# Cross Vision-RF Gait Re-identification with Low-cost RGB-D Cameras and mmWave Radars

# Cross Vision-RF Gait Re-identification with Low-cost RGB-D Cameras and mmWave Radars

DONGJIANG CAO, Southeast University, China
RUOFENG LIU, University of Minnesota, United States
HAO LI, Southeast University, China
SHUAI WANG*, Southeast University, China
WENCHAO JIANG, Singapore University of Technology and Design, Singapore
CHRIS XIAOXUAN LU, The University of Edinburgh, United Kingdom

Human identification is a key requirement for many applications in everyday life, such as personalized services, automatic surveillance, continuous authentication, and contact tracing during pandemics, etc. This work studies the problem of cross-modal human re-identification (ReID), in response to the regular human movements across camera-allowed regions (e.g., streets) and camera-restricted regions (e.g., offices) deployed with heterogeneous sensors. By leveraging the emerging low-cost RGB-D cameras and mmWave radars, we propose the first-of-its-kind vision-RF system for cross-modal multi-person ReID at the same time. Firstly, to address the fundamental inter-modality discrepancy, we propose a novel signature synthesis algorithm based on the observed specular reflection model of a human body. Secondly, an effective cross-modal deep metric learning model is introduced to deal with interference caused by unsynchronized data across radars and cameras. Through extensive experiments in both indoor and outdoor environments, we demonstrate that our proposed system is able to achieve $\sim$ 92.5% top-1 accuracy and $\sim$ 97.5% top-5 accuracy out of 56 volunteers. We also show that our proposed system is able to robustly reidentify subjects even when multiple subjects are present in the sensors' field of view.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Cross Modal, Gait Re-Identification, Millimeter Wave Sensing, Specular Reflection

**ACM Reference Format:**
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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 1, Article 1 (January 2022), 25 pages. https://doi.org/xxx

## 1 INTRODUCTION

Person identification is a key enabler to realize ubiquitous computing and finds many applications in everyday life, such as personalized services, automatic surveillance, continuous authentication, and contact tracing during pandemics, etc. A variety of sensing techniques have been proposed for person identification and are used

*Corresponding author.

Authors' addresses: Dongjiang Cao, djcao@seu.edu.cn, Southeast University, Nanjing, Jiangsu, China; Ruofeng Liu, liux4189@umn.edu, University of Minnesota, Minnesota, Minnesota, United States; Hao Li, 220194384@seu.edu.cn, Southeast University, Nanjing, Jiangsu, China; Shuai Wang, shuaiwang@seu.edu.cn, Southeast University, Nanjing, Jiangsu, China; Wenchao Jiang, wenchao_jiang@sutd.edu.sg, Singapore University of Technology and Design, Singapore, Singapore; Chris Xiaoxuan Lu, xiaoxuan.lu@ed.ac.uk, The University of Edinburgh, Edinburgh, United Kingdom.


2474-9567/2022/1-ART1 $1
https://doi.org/xxx

in different scenarios. For example, in camera-allowed areas where privacy is less a concern, e.g., streets and entrances to buildings, cameras or other vision-based techniques have been extensively adopted for identification. In contrast, radio frequencies (RF) based sensors are widely known to be superior for privacy preservation [47] and are suitable in camera-restricted scenarios where any vision-based techniques (including both RGB and depth imaging sensors) will raise severe privacy concerns and legal issues, e.g., home, restroom, hospital, offices, or some confidential areas in a factory.

Camera-allowed and camera-restricted regions, on the other hand, are often connected with each other in the real world with people regularly moving across them. Moreover, due to the widespread IoT deployments, both of these regions are being increasingly installed with different kinds of sensors [6, 21, 41]. For instance, it is common for workers in a factory to leave from public spaces (e.g., monitored by RGB or RGB-D cameras) and enter confidential warehouses (e.g, monitored by RF sensors) or vice versa. The movement of humans across scenarios with different privacy considerations and sensor deployments leads to a new need for cross-modal human re-identification (**ReID**) - *given a person detected by a RF sensor, the system can identify the same person in camera footage or vice versa.*

Among the wide spectrum of vision-based and RF-based sensors, single-chip millimeter-wave (mmWave) radars and RGB-D cameras respectively emerge as two promising modalities in recent years. Single-chip mmWave radars are a type of sensor designed to have the capability of simultaneously sensing multiple targets in the field of view (FoV), generating precise spatial information (i.e., range and angle of arrival) of each target [43]. Meanwhile, as a RF-based sensing modality, mmWave radars are impervious to any lighting conditions, e.g., darkness, dimness, and glare. These advantages lend the radar itself an effective sensing alternative when cameras fall short [14, 16, 32**?** ]. On the other hand, RGB-D cameras are a type of depth sensing device that work in association with a RGB camera. As they are able to augment the conventional image with depth and scale information, RGB-D cameras significantly outperform other traditional cameras (e.g., RGB cameras) in the domain of vision-based human sensing [13]. With the recent maturity of manufacturing, low-cost RGB-D cameras (e.g., Microsoft Kinect and Intel Real Sense) and mmWave radars (e.g., Texas Instruments IWR series) become available which further strengthen their promise for human sensing tasks.

Motivated by the need for cross-modal identification and the emerging opportunity above, this work proposes the first cross-modal re-identification design among RGB-D cameras and mmWave radar. The key idea underpinning our design is leveraging the human gaits (i.e., how a person walks) across modalities, which is a unique and consistent biometric identifier of each individual [25]. To achieve cross-modal ReID, gait information is captured in both modalities and their similarity is estimated to associate people between cameras and radars.

While gait recognition has been studied separately with mmWave radars [43] or RGB-D cameras [7, 9], new technical challenges come up when leveraging the cross-modal data for human identification. *First*, the sensing data from vision and radar inherently suffer from severe inter-modality discrepancy. Cameras capture detailed shapes and motions of each body part, whereas the radar only provides discrete and sparse detection points, which renders a direct similarity calculation across modalities impossible. *Second*, as the camera and radar in our context are installed in different places, their data are not collected at the same time and thus are often not synchronized in fine-granularity. As a result, a reliable similarity estimation must be robust to various interference factors (e.g., temporal misalignment and minor gait changes).

In order to address the above technical challenges, several novel designs are proposed in this work. *First*, to tackle the inter-modality discrepancy, we propose a modal-unifying method driven by the physical model of specular reflection [3]. We find that the human body acts as a spectral reflector (like a mirror) of incident mmWave signal, and thus only signals arriving close to the normal of the surface on the human body are reflected back towards the receiving antenna [4]. Consequently, when a person is moving relative to the radar, the body parts being captured change over time and form spatio-temporal signatures of human gait. Such gait signatures differ significantly from person to person for their different walking styles (e.g., the stride of legs, arms swing

and motion of torso). Moreover, the gait captured by the RGB-D cameras and mmWave radars can be tightly corresponded via the specular reflection model which bridges the heterogeneity between vision and RF data. Based on this observation, we design a novel algorithm to synthesize signature points from depth images while addressing imperfections of raw RGB-D data and view angle difference of the two sensors). *Next*, to deal with interference factors caused by unsynchronized data collections, we design an end-to-end deep metric learning pipeline for similarity estimation. The model extracts high-level representations of gait features from both synthesized signatures and radar point clouds and estimates their similarity in the high-dimensional latent space, thus addressing the temporal misalignment, minor walking speed change, etc.

To summarize, our work makes the following contributions:

- We present the first-of-its-kind cross-modal re-identification design between mmWave radars and RGB-D cameras, achieving simultaneous multi-person ReID across camera-allowed and camera-restricted scenarios.
- We propose novel designs to address intrinsic technical challenges of cross-modal ReID (e.g., inter-modality discrepancy and similarity estimation).
- We implement our design with commodity mmWave radar (i.e., IWR6843) and RGB camera (i.e., Azure Kinect). Extensive evaluations with 56 volunteers demonstrate that our design is highly accurate (92.5% top-1 accuracy and 97.5% top-5 accuracy ) and robust in handling multi-person ReID. This multimodal dataset collection is released to the community.

## 2 MOTIVATION

### 2.1 The need for cross Vision-RF identification

**The need for cross Vision-RF identification:** Depending on requirements of specific applications and level of privacy sensitivity in various scenarios, RGB-D cameras and RF sensors are deployed at different scenes. For example, nowadays public streets in many smart cities are deployed with surveillance RGB-D cameras to achieve fast and accurate behavior recognition and three-dimensional positioning [28, 33], which overcomes the limitations of traditional RGB camera in dealing with lighting changes, color similarity and shadows. On the other hand, many people prefer to install RF techniques (e.g., WiFi and radar) to monitor their home [2, 20], which avoids privacy invasion (e.g., room layout) while providing various intelligent services such as smart control [20] and fall detection for the elderly [34]. Another example of the hybrid deployment is the enterprise scenario (e.g., office and factory) where RGB-D cameras are increasingly being used in building entrance control systems to improve the accuracy and robustness [26] of identity recognition (e.g., liveness detection [33]), whereas RF sensors are more acceptable at meeting rooms, offices and laboratories because they prevent the leakage of the enterprise's confidential information. This emerging heterogeneous sensing scenario motivates us to study the possibility of cross vision-RF ReID. Specifically, when a RF sensor detects a person of interest in the camera-restricted area, our design aims at retrieving the same person's footage from cameras installed in camera-allowed areas. We describe three broad sets of applications that this system can add values on:

**1) Personalized Services.** By matching a person detected by RF sensor to his/her image, we could exploit the rich information in the image (e.g., age, gender, and apparel) to provide personalized services without invading their privacy (e.g., information about fine-grained activities or confidential behaviors) in camera-restricted areas. For example, we could adjust the temperature in a private office based on the clothes of a person captured by cameras at the entrance.

**2) Tracking and Identification.** Our design also enables seamless tracking when people move between the camera and RF monitored areas, generating complete trajectories across camera-allowed and camera-restricted scenes. Seamless tracking across spaces can be realized and assist in situations such as contract tracing during pandemics.

**3) Security and Surveillance.** Furthermore, our technique enriches RF sensors' capability on security surveillance. For example, when trespassing happens at home, RF-based surveillance monitors could not only detect intruders but also utilize wireless signal as the cues to retrieve the image of the intruders from video footage of the camera on a nearby public street. This will help the investigator to quickly identify the suspects.

### 2.2 Limitation of existing solutions

**Single-modal Person Re-Identification.** Person ReID among homogeneous sensors have been investigated for both vision [36] and RF technologies [24, 29, 37] *separately*. However, these techniques only provide ReID among the specific regions that are equipped with the same type of sensors. Therefore, they cannot accommodate the ReID problem when the subjectives are moving across camera-allowed and camera-restricted scenarios that are covered by different types of sensors. This fundamental limitation motivates us to design a new technology that can effectively bridge these isolated ReID systems, providing a unified identity solution that serves real-world scenarios that contains areas with different camera restrictions.

