the accuracy ranging from top-1 to top-10. We discern that an increase in the number of RGB records for each candidate marginally enhances the ReID accuracy. This can be attributed to the larger number of records, which typically helps mitigate the randomness of gait, such as minor variations in step length. As a result, gait recognition becomes more robust.

5.4.5 Impact of the number of gait frame. A typical gait cycle lasts approximately 1 second. However, there can be extreme cases where only fragments of gait cycles are captured, for instance, due to obstructions. To investigate the impact of duration, we conduct our evaluations repeatedly while altering the number of frames from 5 to 20, corresponding to 0.5 to 2 cycles respectively. With 20 frames (approximately 2 seconds), the top-1 and top-5 accuracies are 86.9% and 90.3% respectively. Remarkably, even with only 5 frames (0.5 seconds) in total, the model achieves a top-1 accuracy of 76.1%, showcasing its robustness even in challenging scenarios

5.4.6 Impact of the number of candidates. The quantity of candidates present in the RGB gallery has a direct impact on performance. Our comprehensive performance is assessed with 58 subjects, simulating the number of candidates in typical scenarios. However, in certain specific scenarios , the number of candidates may notably diminish. Hence, we further evaluate scenarios where the RGB dataset comprises varying numbers of subjects in public areas. Fig.20 reveals that the accuracy of our method gradually declines as the number of candidates rises. For instance, the top-1 accuracy reaches 91% when identifying among 10 different subjects. When the number of candidates surges to 58, the top-1 accuracy remains 84%. It is important to note that the top-5 accuracy consistently remains above 90%, which validates that our method can deliver substantial accuracy for varying quantities of candidates.

6 RELATED WORK

6.1 Single-modal Person Identification

Both vision-based and RF-based person identification are extensively studied in the literature. In computer vision, researchers have delved into the distinctive gait patterns exhibited by individuals as a promising biometric identifier [13, 24, 31]. While effective, cameras can give rise to privacy concerns and struggle with poor light conditions. RF-based techniques (e.g., Wi-Fi and radar) are recently proposed for camera-restricted scenarios. [28, 33] WiFi-based method [41] analyzes the CSI spectrum produced by a single walking person for identification. MU-ID [39] uses raw radar signal to recognize up to four people simultaneously. mID [42] achieves simultaneous tracking and recognition of two subjects by voxelizing point clouds. A common limitation shared by these methodologies is their reliance on pre-collected data and labels, thereby rendering them unsuitable for scenarios involving previously unseen individuals. In contrast, our approach introduces cross-modal visual-RF identification, utilizing the strengths of both modalities.

6.2 Cross-modal Person Identification

Cross-modal ReID is an emerging topic that associates subjects detected by distinct types of sensors. The existing body of work predominantly concentrates on ReID across varying camera types, including RGB-D and RGB images [14, 18], RGB and infrared images [10, 36, 40] and RGB images with different resolution [23]. XModal-ID [21] is the prior work that achieves ReID between camera and Wi-Fi by analyzing the simulated CSI from video and real WiFi CSI. However, XModal-ID’s capabilities are constrained by the limitations of Wi-Fi resolution, thereby confining its utility to

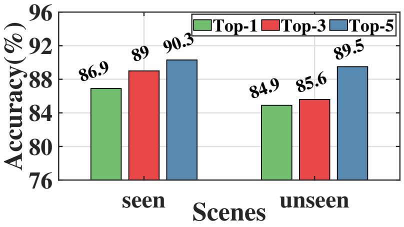


Fig. 18. Impact of the unseen subjects.

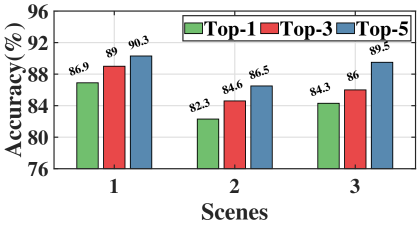


Fig. 19. Effectiveness of cross scenes.

