Learning Longterm Representations for Person Re-Identification Using Radio Signals

Lijie Fan* Tianhong Li* Rongyao Fang* Rumen Hristov Yuan Yuan Dina Katabi
MIT CSAIL

Abstract

Person Re-Identification (ReID) aims to recognize a person-of-interest across different places and times. Existing ReID methods rely on images or videos collected using RGB cameras. They extract appearance features like clothes, shoes, hair, etc. Such features, however, can change drastically from one day to the next, leading to inability to identify people over extended time periods. In this paper, we introduce RF-ReID, a novel approach that harnesses radio frequency (RF) signals for longterm person ReID. RF signals traverse clothes and reflect off the human body; thus they can be used to extract more persistent human-identifying features like body size and shape. We evaluate the performance of RF-ReID on longitudinal datasets that span days and weeks, where the person may wear different clothes across days. Our experiments demonstrate that RF-ReID outperforms state-of-the-art RGB-based ReID approaches for long term person ReID. Our results also reveal two interesting features: First since RF signals work in the presence of occlusions and poor lighting, RF-ReID allows for person ReID in such scenarios. Second, unlike photos and videos which reveal personal and private information, RF signals are more privacy-preserving, and hence can help extend person ReID to privacy-concerned domains, like healthcare.

1. Introduction

Person re-identification (ReID) aims to match a person-of-interest across different cameras, and at different times and locations. It has broad applications in city planning, smart surveillance, safety monitoring, etc. It is challenging because the visual appearance of a person across cameras can change dramatically due to changes in illumination, background, camera view-angle, and human pose. With the success of deep learning, several ReID models [3, 25, 7, 8, 14, 39, 11, 60] have managed to extract appearance features that are view-invariant across cameras, leading to good performance on various person ReID datasets [6].

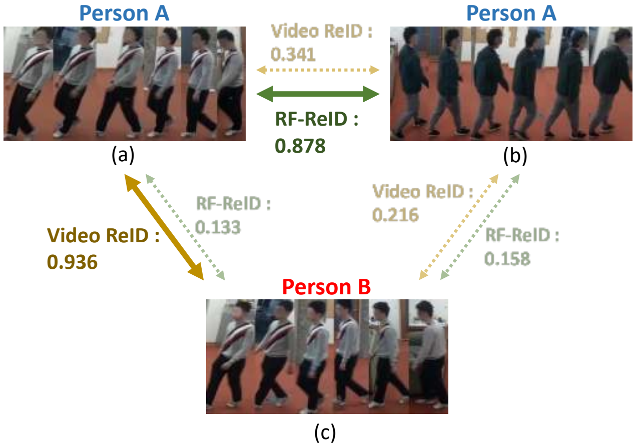

Figure 1. Similarity score between people as computed by RF-ReID and a state-of-the art video-based ReID model (the larger the score, the higher the similarity). (a) and (b) show the same person wearing different clothes, and (c) shows a different person wearing the same pullover as the top-left person. The video-based ReID model relies on appearance features and thus wrongly predicts (a) to be close to (c), while RF-ReID captures body shape and walking style and can correctly associate (a) with (b).

However, another great challenge has rarely been investigated: human visual appearance can also change drastically over time. For example, people may revisit the same shop on different days wearing different clothes and hair styles, and a thief can deliberately change his clothes to mislead the surveillance system. A robust person ReID system should be able to match people despite appearance changes. Unfortunately, existing RGB-based person ReID systems have severe limitations in achieving this goal since they intrinsically rely on appearance information such as clothes, shoes, hair, bags, etc. [39, 49, 52]. All these features are short-lived and can become ineffective the next day. To achieve robust person ReID, the system should be able to extract longterm identifying features that persist for weeks and months.

But how can we capture persistent features suitable for person ReID? Wireless signals present a good solution. Radio frequency (RF) signals in the Wi-Fi frequency range traverse clothes and reflect off the human body. Unlike cameras, wireless signals could extract intrinsic features of the human body, such as body size or shape. These features are relatively stable over days and months, enabling a more

*Indicates equal contribution.

robust longterm ReID system. Moreover, previous works have shown the possibility of tracking people's 3D skeletons and walking patterns using RF signals, which can be used as longterm identifying features [55, 24, 56, 54, 33]. Furthermore, unlike RGB-based person ReID methods which struggle in the presence of occlusion and poor lighting, RF signals traverse walls and can enable ReID through occlusions and in dark settings.