**Cross-modal Person Identification.** Person ReID across different sensing modality is mainly studied by computer vision community with a focus on different types of cameras, including the identification between depth and RGB images [13, 15], infrared and RGB images [40, 44], and images with different resolutions [19]. These solutions are only applicable at camera-allowed areas. On the other hand, the cross-modal ReID between vision and RF signals, despite its importance and ubiquity, receives little attention. The recent work (i.e., XModal-ID) [17] is the pioneering study of cross-modal ReID between the camera and RF. They propose to match video and WiFi signals using channel state information (CSI). Nevertheless, as CSI-based techniques struggle to simultaneously identify multiple people in the same scene, XModal-ID has to assume that there is only one person walking in the WiFi area at a time [17]. This assumption, however, prohibits it from dealing with multi-person scenarios (e.g., contact tracing). In addition, XModal-ID requires a separate WiFi transmitter and receiver, which often introduces extra deployment cost and calibration complexity.

### 2.3 Our proposed approach

**Our proposed approach:** In this work, we utilize the emerging RF and vision sensors (i.e., mmWave radar and RGB-D camera) to achieve ReID between camera-allowed and camera-restricted areas. Recently, both low-cost commercial mmWave radar and RGB-D camera have become widely available and a large body of human sensing applications (e.g., people counting, intruder detection, and health monitoring) have been developed with these two sensing modalities [9, 43]. We envision that both sensors will be massively deployed in the near future. Technically, both mmWave radars and RGB-D cameras provide the precise three-dimension spatial information of multiple targets (e.g., subjects' locations and heights) in the FoV. It is therefore advantageous to exploit this 3D sensing capability to solve multi-person identification problems [1]. Finally, because the radar itself is a transceiver device, only one radar is required for human sensing so its deployment is more flexible than WiFi (i.e., requiring separate WiFi transmitter and receiver). The opportunities above motive us to study the cross-modal ReID problem between mmWave radars and RGB-D cameras.

## 3 SYSTEM OVERVIEW

In this paper, we propose a cross vision-RF person ReID system. Given a person detected by a mmWave sensor, the system re-identifies him/her in a gallery of RGB-D camera footage using the consistent gait features. Fig.1 shows the overall architecture of our design. The RF module processes the received radar signal to obtain the radar points that belong to the person of interest. On the other hand, the vision module synthesizes signature

[1]Note that although less fine-grained 3D information can be obtained by a depth imaging sensor alone, we focus on RGB-D cameras in this work because using RGB images to enhance raw depth images has been widely available as an option [1, 42, 45].

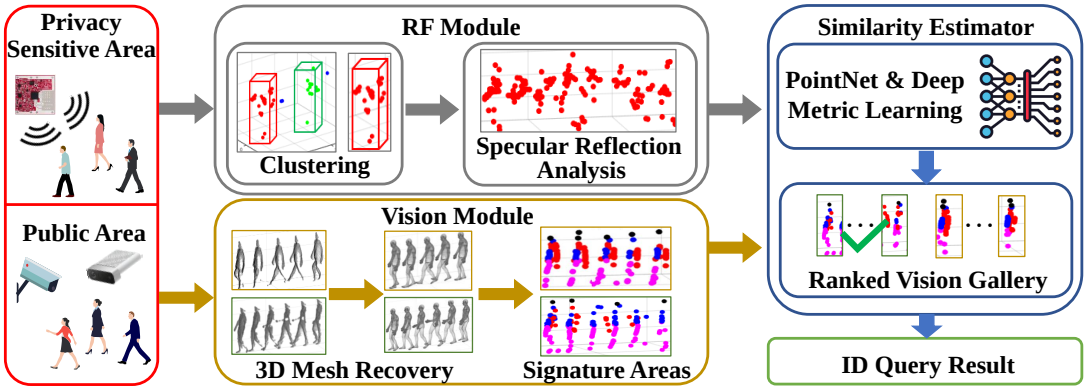



Fig. 1. System architecture.

areas from the reconstructed 3D mesh of a candidate captured by RGB-D cameras, by using the specular reflection phenomenon of mmWave (Section 5). Then, we measure the walking style similarity by comparing the sequence of radar points and synthesized signature areas (Section 6). The similarity of all the candidates are ranked and the one with the highest similarity score, or those with top similarity scores, are ultimately selected as the output.

## 4 BACKGROUND

This section introduces the primer of mmWave radar, gait, and person ReID needed in this work. We refer the interested readers to [45] for the background of RGB-D cameras and avoid repeating the same content here.

### 4.1 Principles of mmWave Radar

The single-chip mmWave radar is based on the principles of frequency modulated continuous wave (FMCW) and has the ability to measure the range, relative radial speed and angle of the target. Specifically, the FMCW radar repeatedly transmits continuous chirp signals for a short period of time which frequency increase linearly with time. When receiving the signal reflected by an object, the radar sensor produces Intermediate Frequency (IF) signal, which is analyzed to obtain the three-dimensional position of the object.

**Range Measurement.** Based on the IF signal, the distance $d$ between the object and the radar is calculated as:

$$d = \frac{f_{IF}\, c\, T_c}{2\, B} \tag{1}$$

where the $c$ is the speed of light, $f_{IF}$ is the frequency of the IF signal, $B$ is the bandwidth swept by chirp, and $T_c$ is the duration of chirp. To measure the range of multiple objects at different ranges, a fast Fourier transform (FFT) is performed on the IF signal (i.e., range-FFT). The result of range-FFT represents the frequency response at different ranges. Thanks to the centimeter level range resolution, it has the ability to detect the position of the torso and limbs.

**Angle of Arrival Estimation.** To depict the exact positions of objects in a spatial Cartesian coordinate system, the angle estimation is indispensable. The mmWave radar utilizes a linear antenna array to estimate the object angle. After emitting chirps with the same initial phase, RF Front-end simultaneously samples from multiple receiver antennas. Because the phases of the received signals are different between receiver antennas, the angle of the reflected signal can be estimated. Formally, the AoA can be calculated as:

$$\theta = \arcsin \frac{\lambda\, \omega}{2\, \pi\, l} \tag{2}$$

where $\omega$ denotes the phase difference, $l$ represents the distance between consecutive antennas and $\lambda$ is the wavelength. Once obtains the range and AoA ($\theta$) of the targets, we can get the exact positions of objects in a spatial Cartesian coordinate system.

### 4.2 Person Re-Identification and Human Gait

**Person ReID:** Originally proposed by the computer vision community, person ReID has been widely studied as a specific person retrieval problem across non-overlapping cameras. Given a query person-of-interest, the goal of ReID is to determine whether this person has appeared in another place at a distinct time captured by a different camera, or even the same camera at a different time instant. The *query* person can be represented by an image and the large collection of recorded images and/or videos is referred to as a candidate *gallery* [18]. In the rest of this paper, we will follow this general terminology of ReID but focus ourselves on a distinct scenario where a query and the gallery collection are obtained by different sensing modalities (mmWave radar V.S. RGB-D camera) and in different places (camera-allowed areas V.S. camera-restricted areas).

**Human Gait Identification:** As a type of biometric when a person is walking, gait information often carries identity-specific patterns that can be utilized to distinguish human identity [25]. A gait-based identification system starts from extracting walking features over a data sequence, such as step length, magnitude of arm swing, pace, etc, and then utilizes these features as gait signatures for identity matching. Because walking is a dynamic sequence of movements [25], the process of gait feature extraction is often applied over a sequence of gait frames. In this paper, we leverage the sequence frames of walking to extract gait features from two different sensors and utilize the gait features to re-identify the same person across two types of data.

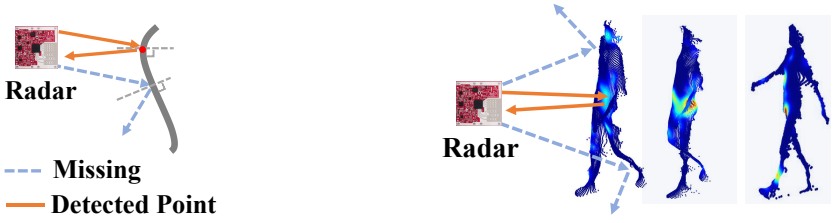
## 5 FEASIBILITY OF CROSS-MODAL REID

This section investigates the feasibility of cross-modal ReID between RGB-D cameras and mmWave radars. We start by introducing the specular reflection model, and then introduce the important notion of spatio-temporal signature utilized for cross-modal identity matching in subsequent sections.

### 5.1 Specular Reflection

*5.1.1 Specular reflection model.* The critical challenge of cross vision-mmWave ReID lies in the significant data discrepancy between the two modalities. While fine-grained shapes of human bodies can be obtained from RGB-D cameras, a mmWave radar only detects a few points of a target person in a noisy and sparse representation.

Our key insight to address the discrepancy issue is inspired by the observation that most points detected by a radar are not arbitrary but meaningful reflections on the human body with salient physical characteristics. Specifically, human skin produces strong specular reflections (similar to a mirror) to the incident mmWave signal because the roughness of skin is considerably smaller than the wavelength of mmWave signal and skin also has high water content [4].




(a) Specular reflection. (b) Radar detected body areas.

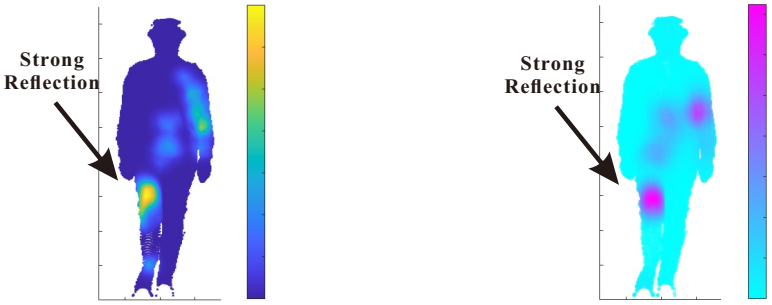
Fig. 2. Specular reflection model.

As shown in Fig.2(a), when the mmWave signal hits the surface of human skin, the majority of energy bounces off in the direction symmetric to the incident angle instead of scattering to all directions. Due to the small antenna aperture of a low-cost mmWave radar, only those signals arriving close to the normal of the surface on the human body can return to the radar, while other signals are deflected away. Consequently, as Fig.2(b) depicts, a radar detects the body area with a unique geometric characteristic - the normal of the surface must point towards the radar device. We hereafter refer to these unique body areas as the "specular reflection signature areas" (or "signature areas" for short).

*5.1.2 Empirical measurement.* To examine the consistency between signature areas on the human body and the areas detected by a radar, we conduct an empirical measurement. We respectively collect the radar point cloud of 11 subjects from a mmWave sensor as well as their 3D mesh from RGB-D camera (preprocessed with recovery algorithm [50] as detailed in Section 6.1). The signature areas are extracted from the 3D mesh using the specular reflection model (details will be given in Section 6.1). Fig.3 demonstrates the comparison of the radar detected areas with the signature areas at a radar frame. Strong reflection is indicated by brighter color. Clearly, we can observe that the signature areas reflected from the 3D body mesh largely agree with the points detected by a radar. Both figures in Fig.3 demonstrate strong reflection on the right thigh (because the right thigh is being lifted) and the left upper arm. We further quantify such cross-modal consistency using a basic and intuitive approach - the intersection ratio, i.e., the percentage of radar points that are inside signature areas. Fig.5(a) depicts the complementary cumulative distribution function (CCDF) of the ratios over 11 people, each of them has 2 records with a total of 50 radar frames (5 seconds). The curves of 11 people show that the majority of radar frames (~90%) have over 70% radar points within the signature areas. This result motivates us to utilize the specular reflection model to associate the vision and RF data.