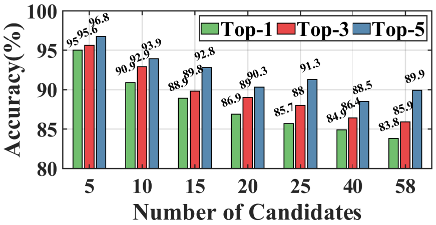


Fig. 20. Impact of the number of candidates.

single individual identification. In contrast, our proposed system excels by enabling the simultaneous identification of multiple individuals. Cao et.al [5] proposed ReID between RGB-D cameras and radars. Although effective, it necessitates the deployment of additional RGB-D cameras (specialized cameras with depth sensors, e.g., Kinect [4]), leading to additional hardware deployments and limited sensing ranges (typically ⩽ 10m). Our system, in comparison, capitalizes on the widespread availability of regular RGB cameras (e.g., existing surveillance cameras). Technically, radar and RGB-D cameras both capture 3D features of human gaits, so [5] focuses on associating 3D point clouds with different granularity. In contrast, our design addresses the unique challenge of dimension discrepancy between 2D RGB images and 3D radar point clouds.

6.3 Multi-modal Data Fusion

Human sensing using RGB cameras has achieved impressive performance, but these systems are vulnerable to harsh environmental conditions. In contrast, mmWave radars have gained attention as a sensor modality that operates reliably in all weather conditions. However, they are hindered by issues such as signal leakage and multi-path effects, which complicate accurate perception. To overcome the limitations of individual sensors, many studies have explored multi-modal data fusion for tasks such as object detection, tracking, and identification [6, 12, 37]. For example, Millieye [29] introduced a lightweight system that fuses mmWave radar and camera data to enable robust object detection. However, these approaches typically require collecting multi-modal data simultaneously from the same location. In contrast to conventional multi-modal fusion, our work focuses on establishing correspondences between individuals captured by different sensor modalities at different times and locations. Unlike traditional setups that collect co-located data, our scenario involves video and radar point clouds gathered at distinct times and from separate, non-overlapping positions. This results in query (radar) and candidate (RGB) data representing different gait instances. Such a setup introduces unique technical challenges, including temporal misalignment, viewpoint differences, and subtle variations in gait across instances. Addressing these challenges is the primary contribution of our work.

7 DISCUSSION

Scalability in a large population. `Mission` reidentifies a person detected by radar in the widely deployed RGB camera, which significantly improves the scalability of person identification techniques with radar. The proposed method provides a metric to estimate the similarity of gait in RGB video and radar point clouds so it can be extended to reidentify multiple people simultaneously. To adopt `Mission` in large population scenarios, we can use the proposed method to estimate the similarities between each pair of radar and RGB candidates. Restricted by the low-level configurations of the current mmWave radar (e.g., the maximum number of points per frame), we use a two-person experiment to demonstrate the feasibility of multi-person ReID (Section 5.2.2). In our future work, we will enhance point cloud generation from raw signal [20] to evaluate our design on a greater number of people. Scaling the techniques in a very large population could incur two challenges. First, gaits are soft biometrics (compared to the face and fingerprint). A large number of candidates will impact the Top-1 ReID accuracy. To handle the situation where a person could be misidentified as two different entities, our design provides the users with Top K candidates (the K most similar candidates). This offers the user several suggestions rather than a single result. We believe this will be useful (e.g., for investigators to narrow down the search space). Furthermore, simultaneously reidentifying multiple subjects increases the computation complexity. We plan to investigate the computation cost in our future work.

Generalization to various activities. Our design utilizes gait characteristics (i.e., the unique patterns of walking) as key features for identification. In real-world scenarios, subjects may engage in a variety of activities. To ensure that the data used for identification corresponds to walking, we can incorporate activity classification models, such as those from [30], to first detect walking segments and then extract the relevant data for re-identification (ReID). Reidentification during other activities (e.g., running) or when the subject is stationary presents additional challenges for `Mission`, which we plan to explore in future work.