However, applying RF signals to person ReID presents several challenges. First, training an end-to-end model only using people's IDs as labels leads to overfitting problems. ID labels provide a rather weak supervision. The model will tend to learn environment-dependent short-cuts such as the position a person usually stays at. Second, unlike RGB images, a single RF snapshot contains reflections from few body parts, and misses the rest of the body. This is due to a special property called specularity [4]. The human body acts like a mirror for RF signals in the Wi-Fi range, and signals reflected away from the receiver will not be captured. As a result, a single RF snapshot does not contain enough information about the whole body to identify the person.

To solve the overfitting problem, we propose a multi-task learning framework and an environment discriminator. Apart from predicting the identity of the person, we force features from our model to contain enough information to predict the 3D skeleton of the person. We further add an environment discriminator to force the features to be environment independent. This discriminator is co-trained with the feature extraction network in an adversarial way, making features from different environments indistinguishable by the discriminator. To solve the specularity of RF signals, we add a hierarchical attention module that effectively combines information regarding the human shape and walking style across time, i.e., across multiple RF snapshots.

We introduce RF-ReID, an RF-based person ReID model that can extract longterm identifiable features, and work under occluded or inadequate lighting conditions. RF-ReID takes wireless signals as input, extracts identifiable features from the reflection of the human body, and identifies the person with the extracted feature. It performs robust person ReID across both time and space. Figure 1 shows an examples that demonstrates the effectiveness of RF-ReID. The same person wearing different clothes in (a) and (b)) is mistaken as a different individual by state-of-the-art video ReID [11], while two different people wearing the same clothes in (a) and (c) are wrongly identified to be the same person. In contrast, RF-ReID can accurately identify (a) and (b) to be the same person, and (a) and (c) to be different.

We evaluate RF-ReID on two datasets. **(A) RRD-Campus:** The first dataset is collected using five radios deployed in different locations on our campus. Ground truth ID labels are collected using RGB video cameras colocated with each radio. The resulting dataset contains 100 different IDs, and spans 15 days. People appear multiple times in different clothes and different people may wear similar clothes. **(B) RRD-Home:** The second dataset was originally collected to assess the viability of tracking movements of Parkinson's patients in their homes using RF signals. It includes data from 19 homes, and an average of one week per home. Ground truth IDs are obtained through manual labeling by comparing the movements from wearable accelerometers with movements from RF signals. More details about the datasets are available in section 5.

We train a single model to ReID people in both datasets. As described above we use a discriminator to ensure that the representation is environment independent. Since RRD-Campus includes colocated RGB videos, we compare our model with state-of-the-art video-based ReID[11]. The results show that RF-ReID outperforms the state-of-the-art RGB-based ReID approaches by a large margin. They also show RF-ReID ability to work through occlusions and in poor lighting conditions when the camera fails completely. RF-ReID also works well on RRD-Home, which contains RF signals from the homes of Parkinson's patients. This result shows that our method can ReID people in private settings like homes, locker rooms, and other private locations where one cannot deploy cameras. Interestingly, this result leads to a new concept – privacy-conscious ReID, where people may be re-identified without capturing their personal information, e.g., pictures or audio clips.

To summarize, this paper makes two key contributions:

- First, it investigates the task of longterm person ReID, which identifies a person regardless of appearance changes over time. It proposes a novel model that leverages RF signals to achieve longterm person ReID. It further demonstrates that the model is robust to occlusion and poor lighting.
- Second, it introduces the concept of privacy-conscious ReID as the ability to identify encounters with the same person without collecting personal or private data like pictures, videos, or audio clips. The paper also demonstrates the first such privacy-conscious ReID model.

2. Related Works

(a) RGB-based ReID. There are mainly two categories of RGB-based ReID: image-based and video-based. Early approaches for image-based ReID rely on hand-crafted features based on color descriptors, and optimize some distance metric relative to those descriptors [12, 26, 37, 28, 42]. Early video-based ReID models use spatio-temporal descriptors like HOG3D [21] and gait energy image (GEI) [13] to extract additional temporal information.