## 5.2 Spatio-temporal Signature of Gait

*5.2.1 Spatio-temporal Signature.* We exploit the specular reflection model to compare the walking pattern of different people. In specific, when a human is walking towards the radar, the orientation of each body part w.r.t the radar (the angle formed by each part of the body and the radar) changes over time. Therefore, the signature areas (i.e. the subset of body parts detected by the radar) also temporally alternate. Notably, as each person has distinct body shapes (e.g., height and body part lengths) and unique gait (e.g., gait period, stride length, and the angle of arm swing), walking towards a radar will produce a sequence of signature areas that possesses a spatio-temporal pattern unique to human identity. It is this unique spatio-temporal pattern that allows us




(a) mmWave radar detection. (b) Synthesized signature areas.

Fig. 3. Comparison of radar detection and signature areas on 3D mesh. The intensity of reflection is indicated by brightness. (a) The body areas detected by mmWave radar when a subject is walking. (b) Specular reflection signature areas synthesized based on the specular reflection model.

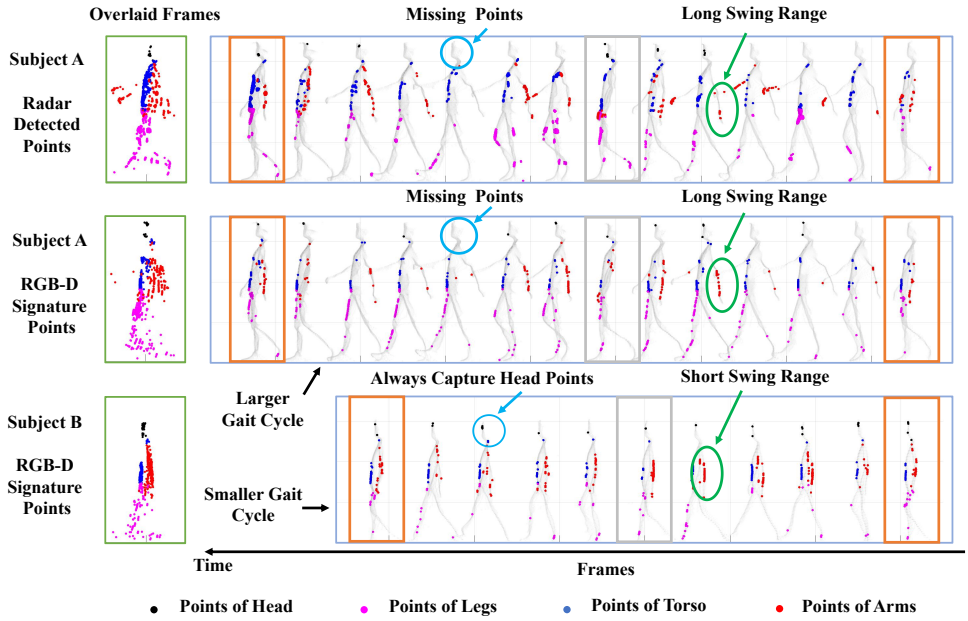
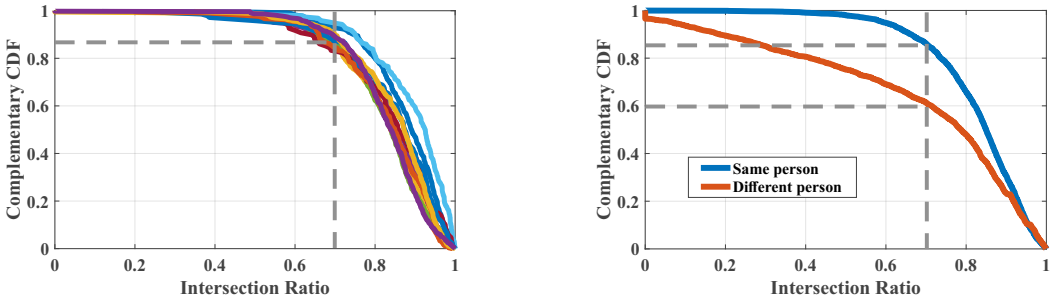



Fig. 4. Radar points and spatial-temporal signatures of two subjects. Radar detected points of subject A ($1^{st}$ row). Signature points on 3D mesh of subject A ($2^{nd}$ row). Signature points on 3D mesh of subject B. ($3^{rd}$ row).

to locate the consistent person's ID across vision and RF modality. We hereafter refer to this consistent ID as "spatio-temporal signature". Fig.4 demonstrates the side views of spatio-temporal signatures of two different subjects during a gait cycle. For ease of visualization, each signature area is aggregated and depicted as a point and the signature areas on different body parts are plotted with different colors. The first and second row of the figure depicts the radar points and signature points of subject A, while the last row is the signature points of subject B. To make it more intuitive, the true mesh of a body obtained from the RGB-D cameras is also given in the grey background.

**Observation 1:** Several notable gait discrepancies among two subjects can be observed from two signatures, including dynamic kinematic characteristics and static shape characteristics e.g., 1) *stride:* The stride or step length could be observed by trajectories of the leg (pink points). Both the radar and signature points of subject A show a larger step length than subject B. 2) *arm swing:* The distribution of the red points represents the amplitude of arm wings. The swing amplitude of subject A is visibly larger than subject B, as seen from the overlaid frames. 3) *gait cycle:* Fig.4 shows a sequence of frames on the whole gait cycle. Looking into the gait cycle, Subject A has a longer gait cycle than Subject B, implying that Subject B has a faster cadence. 4) *height:* The heights of both subjects can be obtained from the range of signature areas in the vertical direction. Also, the head of subject B is detected in all radar frames due to the person's unique height while subject A's head is detected intermittently.

**Observation 2:** It is also observed that signature areas on various body parts have uneven contributions in differentiating subjects. As Fig.4 shows, for both subjects, the torso can be detected by the radar at a similar horizontal position in almost every frame. This, however, means that the signature areas on the torso carry little identity-specific information. In contrast, the patterns of signature areas on the arm and leg vary dramatically from person to person because the limbs have a larger degree of changes than the torso when a person is walking and thus can capture the most salient gait features (e.g., step length and arm swing).




(a) CCDF of the intersection ratios of 11 people, legend omitted for readability.

(b) CCDF of the intersection ratios: same person vs. different person

Fig. 5. CCDF of the intersection ratios. (a) Ratios of a same person (b) Same person (mean = 82.53%) vs. different persons (mean = 67.39%).

*5.2.2 Empirical measurement.* We now quantitatively examine the feasibility of using the signature areas to differentiate different people across vision and radar. Specifically, we analyze the intersection ratio (the percentage of radar points that are inside signature areas) of 11 people with each of them having 50 frames (~ 4 gait cycles). Fig.5(b) shows the result. The blue curve depicts the CCDF of intersection ratios when the radar points and signature areas are from the same person. The orange curve is the average result of different people. As shown in Fig.5(b), for the case of different people, only 60% of radar frames have over 70% intersection ratios with the signature areas - a level significantly lower than the intersection ratio measured on the same person (> 86%). In addition, there are ~25% radar frames that have less than 50% intersection ratio between different people. In contrast, this number reduces from 25% to <5% for the case of the same person, demonstrating that the effectiveness of the spatio-temporal signature for cross vision-RF ReID. Fig.6 further shows the average intersection ratio for these 11 different persons, breakdown in different body parts. As we can see, the intersection ratio of the torso across different people is significantly larger than the other two, implying that the torso should not be considered as salient area when differentiating persons. We hypothesize that it is the torso that contributes the most false intersection ratios among different people in Fig.5(b). Limbs, on the other hand, have a much smaller intersection ratio across different people, indicating their potential effectiveness and should have more weights compared with the torso.

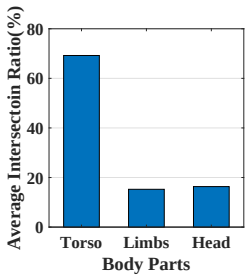



Fig. 6. Comparison of average intersection ratios of different body parts from different people.

## 5.3 Summary and Challenges

In summary, cross-modal ReID is feasible based on the above observations and empirical studies: 1) specular reflection can accurately model radar detection of human body; 2) a unique spatio-temporal signature can be produced from 3D mesh using the observed specular reflection model, which captures distinct gait features of the subject. Also, signature areas on various body parts have uneven contributions in differentiating subjects; 3) specular reflection can effectively bridge radar and vision modality for cross-modal ReID.

However, to develop a practical ReID algorithm based on these observations, we have to further address several technical challenges. First, we assume a perfect 3D mesh of a subject in spatio-temporal signature synthesis, whereas the raw outputs of a RGB-D camera often suffer from self-occlusion and noise (Section 6.1). Second, the 3D mesh is obtained from the camera's view angle. To synthesize the radar signature, we need to transform it to the radar coordinate system, which requires the knowledge of the location and orientation of the subject (Section 6.1) in radar. Third, RGB-D image and radar point clouds are collected in a unsynchronized manner. The similarity

estimation algorithm needs to be robust against the temporal misalignment and minor gait difference when the subject walks in disjointed areas (Section 6.2). Finally, since various body parts contain different amounts of identify-specific information, the algorithm is expected to dynamically adjust their contribution to the similarity estimation (Section 6.2).

## 6 CROSS VISION-RF GAIT RE-IDENTIFICATION

This section introduces the design of cross-modal ReID. We start by introducing the signature synthesis based on specular reflection model, and then discuss the similarity estimation.

### 6.1 Signature Synthesis

**Mesh Preprocessing.** The feasibility analysis in Section 5.1 indicates that the radar capture has a unique geometric characteristic due to the specular reflection property of human skin. Based on this phenomenon, one can figure out these signature areas on the 3D mesh of a human body and thereby synthesize spatio-temporal signatures for cross-modal ReID. However, using the body mesh obtained from a single RGB-D camera to synthesize the signature has two issues: (i) The RGB-D camera only outputs the surface of the human body directly facing the device while other surfaces are self-occluded. Since the camera and mmWave radar are installed in different locations, they may view the person from different angles. Therefore, synthesizing the radar signature requires those self-occluded body mesh, which is missing in the raw RGB-D output. (ii) The raw output from RGB-D camera contains a lot of holes and noisy points due to undesired artifacts [42]. Without handling, these noises and holes could lead to errors in the spectral reflection synthesis.

To address these issues, we propose to reconstruct the full-body mesh first from raw RGB-D output before any cross-modal matching is performed. Concretely, we input each frame of RGB-D capture with self-occlusion and noise into the state-of-the-art human mesh recovery algorithm [50]. The algorithm fits the data into a human body model that recovers the self-occluded body mesh and gets rid of the noise. Fig.7 shows a few raw RGB-D camera snapshots and their corresponding reconstructed full-body 3D mesh, demonstrating that the algorithm can effectively recover the missing body mesh and reduce the noise.