8 CONCLUSION

This paper proposes a novel cross Vision-RF identification with RGB cameras and mmWave radar. To address the modal discrepancy between different sensors, we present a novel heterogeneous feature representation method based on self-mutual attention. Furthermore, `Mission` also exploit a multi-task learning architecture to help extract more representative features. It is suggested that `Mission` demonstrates the feasibility of cross Vision-RF identification and illustrates the potential for leveraging widely deployed RGB cameras in public areas for large-scale recognition and tracking across both camera-allowed and camera-free zones.

9 ACKNOWLEDGE

We sincerely thank the anonymous reviewers and shepherd for their insightful feedback, which significantly helped us improve the quality of this work.

REFERENCES

[1] 2023. A Hong Kong hospital integrated cloud, data and 5G to improve care and save lives. https://www.fierce-network.com/sponsored/hong-kong-hospital-integrated-cloud-data-and-5g-improve-care-and-save-lives-0.

[2] 2023. IWR6843ISK-ODS. https://www.ti.com.cn/tool/cn/IWR6843ISK-ODS.

[3] 2023. Radar Sensors Improve Smart Home Security, Safety, Comfort, and More. https://www.iotworldtoday.com/smart-cities/radar-sensors-improve-smart-home-security-safety-comfort-and-more.

[4] 2023. Sensor Capture + Azure Kinect + Refinement Workflow. https://www.depthkit.tv/tutorials/azure-kinect-microsoft-volumetric-capture-depth-workflow-depthkit.

[5] Dongjiang Cao, Ruofeng Liu, Hao Li, Shuai Wang, Wenchao Jiang, and Chris Xiaoxuan Lu. 2022. Cross vision-rf gait re-identification with low-cost rgb-d cameras and mmwave radars. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6, 3 (2022), 1–25.

[6] Siyuan Cao and He Wang. 2018. Enabling Public Cameras to Talk to the Public. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2, 2, Article 63 (jul 2018), 20 pages. https://doi.org/10.1145/3214266

[7] Xin Chen, Xizhao Luo, Jian Weng, Weiqi Luo, Huiting Li, and Qi Tian. 2021. Multi-view gait image generation for cross-view gait recognition. IEEE Transactions on Image Processing 30 (2021), 3041–3055.

[8] Yuwei Cheng and Yimin Liu. 2021. Person reidentification based on automotive radar point clouds. IEEE Transactions on Geoscience and Remote Sensing 60 (2021), 1–13.

[9] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. 2021. Beyond static features for temporally consistent 3d human pose and shape from a video. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 1964–1973.

[10] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. 2020. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 10257–10266.

[11] Patrick Connor and Arun Ross. 2018. Biometric recognition by gait: A survey of modalities and features. Computer vision and image understanding 167 (2018), 1–27.

[12] Kaikai Deng, Dong Zhao, Qiaoyue Han, Shuyue Wang, Zihan Zhang, Anfu Zhou, and Huadong Ma. 2022. Geryon: Edge Assisted Real-time and Robust Object Detection on Drones via mmWave Radar and Camera Fusion. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6, 3, Article 109 (sep 2022), 27 pages. https://doi.org/10.1145/3550298

[13] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. 2020. Gaitpart: Temporal part-based model for gait recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 14225–14233.

[14] Frank M Hafner, Amran Bhuiyan, Julian FP Kooij, and Eric Granger. 2019. RGB-depth cross-modal person re-identification. In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 1–8.

[15] Albert Haque, Alexandre Alahi, and Li Fei-Fei. 2016. Recurrent attention models for depth-based person identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1229–1238.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7122–7131.

[18] Nikolaos Karianakis, Zicheng Liu, Yinpeng Chen, and Stefano Soatto. 2018. Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV). 715–733.

[19] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In Proceedings of the IEEE/CVF international conference on computer vision. 2252–2261.