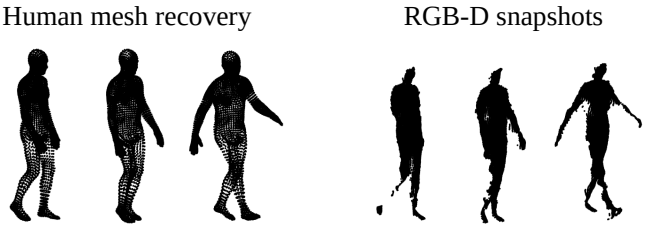



Fig. 7. Samples of meshes output from recovery algorithm (left) for snapshots of a walking person (right).

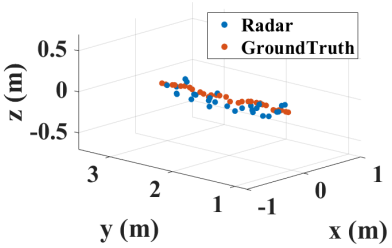



Fig. 8. Estimation of location and trajectory of subject's center using radar vs. GT using RGB-D.

**Coordinate Transformation.** Before using the reconstructed full-body 3D meshes to synthesize the signature areas, we must first shift the 3D mesh to the cartesian coordinate system of radar. To achieve this, we exploit the rich 3D information of radar to find the location and orientation of the subject. Specifically, we first obtain a rough estimation of the initial position of the subject directly from the first frame of radar point clouds. Then, we fine-tune the position by matching the geometric center of 3D mesh to the one of the radar point cloud, which are figured out by stacking the radar frames in the walking direction. Furthermore, we estimate the orientation of the subject from a short walking trajectory, which could be considered as a straight line for a very short time. With the initial location and trajectory, we can shift the sequence of 3D mesh to the radar coordinate by mapping

the first mesh to the location of the first radar frame and aligning its orientation. Other meshes are also shifted using the first one as the reference. As Fig. 8 demonstrates, the estimation of location and trajectory is very close to the ground truth. It's worth noting that although our estimation is coarse-grained, small error can be tolerated by the redundancy of signature synthesizer and resiliency in our deep neural network based similarity estimator.

**Signature Point Synthesis.** We utilize spectral reflection model to synthesize the radar signature. Note that RGB-D cameras are digital sensors, so our algorithm obtains the discrete samples of these signature areas, which are referred to as "signature points" in the following sections. Formally, we denote the set of shifted dense 3D mesh points of a human body at time $t$ by $\mathcal{P}(t)$, while the location of the (virtual) mmWave radar is denoted by $\boldsymbol{x}_m = (0, 0, 0) \in \mathcal{R}^3$. For each point in $\mathcal{P}(t)$, we find two vectors - the normal vector of the point and the vector from the point to the radar. The normal vector $\boldsymbol{n}_s(t)$ of the point $\boldsymbol{s}(t)$ is determined by the local plane fitted of six neighboring points. And the vector from the point to the radar is $(\boldsymbol{s}(t) - \boldsymbol{x}_m)$. If the angle between the two vectors is sufficiently small, then that point is considered as a signature point. As shown in equation 3, we obtain the spatio-temporal signature $\mathcal{P}_s(t)$, which is a set of signature points. $\epsilon$ is the threshold of the angle, which is set to a non-zero value to tolerate small errors in the subject localization.

$$\mathcal{P}_s(t) = \{\boldsymbol{s}(t) \in \mathcal{P}(t) | \arccos \frac{(\boldsymbol{s}(t) - \boldsymbol{x}_m) \cdot \boldsymbol{n}_s(t)}{|\boldsymbol{s}(t) - \boldsymbol{x}_m||\boldsymbol{n}_s(t)|} < \epsilon\} \tag{3}$$

## 6.2 Similarity Estimation

*6.2.1 Similarity Estimation via Deep Metric Learning.* Section 5.2 shows that we can synthesize mmWave signature points from a subject's RGB-D images which demonstrates significantly higher similarity with the subject's own radar point cloud than the one synthesized using the images from a different person. We leverage the key insight to reidentify the radar target among the RGB-D candidates. This section aims at design a reliable estimation algorithm that satisfies the following objectives. First, the radar and RGB-D data are collected in the disjointed areas and thus are not synchronized in the time. Because an explicit alignment of two modalities is non-trivial due to the sparsity of point clouds, we desire that the estimation algorithm can be invariant to the unknown temporal misalignment. Second, the walking speed and step length might slightly change when people are walking in different areas. Our algorithm must be robust against these minor gait variations. Finally, as we introduce in Section 5.2.2, various body parts (e.g., torso, head and limbs) carry different amounts of identity-specific features. Presumably, an effective similarity estimation needs to be anatomy-weighted - body parts containing more salient gait features (e.g., head and limbs) should have more impact on the similarity estimation result than the parts with less salience (e.g., torso).

To achieve these goals, we design a deep metric network to estimate the similarity, which have several benefits. The neural network can extract spatio-temporal features from a sequence of point cloud as well as synthesized signature points and map them into the same feature space, which addresses unknown time misalignment. In addition, by comparing the extracted feature instead of the original input data (e.g., by the location of point cloud), the estimator is more tolerant of minor changes in gait and imperfection in the signature synthesis. Finally, the neural network can learn to dynamically adjust the weight of various body parts and frames based on the context (e.g., pose of the subject).

*6.2.2 Similarity Estimation Model.* Fig. 9 depicts the architecture of our designed similarity estimation network based on deep metric learning. Deep metric learning [10] is to learning deep feature embeddings that better fits a simple distance function such as Euclidean distance or cosine distance, which has been an active research topic in computer vision community. Given a deep metric model, samples with similar content are projected onto neighboring locations on a manifold, while samples with different semantic context are mapped apart from each other. As for this work, We train the network with triples: an anchor instance (i.e., mmWave points of a subject), a positive instance (i.e., RGB-D signature points of the same subject), and a negative instance (i.e., RGB-D signature

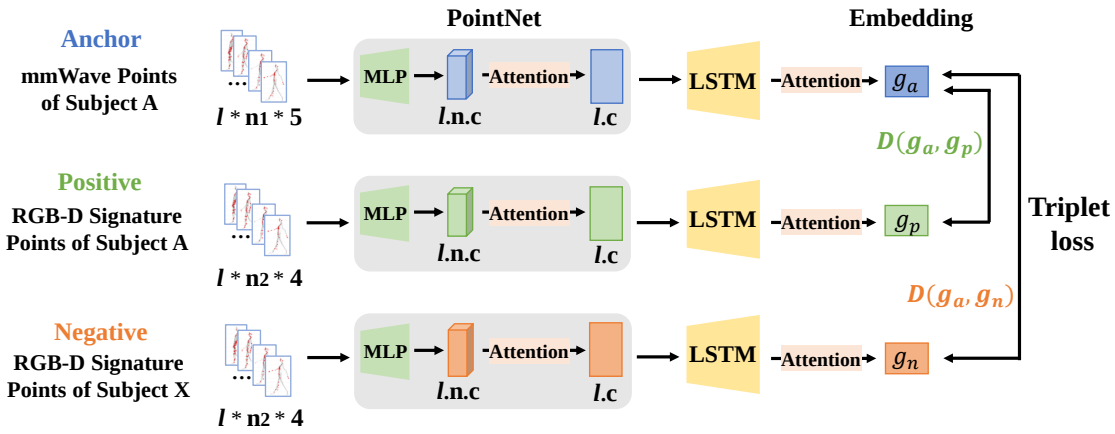



Fig. 9. Similarity Estimation via Deep Metric Learning.

points from a different subject). The network learns to map the both mmWave and RGB-D signature into an identical feature space in which the distance between radar point cloud and signature points of the same person is minimized whereas the distance is maximized when they come from different subjects.

**Feature Extraction: mmWave.** For mmWave point cloud, we first extract the feature from each individual radar frame with PointNet [30]. PointNet directly ingests point cloud $P_t = \{p_{i,t}\}_{i=1}^{N}$ contains $N$ mmWave points each of which consists of five features, i.e., 3D coordinates $(x_{i,t}, y_{i,t}, z_{i,t})$, intensity $(s_{i,t})$ and velocity $(v_{i,t})$. Then, it utilizes permutation-invariant operators (e.g., pointwise MLP and attention) to deal with these unordered points. More specifically, it encodes each point $p_{i,t}$ independently using a shared-weighted MLP (Multi-layer Perception) and outputs a high-level representation of each point denoted as $m_{i,t} = MLP(p_{i,t}; \theta_m)$, where $\theta_m$ is the learnable parameters of the MLP. Then, we aggregate the features of all the points in a frame using attention mechanism [35]. The attention computes a score for each point and then calculates a weighted sum of all scores. Benefited from this mechanism, the estimator dynamically adjusts the contribution of each point based on the amount of identity-specific features the corresponding body part carries. The attention procedure is formally given as follows:

$$f_t = \sum_{i=1}^{N} A_p(m_{i,t}; \theta_a) \times m_{i,t} \tag{4}$$

where $N$ is the number of points in $t^{th}$ frame, $A_p()$ is a learnable attention function implemented by a full connection layer, and $\theta_a$ is the parameter of attention function.

Since the gait is a sequence of motion, we need to further exploit the spatio-temporal correlation among each individual frames. To achieve this, we pass the sequence of the extract feature $f_t$ throughput a LSTM (Long short-term memory network) and obtain $r_t = LSTM(f_t, r_{t-1}; \theta_r)$, where $\theta_r$ denotes the parameters of LSTM. Finally, in order for the result to be invariant to unknown temporal misalignment between two inputs, we aggregate the LSTM output of the all frames using another attention network. As equation 5 illustrates, the attention function computes a score for each frame in a walking sequence and then aggregates them by a weighted sum of the all scores where the weights are dynamically adjusted based on the significance to the identification.

$$g_a = \sum_{i=1}^{L} A_g(r_t; \theta_g) \times r_t \tag{5}$$

where $L$ is the length of frame sequence, $\theta_g$ is the parameter of attention function and $g_a$ is the gait feature embedding of mmWave points of an anchor instance.

**Feature Extraction: RGB-D.** We adopt a similar feature extractor for RGB-D signature points. In addition to the pointNet and attention network designed for mmWave, we further exploit the availability of rich amount of information from RGB-D camera to accurately identify the body segment label of each signature point. By doing this, we can guide the network to learn the different salience of body parts in identification task and adjust the weights to the final results. Therefore, the input RGB-D signature point $q_{i,t}$ is a vector of four features, i.e., 3D coordinates $(x_{i,t}, y_{i,t}, z_{i,t})$, and body parts index $(b_{i,t})$. We obtain the gait feature embedding of positive and negative RGB-D signature points denoted as $g_p$ and $g_n$.

**Loss Function.** Each input of the deep metric objective function is a triple consist of anchor, positive, negative instance denoted as $< g_a, g_p, g_n >$. In order for the samples with the same identities to be as close as possible in the embedding space, and the samples with different identities to be as far away as possible, the objective function is designed as follows:

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

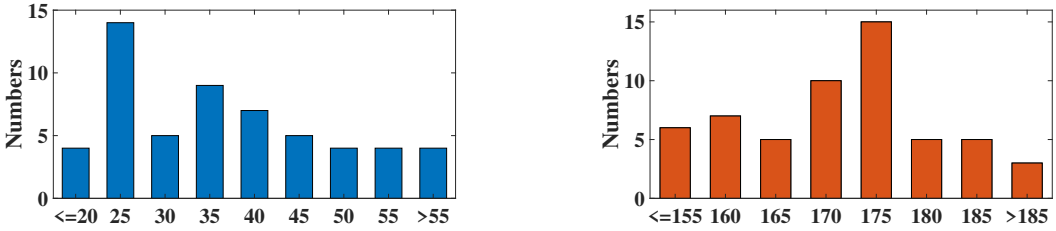
where $D(g_a, g_p)$ is the euclidean distance between embedding of anchor(e.g., mmWave points of the subject) and embedding of positive (e.g., RGB-D signature points of the same subject), $D(g_a, g_n)$ is the distance between embedding of anchor and embedding of negative (e.g., RGB-D signature points of subject X). *margin* is the hyper-parameter, which makes the distance value of anchor and negative samples larger, while making the distance value of anchor and positive samples smaller. The whole framework is trained end-to-end.

## 7 IMPLEMENTATION AND DATA COLLECTION

This section presents the implementation to validate our proposed methodology. We first introduce the experimental setup for collecting data (both of radar and RGB-D camera) and then introduce the preprocessing method of radar data.

### 7.1 Experiment Subjects

We recruited a total of 56 participants to collect mmWave and vision data[2]. The participants consist of 26 males and 30 females, aging from 17 to 60 with various heights from 152cm to 188cm and weights from 45kg to 96kg. Fig.10 shows the cohort stats regarding the age and height, which are generally believed to have significant influence on gait patterns. As for the details on the age of subjects, we divide all subjects into small groups every five years and the boundaries are open at the front and closed at the back (i.g., ( , ]). For example, the 30 of abscissa axis in Fig.10(a) represents (25,30]. We also divide all subjects into small groups every 0.05m on the height. The cohort stats of the participates depicted in Fig.10 demonstrates that the participates in our evaluation have diversified ages and body characteristic (e.g., heights).




(a) Age groups. (b) Height groups.

Fig. 10. The cohort stats regarding the age and height. (a) Age groups. (b) Height groups.

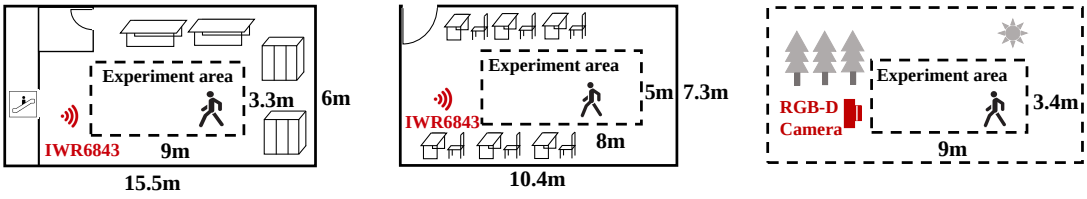
[2]The study has received the ethical approval XXX (no name for blind review).

## 7.2 Data Collection

**mmWave Radar Platform.** For the radar data collection, we utilize a commercial and off-the-shelf millimeter-wave radar IWR6843-BOOST. The radar operates in a frequency band from 60GHz to 64GHz whose wavelength is ~ 4mm. It has three transmitting antennas and four receiving antennas that form a 60 degree azimuth FoV and 60 degree elevation FoV whose angle resolution is ~ 15°. We utilize the FMCW processing chain provided by TI and the radar outputs 3D point cloud. For reproduction, the detailed configuration parameters of the device are provided as follows: the device is set to transmit 32 chirps per frame. The start frequency of the chirp is set to 60.065GHz. The frequency bandwidth is set to 3194.88MHz. The Frequency slope is set to be 12.5MHz/us.

**Experiment Site and Data Collection.** For experiments, we collect radar data in two scenes: a corridor and a room. As shown in Fig.11 (a) and (b), the millimeter-wave device is placed at one end of the corridor or room with a height of 0.9m, and participants on the other end walk towards the radar device. For multi-person scenario, the closest distance between two concurrent people is about 0.3m which is a reasonable social distance as adopted similarily in the relevant studies [24, 43]. In each experiment, a participant walks for at least 7 meters. Note that because of different people's walk speeds, the recording duration of a radar capture is different and fluctuates ranging from 4.5s to 6.5s. In order to create a valuable data set and avoid data corruption, participants are required to walk ten times or more. In summary, the dataset contains 45 different identities with each identity having 17 radar records on average. It took about 30 days to collect the data and people were free to change clothes across during this period.

**RGB-D Camera and Experiment.** To collect the dense mesh of people, we utilize Azure Kinect which is a representative type of RGB-D camera. As shown in Fig.11 (c), the RGB-D camera is placed at one end of the experiment area with a height of 1m. The sites for camera data collection are completely disjoint from the locations used for radar data collection, mimicking the real-world public and privacy-private scenarios respectively. We collected the camera data of 45 participants walking in the entrance area and outdoor environments. Each participant was required to walk a total of 20 times. Finally, we convert the camera data into the same Cartesian coordinate system as the radar data.




(a) Scene1. (b) Scene2. (c) Scene3.

Fig. 11. Experimental scenarios. (a), (b)MmWave data collection areas indoors, simulating the camera-restricted scenario. (c) Vision data collection area outdoors, simulating the camera-allowed scenario.

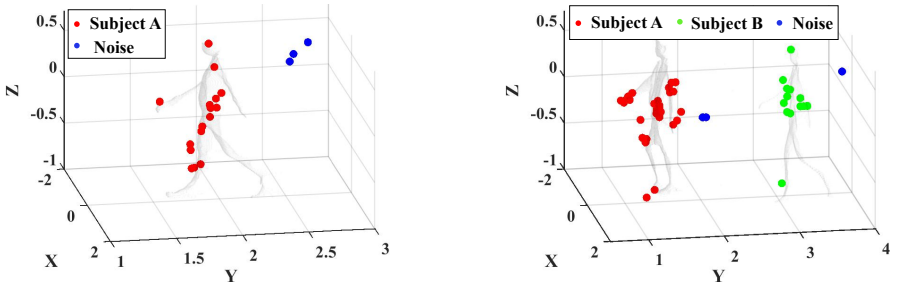
## 7.3 Preprocessing for Radar Data

While the human subjects are easy to segment from RGB-D camera data, the segmentation with radar points is much more tricky. The radar data was collected in confined indoor environments that include static obstructions such as walls, floors, and ceilings, together with the walking people. In such an environment, however, multi-path noise is non-negligible, which is a common issue for almost all RF technologies. Due to the reflection of ambient objects and beam spreading [23], the propagation of mmWave signals between objects and transceivers tends to travel through multiple paths. Consequently, unwanted points often appear in the radar point cloud which are widely known as the 'ghost points' [22]. In order to mitigate the impact of these noisy points and segment human subjects out, we implement the following two steps (1) Height Heuristics based Denoising (2) Clustering-based Point Segmentation.

*7.3.1 Denoise with Height Heuristics.* we first remove some unlikely points based on their 3D locations as we know the normal areas where people walk. For instance, when the device is placed 1m high above the ground and the coordinate system is based on the device as the origin, points with their heights greater than 2.5m or below than -1m are unlikely to be human bodies that can be safely removed.

*7.3.2 Segmentation via Clustering.* Next we apply the DBScan algorithm [5] to acquire the cluster of points of a person such that the near-person noise can be suppressed. DBScan is a density-aware clustering algorithm that can divide a point cloud based on the distance and the density described based on a set of neighborhoods in the 3D space. As it does not require the number of clusters to be specified a priori and can automatically mark outliers that are noise, DBScan has been utilized by [47] to separate individual human objects from mmWave radar point clouds. Our implementation carefully follows [47] to separate the radar points belonging to different individuals. Regarding the hyperparameters settings of DBScan, we empirically set the maximum distance (radius) between two points falling into the same cluster to 0.35 and set the minimum point number in a cluster to 3.

As shown in Fig.12, after applying the two denoising steps, radar points belonging to two different individuals can be clearly segmented.




(a) Single subject. (b) Two subjects.

Fig. 12. Segmented point clouds belonging to individual people using height heuristics and clustering.

## 8 SYSTEM EVALUATION

This section presents the performance evaluation of our system. We start with the evaluation methodology including the training and test procedure, evaluation metrics and competing approaches (Section 8.1). The overall performance is reported in Section 8.2, followed by separate evaluations of each critical design component in Section 8.3. We then extensively study the impact of various factors on the performance in Section 8.4.

### 8.1 Evaluation Methodology

As introduced in Section 7, we establish the databases for radar and RGB-D data of 56 subjects. To evaluate the identification accuracy for both single-person presence in the scene and the co-presence of multiple people at the same time, a single-person database and a multi-person database are respectively built. The former contains 655 radar records with one person in the FoV, while the latter consists of 248 records with 2 people in the FoV. In addition, to evaluate the impact of the angle of view, we collect a total of 1600 radar records with varying angles. On the other hand, the RGB-D gallery contains 1072 records in total. We have up to 17 RGB-D records for each subject, representing the real-world situation where people are typically captured by camera several times. Also, as each gait cycle typically spans about 1s, all records of both modalities contain the sensing data of several gait cycles. *For reproducibility, this multi-modal dataset used in our evaluation will be released to the community.*

*8.1.1 Model Setting and Model Training/Testing.* In this section, we describe the details of the deep metric learning model, training, and testing procedure. In pointNet, we implement a multi-layer perception (MLP), the layer sizes of which are $(x, 12, 24, 48, 64)$ with $x$ being 5 and 4 for mmWave points and RGB-D signature respectively. We use batch normalization followed by ReLU activation functions after all layers. The attention operation is implemented by fully connected layer (64,1). The LSTM has 3 layers and the size of each layer is 64.

Following the conventional strategy adopted by many cross-modal ReID studies in computer vision [13, 15], we use 75% of the data records collected from each subject for model training, and the remaining 25% serve as the testing set. The learning rate is set to 0.0002 and the batch size is 32. The number of training epochs is 50000. The hyper-parameter assigned to loss function in Equation 6 (i.e., *margin*) is set to 0.3. We implement our deep learning model in PyTorch and train the model with NVIDIA RTX 3090.