[20] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. m3track: mmwave-based multi-user 3d posture tracking. In Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services. 491–503.

[21] Belal Korany, Chitra R Karanam, Hong Cai, and Yasamin Mostofi. 2019. XModal-ID: Using WiFi for through-wall person identification from candidate video footage. In The 25th Annual International Conference on Mobile Computing and Networking. 1–15.

[22] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2285–2294.

[23] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. 2019. Recover and identify: A generative dual model for cross-resolution person re-identification. In Proceedings of the IEEE/CVF international conference on computer vision. 8090–8099.

[24] Athira Nambiar, Alexandre Bernardino, and Jacinto C Nascimento. 2019. Gait-based person re-identification: A survey. ACM Computing Surveys (CSUR) 52, 2 (2019), 1–34.

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

[26] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Yi Guo, Jun Cheng, Xinwang Liu, and Bin Hu. 2021. A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 10 (2021), 6649–6666.

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition. 815–823.

[28] Muhammad Shahzad and Shaohu Zhang. 2018. Augmenting User Identification with WiFi Based Gesture Recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2, 3, Article 134 (sep 2018), 27 pages. https://doi.org/10.1145/3264944

[29] Xian Shuai, Yulin Shen, Yi Tang, Shuyao Shi, Luping Ji, and Guoliang Xing. 2021. milliEye: A Lightweight mmWave Radar and Camera Fusion System for Robust Object Detection. In Proceedings of the International Conference on Internet-of-Things Design and Implementation (Charlottesvle, VA, USA) (IoTDI '21). Association for Computing Machinery, New York, NY, USA, 145–157. https://doi.org/10.1145/3450268.3453532

[30] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. 2019. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems. 51–56.

[31] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. 2019. Gaitnet: An end-to-end network for gait based human identification. Pattern recognition 96 (2019), 106988.

[32] Sudeep Tanwar, Sudhanshu Tyagi, Neeraj Kumar, and Mohammad S Obaidat. 2019. Ethical, legal, and social implications of biometric technologies. Biometric-based physical and cybersecurity systems (2019), 535–569.

[33] Raghav H. Venkatnarayan, Muhammad Shahzad, Sangki Yun, Christina Vlachou, and Kyu-Han Kim. 2020. Leveraging Polarization of WiFi Signals to Simultaneously Track Multiple People. 4, 2, Article 45 (jun 2020), 24 pages. https://doi.org/10.1145/3397317

[34] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7794–7803.

[35] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. 2019. Dynamic Graph CNN for Learning on Point Clouds. ACM Trans. Graph. 38, 5, Article 146 (oct 2019), 12 pages. https://doi.org/10.1145/3326362

[36] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-infrared cross-modality person re-identification. In Proceedings of the IEEE international conference on computer vision. 5380–5389.

[37] Jingao Xu, Hengjie Chen, Kun Qian, Erqun Dong, Min Sun, Chenshu Wu, Li Zhang, and Zheng Yang. 2019. iVR: Integrated Vision and Radio Localization with Zero Human Effort. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3, 3, Article 114 (sep 2019), 22 pages. https://doi.org/10.1145/3351272

[38] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services. 269–282.

[39] Xin Yang, Jian Liu, Yingying Chen, Xiaonan Guo, and Yucheng Xie. 2020. MU-ID: Multi-user identification through gaits using millimeter wave radios. In IEEE INFOCOM 2020-IEEE Conference on Computer Communications. IEEE, 2589–2598.

[40] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. 2018. Visible thermal person re-identification via dual-constrained top-ranking.. In IJCAI, Vol. 1. 2.

[41] Yunze Zeng, Parth H Pathak, and Prasant Mohapatra. 2016. WiWho: WiFi-based person identification in smart spaces. In 2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN). IEEE, 1–12.

[42] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. 2019. mID: Tracking and Identifying People with Millimeter Wave Radar. In 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS). 33–40. https://doi.org/10.1109/DCOSS.2019.00028