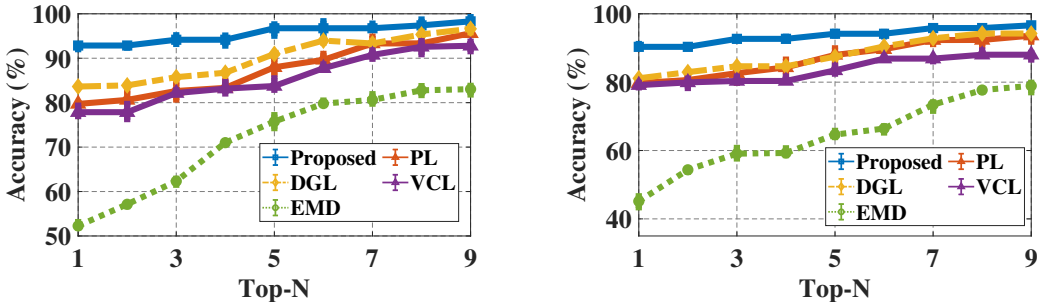
In each ReID test, we randomly select one radar record from test set as the query and our system retrieves the RGB-D record of the same identity from the RGB-D gallery. Note that these query and gallery records are not used in the training phase. Single-person ReID and multi-person ReID experiments are conducted using radar records with one and two subjects respectively. To make the result more accurate, we perform 4-fold cross-validation, where each time we randomly assign records to the training set and test set.

*8.1.2 Evaluation Metrics.* Our evaluation follows [18] and adopts the cumulative matching curve (CMC), which is a widely adopted metric in ReID studies. Specifically, for each test sample, we calculate the similarity between the radar query and every candidate's RGB-D record. The results are ranked and the top-N RGB-D records (the N most similar records) are retrieved. We report the top-N accuracy, which is defined as the percentage of test cases where the RGB-D record of the target person is ranked among the top N positions among all the RGB-D records in the test. N varies from 1 to 9 in our evaluation given the number of our volunteers.

*8.1.3 Competing Approaches.* We compare our approach with the following four baselines. Since this is the first work for cross-modal ReID among vision and radar, we port several single-modal mmWave or RGB-D (Re)ID approaches in the recent literature to our problem.

**PointNet + LSTM (PL)** [8, 30]. We adapt the single-modal mmWave points re-identification method [8] to our problem. Specifically, the method first uses PointNet [30] to extract features from dense points from raw RGB-D output as well as raw mmWave points. Secondly, the feature vector is fed to LSTM to extract temporal and spatial features. Lastly, a mean pooling operation aggregates the features to get the final gait embedding. The training method is also based on triplet loss.

**Voxelization + 3DCNN + LSTM (VCL)** [47]. We also implement mmWave person identification method [47] based on voxelization and 3DCNN. In this baseline method, a point cloud is first mapped to a 3D voxel grid, and then the 3D voxel grid is converted into a feature vector using 3DCNN. Finally, LSTM and mean pooling are utilized to obtain the gait embedding. **DGCNN + LSTM (DGL)** [38, 48]. We port the single-modal RGB-D person ReID method [48] to our problem. Specifically, the method utilizes DGCNN [38], a graph CNN to extract features from raw RGB-D outputs and mmWave points. DGCNN consumes the point cloud directly and applies the proposed EdgeConv which takes $k$ adjacent points as graph structure to extract local features. Finally, LSTM and mean pooling are utilized in the same way as VCL to extract gait embedding. **Earth Mover's Distance (EMD)** [31]. This baseline utilizes the traditional similarity estimation method (i.e., EMD) to measure the similarity between a mmWave capture and a synthesized RGB-D signature. Earth Mover's Distance (EMD) [31] is a popular method to qualify the distance between two distributions. In short, it models two distributions as a mass of earth and a collection of holes and it measures the least amount of work needed to fill the holes with earth. In our problem, both the synthesized signature points and radar points essentially capture the distribution of specular reflection energy in the space over time. Therefore, their similarity can be estimated by calculating the distance between two distributions.




(a) Single subject. (b) Multiple subjects.

Fig. 13. Overall performance of ReID (Cumulative Matching Curve). (a) Single-person ReID. (b) Multi-person ReID.

## 8.2 Overall ReID Performance

We now report the overall ReID accuracy of single-person (i.e., only one subject appears in the radar FoV) and multi-person (i.e., more than one subjects appears in the radar FoV) experiments.

*8.2.1 Single-person ReID accuracy.* As Fig.13(a) depicts, our design achieves 92.86% top-1 accuracy, 94.16% top-3, and 96.75% top-5 accuracy out of 56 subjects in the single-person ReID. In other words, given the radar capture of a person of interest, we rank 56 candidates' RGB-D meshes based on cross-modal similarity. We have a 92.86% chance to identify the target subject as the most similar candidate. The chances to have the target among the most similar 3 and 5 candidates are 94.16% and 96.75% respectively.

The performance of our method significantly outperforms the baseline approaches. First, our approach achieves 13% higher top-1 and 7% higher top-5 accuracy than the **PL** using raw RGB-D data (i.e., without being preprocessed by specular reflection model), showing that the importance of the signature synthesis (Section 6.1) in addressing modality discrepancy. Our method also outperforms other two methods ported from single-modal re-identification (i.e., **VCL** and **DGL**) especially for top-1 accuracy, improving by at least 9%. Second, compared with the traditional similarity estimator (i.e., **EMD**), our method demonstrates dramatic improvement (e.g., improving for top-1 accuracy by 39%) indicating that our similarity estimation algorithm using deep metric learning more robust against interference factors (e.g., temporal misalignment and minor variation in gaits). Finally, we also compare our method with an unsupervised learning method for cross-modal ReID. Specifically, we utilize representative methods (i.e., PointNet [30] and RNN) to extract features of internal and inter frames for radar and RGB-D point cloud, and multi-label classification loss [36] as the loss function. However, because of the manifestation gaps among heterogeneous sensor modalities, we found that the pseudo labels predicted by the unsupervised method are very error-prone on our cross-modal Re-ID problem. The top-5 accuracy of unsupervised learning evaluation on our cross-modal task is only 40% on average. This indicates that supervised learning is still a better learning fashion over unsupervised learning for cross-modal Re-ID.

*8.2.2 Multi-person ReID accuracy.* Fig.13(b) shows that our method achieves a comparable accuracy in the multi-person ReID with 90.33% top-1 accuracy, 92.67% top-3, and 94.16% top-5 accuracy. This result validates that our method is very robust even when there is more than one subject co-present in the radar FoV, demonstrating the unique advantage of our design against the WiFi-based solution [17] in the capability of simultaneously handling multiple people of interest in the real-world settings.

## 8.3 Effectiveness of Key Design Components

*8.3.1 Ablation Study.* In Section 6, several critical designs were introduced. To demonstrate their effectiveness, we measure the performance of the system when specific components are disabled. We conduct both single-person and multi-person ReID experiments with the following 4 different settings and top-5 accuracy is provided.

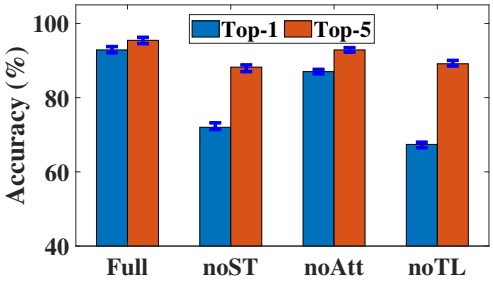



Fig. 14. Effectiveness of diff design components.

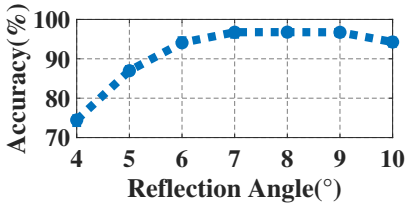



Fig. 15. Impact of reflection angle threshold.

**Full version (Full):** All the components in Section 6 are activated.
**w/o spatio-temporal signature (noST):** The similarity is computed on raw input data. The RGB-D mesh is not processed by specular reflection model (Section 6.1) to synthesize signature. This setting aims to highlight the importance of using signature points rather than raw inputs.
**w/o attention (noAtt):** The attention operation in similarity estimation module (Section 6.2) is replaced with max-pooling operation utilized in the original PointNet paper. This setting aims at examining effectiveness using the attention operation to our problem (e.g., dynamically adjust the weights of various body parts).
**w/o triplet loss (noTL):** The training strategy is replaced by two-tuples inputs with contrastive loss. In specific, the network is trained with two-tuples consisting of one sample from each modality (denoted by $< A, B >$) and the loss function is the contrastive loss [10] defined as:

$$L_c = yd^2 + (1-y)max(margin_c - d, 0)^2 \tag{7}$$

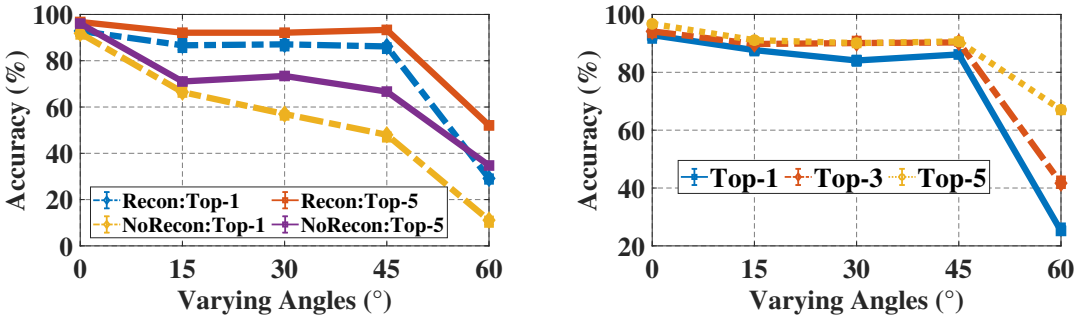
where $d$ represents the distance of $A$ and $B$. $margin_c$ is a pre-defined threshold. $y$ is the label of of the two samples, indicating whether A and B from two modalities belong to the same identity ($y = 1$) or not ($y = 0$). This setting aims to evaluate the importance of the triplet loss during training stage to the performance (Section 6.2.2).

The results are shown in Fig.14. The full version setting achieves the best results, confirming that every design component is effective to cross-modal ReID. Furthermore, compared with the results of **noST**, we observe that the signature points plays an very important role in our design. We therefore hypothesize that spatio-temporal signature points preserve the most salient characteristic of an identity while unifying the modality representation. Moreover, it eliminates the the redundancy in raw RGB-D data, and thus is easier for the network to extract human gait features. The result of **noAtt** demonstrates that attention operation also effectively improves the accuracy while the improvement on top-1 is most significant. This is because by adopting the attention mechanism, we can exploit our observation in Fig.4 - the network paies more attention to features on the limb parts which are more identity-specific than the torso that has more points due to larger RCS but contains limited identify information. Finally, the performance drop shown in **noTL** proves that the triplet loss is more effective than contrastive loss.

*8.3.2 Reflection angle threshold.* In signature synthesis (Section 6.1), we utilize a parameter $\epsilon$ as the threshold to determine signature points. An improper setting of $\epsilon$ might lead to incorrect signature points, whereas there will be fewer points if $\epsilon$ is too small. To find the optimal threshold, we examine the performance over various threshold values. As Fig.15 depicts, the optimal value of $\epsilon$ is empirically found at 7°.

## 8.4 Sensitivity Analysis

*8.4.1 Impact of view angles.* Our approach can work in real-world scenarios where RGB-D and radar sensors are installed in disjointed locations and thus have different view angles of a walking subject. This mainly benefits from our body mesh reconstruction (discussed in 6.1) on RGB-D data and rich spatial information available from radar. We therefore evaluate the performance against various view angles of both sensors.




(a) Various view angles of RGB-D camera.

(b) Various view angles of radar.

Fig. 16. Evaluation of various view angles.

**RGB-D camera view angles.** To simulate different view angles of the camera, candidates are asked to walk following a series of trajectories with varying offsets (from 0 to 1.2 meters) from the camera's mid-line (i.e., 0 meter). By doing this, the angle between the target and the device varies from 0° to 60° with the interval of 15°. On the other hand, the target subject walks towards the mmWave radar along the mid-line that is perpendicular to the radar. This simulates common RF sensing scenarios (e.g., walking in camera-restricted areas such as corridor) and it also allows the radar to perceive as much gait information as possible. The results in Fig.16(a) demonstrate that our system shows stable top-1 accuracy (86.18%) and top-5 accuracy (93.3%) at a large angle of 45° (0.9 meters offset), while accuracy drops dramatically when the angle is greater than 60° because subjects are out of the FoV of the camera and the data captured is incomplete. To further demonstrate the benefits of mesh reconstruction in mitigating view angle issues (e.g., self-occlusion), Fig.16(a) compares accuracy with and without performing reconstructions. The reconstruction improves both top-1 and top-5 accuracy by up to 26% at large view angles (i.e., 45°), showing its effectiveness in mitigating self occlusions.

**mmWave radar view angles.** We also measure the impact of radar view angles. In the experiments, the target subject walks toward the radar following the trajectories with varying offsets from the mid-line (from 0 to 1.2 meters). The angle between the subject and the radar varies from 0° to 60° with the interval of 15° to simulate different view angles. The results depicted in Fig.16(b) demonstrates that robust top-1 (> 86%) and top-3 (> 90%) accuracy across various view angles from 0° to 45° (from 0 to 0.9 meters offset). The problem becomes more difficult when the angle between subject and radar is greater than at 60° due to two challenges. First, the angular resolution and signal strength decrease dramatically at large angle. Second, identify-specific signature (e.g., arm swings) may be lost due to self-occlusion, making the signature less distinguishable. The result suggests that we can install multiple radar sensors to provide various view angles, which can be combined to improve the accuracy.

*8.4.2 Performance of subject walking away from the device.* We also evaluate the performance when the targets are walking away from the devices. Specifically, we additionally collect data of 15 candidates who asked to walk away from the devices during data collection. As Fig.17 depicts, our design achieves 90.63% top-1 accuracy, 93.75% top-3, and 96.88% top-5 accuracy. The performance of our method outperforms the baseline approaches, indicating that our design is effective when the target is walking away from the devices.

*8.4.3 Impact of the number of individual RGB-D records.* We change the number of records of a subject in the RGB-D database. This experiment mimics the situation in real life that cameras are able to capture one same subject multiple times and thereby forms a comprehensive candidate gallery. To this end, we randomly selected a different number of RGB-D records for each subject and analyze their accuracy from top-1 to top-9. Fig.18 shows the results for single-person settings. We found the availability of more RGB-D records for each subject can

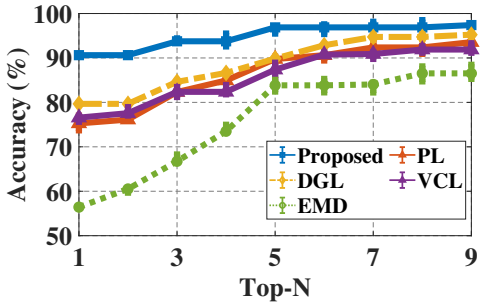



Fig. 17. Performance of subject walking away from device.

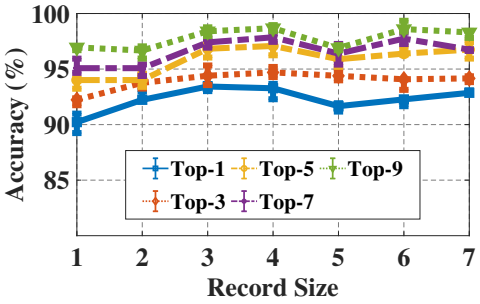
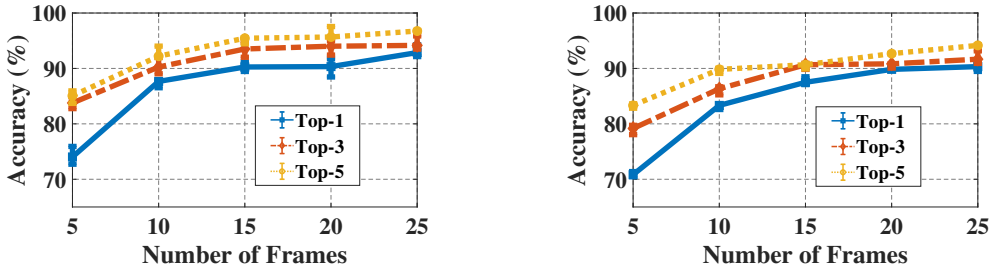



Fig. 18. Impact of the RGB-D records per subject.

increase the chances of identifying the target person. This is reasonable because more records can mitigate the randomness in a person's walking style, e.g., the small changes of the walking speeds. However, even in the case where every subject is captured by camera only once, our method still achieves a robust result (90.22% top-1 and 92.23% top-3 accuracy).

*8.4.4 Impact of number of gait frames.* Since the duration of a gait cycle is typically around 1 second, it is common that radar and RGB-D can provide the footage of the subject with more than one gait cycles. On the other hand, there are corner cases (e.g., blockage) where only a fragment of gait cycle is captured. Presumably, an increasing the number of gait frames can capture more identity-specific gait features and mitigate random errors. Thus, we repeat our evaluations by varying number of captured frames from 5 to 25 to simulate 0.5 to 2.5 cycles. Fig.19 shows that when the number of gait frames is 25 (2.5 seconds), the top-1 accuracy accuracy of single and multi-person ReID are 92.86% and 90.33%. Since it is a very mild assumption in the real world to capture 2.5 second footage in each record, the results show that our model are accurate in most regular use cases. Even with half gait cycle, the model can still achieve 74.03%, showing its robust in challenging situations.

*8.4.5 Impact of the number of candidates.* The number of candidates in the RGB-D database impacts the performance. Presumably, the identification task becomes tougher with the growing number of candidates. Our overall performance is reported with 56 subjects while in reality the number varies in the different scenarios. For example, in the domestic settings, the scale could be much smaller. Therefore, we further simulate the scenario where RGB-D captures different numbers of subjects in public areas. Fig.20 shows that the number of candidates impacts top-1 accuracy most. For example, our method can achieve more than 97.4% of the top-1 accuracy when handling 10 candidates. As the number of candidates increases to 56, the top-1 accuracy decreases to 92.86%. In contrast, the top-2 and top-3 accuracy only decrease moderately on a large candidate set. Top-3 accuracy is kept above 94% across all the settings. We can thus provide the user of the system a confidence value of the identification result based on the number of candidates.




(a) Single subject.

(b) Multiple subjects.

Fig. 19. Impact of the number of gait frames.

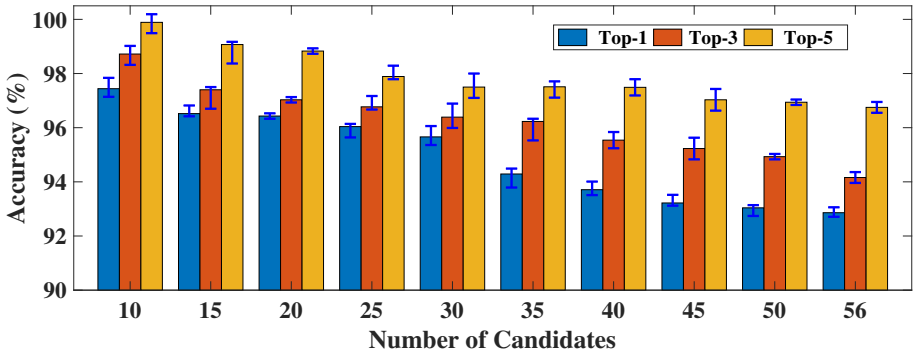



Fig. 20. Impact of the number of candidates.

*8.4.6 Impact of different mmWave radar scenes.* To evaluate the generalization performance of the model, we conduct experiments with radar data collected in two different scenes as shown in Fig.11 (a) and (b). In particular, radar data collected in scene 1 and 2 consists of 26 and 30 participants respectively while RGB-D gallery data is a complete set 56 participants. The results show that the change of scenes has little effect on ReID accuracy. In scene 1, our method achieves accuracy rate top-1 93.16% and top-5 96.06%. In scene 2, our method achieves accuracy rate top-1 92.56% and top-5 97.44%. The evaluation validates that the specular reflection signature of our design is resilient to environmental heterogeneity. The reason lies in that the mmWave sensor and our preprocessing algorithm is able to remove static reflection points from the different layout in different environments.

## 9 RELATED WORK

### 9.1 Single-modal Person Identification

Both vision-based and RF-based identification techniques have been extensively studied in the literature. In the computer vision, gait energy image (GEI) [25] is commonly utilized to extract a person's gait features (i.e., walking style) from RGB images. Depth information from RGB-D camera is also exploited to improve the identification accuracy [13, 39]. On the one hand, RF techniques (e.g., WiFi and radar) are proposed as privacy-preserving identification methods. WiFi CSI spectrum is analyzed in [46] as a gait feature for identifying a single user, while mmWave radars are adopted for multi-person identification. MU-ID [43] simultaneously recognizes up to 4 person by estimating their step lengths and duration from radar raw signal. mID [47] trains a RNN model to concurrently track and identify 2 people via radar point clouds.

In contrast to these works, our design addresses an emerging problem of re-identifying people across vision and RF sensor, which gives rise to new challenges such as inter-modality discrepancy and misalignment.

### 9.2 Cross-modal Person Identification

Cross-modal person identification has been studied by the computer vision community for years, though the majority of them focus on different types of cameras. These include the identification between depth and RGB images [12, 15], visible-to-infrared [40, 44] and across different image resolution [19], which are only applicable at camera-allowed areas. On the other side, despite its importance and ubiquity, the cross-modal ID between vision and RF signals receives little attention. XModal-ID [17] is the pioneering study of cross-modal ReID between the camera and RF. However, XModal-ID is a CSI-based technique and struggles to simultaneously identify multiple people in the same FoV. In addition, XModal-ID requires a separate WiFi transmitter and receiver, which often introduces extra deployment cost and calibration complexity. By investigating the emerging low-cost mmWave radars and RGB-D cameras, our proposed method leverages the scale and metric information provided by them and is able to achieve cross-modal ReID even in multi-person scenarios. The detailed comparison of cross-modal Re-ID system is shown in Table.1. Our system unfolds the potential of cross vision-RF human sensing and seamless human identification across camera-allowed and camera-restricted areas.

Table 1. Comparison of cross-modal Re-ID system.

| System | Scenarios of the cross-modal combination | Accuracy | Field of view coverage of this combination | Multi-person | Visual privacy protection | Ease of instal. | Power consum. |
|---|---|---|---|---|---|---|---|
| RGB – IR [40, 44] | Camera-allowed – Camera-allowed | High | Medium Wide | Yes | No | Easy | Low |
| RGB – RGBD [12, 15] | Camera-allowed – Camera-allowed | High | Medium Wide | Yes | No | Easy | Low |
| RGB – WiFi [17] | Camera-allowed – Camera-restricted | Medium | Wide | No | Yes (WiFi) | Hard | Low |
| **RGBD – Radar (Ours)** | **Camera-allowed – Camera-restricted** | **High** | **Wide** | **Yes** | **Yes (Radar)** | **Easy** | **Low** |

### 9.3 Multi-modal fusion

Our work is also related to multi-modal fusion between RF and vision sensor that has been investigated [6, 11, 41] in the topics of location, detection and tracking, etc. However, our work is distinct from sensor fusion in two aspects. First, regarding application scenarios, these work focus on the problem where different sensors are installed or appear in the same site simultaneously whereas our work allows RF and camera to be installed in the disjoint areas to satisfy different camera restrictions. Second, in technical design, fusion works commonly differentiate and associate identities through the correlation in the location and trajectory [6]. In contrast, our work studies the problem of cross-modal human re-identification (ReID) between different places where deployed with heterogeneous sensors (i.e., cross-modal querying between different places). Therefore, the challenges and technologies are different. First, we need to tackle the challenges of inter-modality discrepancy and robust similarity estimation against practical factors (e.g., minor gait change and temporal misalignment).

## 10 DISCUSSION AND FUTURE WORK

This work focuses on proof-of-principle cross vision-RF human re-identification by using low-cost mmWave radars and RGB-D cameras. There are limitations and future extensions.

**Number of people in the FoV.** Constrained by the low-level HW/SW configurations of the mmWave radar used in this work (e.g., the maximum number of points per frame), we use a two-person experiment to demonstrate the effectiveness against multiple persons present in the sensors' FoV. In future work, we will consider using more powerful commodity mmWave radars and evaluate our design on a greater number of people in the FoV.

**Blockage.** During multi-person ReID, different persons and the mmWave device may be in a straight line, which would block each other in some frames. We observe there are two categories of blockage scenarios: partial occlusion (some frames are not occluded in a sequence of gait frames) and complete occlusion (all the frames suffers from occlusion when subject is walking). As for the partial occlusion, we could use the gait frames that the subject is not occluded for ReID. These frames could be chosen using recent works of multi-subject localization and tracking with mmWave radar [47]. Through multi-target trajectory tracking, the segments in which the subjects block each other can be detected, and these gait frames can be removed accordingly. The system performance under various number of gait frames is evaluated in Section 8.4.4 - as shown in Fig.19, our method achieves a robust result (87% top-1 and 90% top-3 accuracy) when there are only 10 frames (1 second) available. The performance could drop when the number of frames is insufficient to capture the motion of one gait cycle. As for the complete occlusion which is fundamentally challenging to any kind of ReID systems due to insufficient sensory information, our proposed system cannot recognize the identities outside of its field of view. However, recent works [49] propose multi-sensor architecture where multiple sensors are installed to form a

complementary field of view. The multi-radar setup is outside the current scope of this proof-of-concept work, but in the future we plan to expand our current algorithm to multi-radar scenarios for the robust whole-scene ReID against blockage.

**Cross mmWave radar and RGB camera scenario.** We use low-cost RGB-D cameras as the visual modality, because they emerge as a promising sensor and can obtain the 3D information of walking humans. While, in real life, RGB cameras have been more widely deployed. Re-ID cross radars and RGB cameras will be further explored. We believe this is very feasible to extend as reconstructing a 3D mesh from a single or multiple RGB images has recently been an established topic [27].

## 11 CONCLUSION

This paper presents a novel system design that re-identifies a person across the data captured by mmWave radar and RGB-D cameras. To address the fundamental data discrepancy across heterogeneous sensors, we proposed a novel signature synthesis method based on the observed specular reflection model. Our system also features an effective cross-modal deep metric learning method to handle the interference factors caused by unsynchronized data across radars and cameras. We believe our system unfolds the potential of cross vision-RF human sensing and envision it to serve as a key solution to seamless human identification across camera-allowed and camera-restricted areas.

## ACKNOWLEDGEMENTS

This work was supported in part by China National Key R&D Program 2018YFB2100302, National Natural Science Foundation of China under Grant No. 61902066 and Natural Science Foundation of Jiangsu Province under Grant NoBK20190336.

## REFERENCES

[1] 2022. Kinect + Refinement. https://www.depthkit.tv/tutorials/azure-kinect-microsoft-volumetric-capture-depth-workflow-depthkit.
[2] 2022. wholehome-ai-sensor. https://consumer.huawei.com/cn/wholehome/ai-sensor/.
[3] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–13.
[4] Sherif Sayed Ahmed and Lorenz-Peter Schmidt. 2012. Illumination of humans in active millimeter-wave multistatic imaging. In *2012 6th European Conference on Antennas and Propagation (EUCAP)*. IEEE, 1755–1757.
[5] Derya Birant and Alp Kut. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & knowledge engineering* 60, 1 (2007), 208–221.
[6] Siyuan Cao and He Wang. 2018. Enabling public cameras to talk to the public. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–20.
[7] Pratik Chattopadhyay, Shamik Sural, and Jayanta Mukherjee. 2014. Frontal gait recognition from incomplete sequences using RGB-D camera. *IEEE Transactions on Information Forensics and Security* 9, 11 (2014), 1843–1856.
[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] Muqing Deng, Cong Wang, Fengjiang Cheng, and Wei Zeng. 2017. Fusion of spatial-temporal and kinematic features for gait recognition with deterministic learning. *Pattern Recognition* 67 (2017), 186–200.
[10] Weifeng Ge. 2018. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 269–285.
[11] Xiao-peng Guo, Jin-song Du, Jie Gao, and Wei Wang. 2018. Pedestrian detection based on fusion of millimeter wave radar and vision. In *Proceedings of the 2018 International Conference on Artificial Intelligence and Pattern Recognition*. 38–42.
[12] Frank Hafner, Amran Bhuiyan, Julian FP Kooij, and Eric Granger. 2018. A cross-modal distillation network for person re-identification in rgb-depth. *arXiv preprint arXiv:1810.11641* (2018).
[13] 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.
[14] Chengkun Jiang, Junchen Guo, Yuan He, Meng Jin, Shuai Li, and Yunhao Liu. 2020. mmVib: micrometer-level vibration measurement with mmwave radar. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[15] 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.

[16] Abdelwahed Khamis, Branislav Kusy, Chun Tung Chou, Mary-Louise McLaws, and Wen Hu. 2020. RFWash: a weakly supervised tracking of hand hygiene technique. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 572–584.

[17] 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.

[18] 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.

[19] 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.

[20] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kunpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. 2020. Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4, 4 (2020), 1–28.

[21] Xiaochen Liu, Yurong Jiang, Puneet Jain, and Kyu-Han Kim. 2018. Tar: Enabling fine-grained targeted advertising in retail stores. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 323–336.

[22] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. 2020. See through smoke: robust indoor mapping with low-cost mmWave radar. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*. 14–27.

[23] Chris Xiaoxuan Lu, Muhamad Risqi U Saputra, Peijun Zhao, Yasin Almalioglu, Pedro PB de Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. 2020. milliEgo: single-chip mmWave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 109–122.

[24] Zhen Meng, Song Fu, Jie Yan, Hongyuan Liang, Anfu Zhou, Shilin Zhu, Huadong Ma, Jianhua Liu, and Ning Yang. 2020. Gait recognition for co-existing multiple people using millimeter wave sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 849–856.

[25] 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.

[26] Mauricio Pamplona Segundo, Sudeep Sarkar, Dmitry Goldgof, Luciano Silva, and Olga Bellon. 2013. Continuous 3D face authentication using RGB-D cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 64–69.

[27] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. 2019. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9964–9973.

[28] Amol S Patwardhan. 2017. Hostile behavior detection from multiple view points using RGB-D sensor. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 1–6.

[29] Jacopo Pegoraro, Francesca Meneghello, and Michele Rossi. 2020. Multiperson continuous tracking and identification from mm-wave micro-Doppler signatures. *IEEE Transactions on Geoscience and Remote Sensing* 59, 4 (2020), 2994–3009.

[30] 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.

[31] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 1998. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE, 59–66.

[32] 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*. 145–157.

[33] Xudong Sun, Lei Huang, and Changping Liu. 2018. Multimodal face spoofing detection via RGB-D images. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2221–2226.

[34] Yangfan Sun, Renlong Hang, Zhu Li, Mouqing Jin, and Kelvin Xu. 2019. Privacy-preserving fall detection with deep learning on mmWave radar signal. In *2019 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 1–4.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[36] Dongkai Wang and Shiliang Zhang. 2020. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10981–10990.

[37] Xinglu Wang. 2020. Adversarial Multi-scale Feature Learning for Person Re-identification. *arXiv preprint arXiv:2012.14061* (2020).

[38] 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 Transactions On Graphics (tog)* 38, 5 (2019), 1–12.

[39] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. 2017. Robust depth-based person re-identification. *IEEE Transactions on Image Processing* 26, 6 (2017), 2588–2603.

[40] 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.
[41] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–22.
[42] Shi Yan, Chenglei Wu, Lizhen Wang, Feng Xu, Liang An, Kaiwen Guo, and Yebin Liu. 2018. Ddrnet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns. In *Proceedings of the European conference on computer vision (ECCV)*. 151–167.
[43] 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.
[44] 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.
[45] Ling-Fung Yeung, Zhenqun Yang, Kenneth Chik-Chi Cheng, Dan Du, and Raymond Kai-Yu Tong. 2021. Effects of camera viewing angles on tracking kinematic gait patterns using Azure Kinect, Kinect v2 and Orbbec Astra Pro v2. *Gait & Posture* 87 (2021), 19–26.
[46] 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)*. 1–12. https://doi.org/10.1109/IPSN.2016.7460727
[47] 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)*. IEEE, 33–40.
[48] Zhedong Zheng and Yi Yang. 2020. Person re-identification in the 3d space. *arXiv preprint arXiv:2006.04569* (2020).
[49] Anfu Zhou, Shaoyuan Yang, Yi Yang, Yuhang Fan, and Huadong Ma. 2019. Autonomous environment mapping using commodity millimeter-wave network device. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 1126–1134.
[50] Xinxin Zuo, Sen Wang, Minglun Gong, and Li Cheng. 2021. Unsupervised 3d human mesh recovery from noisy point clouds. *arXiv preprint arXiv:2107.07539* (2021).