Recent approaches rely on deep learning and can be divided into two categories. The first category uses classification models similar to image or video classification

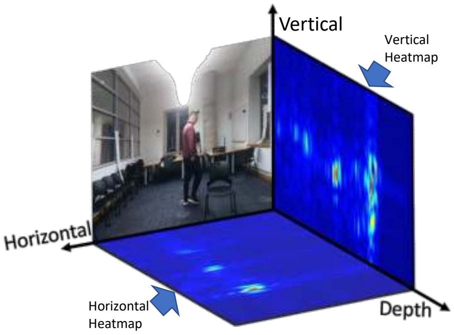

Figure 2. RF heatmaps and an RGB image recorded at the same time.

tasks [23, 5]. The second category uses siamese models which take a pair or triplet of images or videos as input, and uses pairwise or triplet loss to train the model [3, 14, 8]. Some Video-based ReID models further aggregate temporal information through temporal attention or RNN networks [11, 29, 32, 47, 48].

Both image-based and video-based ReID methods tend to extract short-lived features such as clothes, shoes, bags, hair styles, etc. [6] and hence struggle to re-identify a people across days and weeks.

(b) RF-based Person Identification. Research in wireless systems has explored person identification using radio signals. Previous work can be divided into two categories. The first category uses the signals transmitted by portable devices (e.g., cellphones) to track and identify each person [22, 46, 41]. Such systems require a person to wear or carry sensors, which limits their utility and robustness. The second category analyses signal reflections off people's bodies to identify each person [17, 2, 16, 18, 35, 36, 44, 45, 15, 38, 43, 50]. Past systems however classify small number of people (<10) in the same environment, cannot generalize to new identities unseen in the training set, and typically require the person to walk on certain constrained paths [43, 50]. In contrast to all past work, we are the first to achieve person ReID in the wild with RF signals, and without requiring people to wear sensors or move on specific paths. Furthermore, our model generalizes to new people and new environments unseen during training.

3. Radio Frequency Signals Primer

We use an FMCW radio widely used in previous work on RF-based human sensing [56, 27, 51, 9, 34, 40, 17, 53, 57, 55, 24]. The radio is equipped with two antenna arrays: horizontal and vertical. It operates between 5.4 and 7.2 GHz and can sense people up to 12m away from the device.

(a) RF Heatmaps: The RF signal at the output of the radio takes the format of two 2D heatmaps: one from the horizontal array and the other from the vertical array, as shown in Figure 2 (red refers to large values while blue refers to small values). The horizontal heatmap is a projection of RF signals on the plane parallel to the ground, and the vertical heatmap is a projection of RF signals on a plane perpendicular to the ground. Intuitively, we can treat these heatmaps as depth maps, where higher values correspond to higher strength of signal reflections from a location. The radio generates 30 horizontal-vertical heatmap pair per second; we call each pair an RF frame.

Figure 2 reveals that RF signals have different properties from vision data. The human body is specular in our frequency range [4]. RF specularity occurs when the wavelength of the signal is larger than the roughness of the object's surface. In this case, the object acts like a mirror as opposed to a scatterer. The signal from each body part may be reflected towards our sensor or away from it depending on the orientation. Therefore, each RF frame contains reflections from a subset of body parts, making it hard to obtain identifiable information from a single RF frame.

(b) RF Tracklets: Prior work has demonstrated that RF signals can be used to detect, localize and track people [1]. We use this technique to extract RF tracklets from the RF heatmaps. As shown on the left of Figure 3, a tracklet extracts from the horizontal and vertical heatmaps the RF signals reflected off a person, and the bounding box of that person at each time step (white box in figure). Since one RF tracklet always corresponds to one person, the ReID task is performed across different RF tracklets.

(c) Skeletons from RF: 3D Human skeletons can be generated from RF signals using the approach in [56]. The generated 3D skeleton data contains the 3D coordinates of 18 major human body joints at each time step, as specified in [10], which can be used to assist the task of person ReID.

4. RF-ReID

RF-ReID is an end-to-end model for person ReID using RF signals. As shown in Figure 3, our model takes an RF tracklet as input. It then extracts features from the tracklet using an RF feature extraction network. It then aggregates temporal information through a learnable hierarchical attention module to generate a feature map. During training, these features are supervised in a multi-task learning manner using identity classification loss, triplet loss and skeleton loss. We further add an additional environment discriminator loss to force the model to learn environment-invariant features. This allows our model to generalize to new environments not seen during training. Below, we explain each RF-ReID component in detail.

4.1. RF Feature Extraction Network

Since RF tracklets can have different durations, we first perform temporal sampling on each tracklet before extracting features from it. For each RF tracklet, we uniformly sample 25 segments from it, where each segment contains 3 seconds (90 frames) of RF heatmaps.

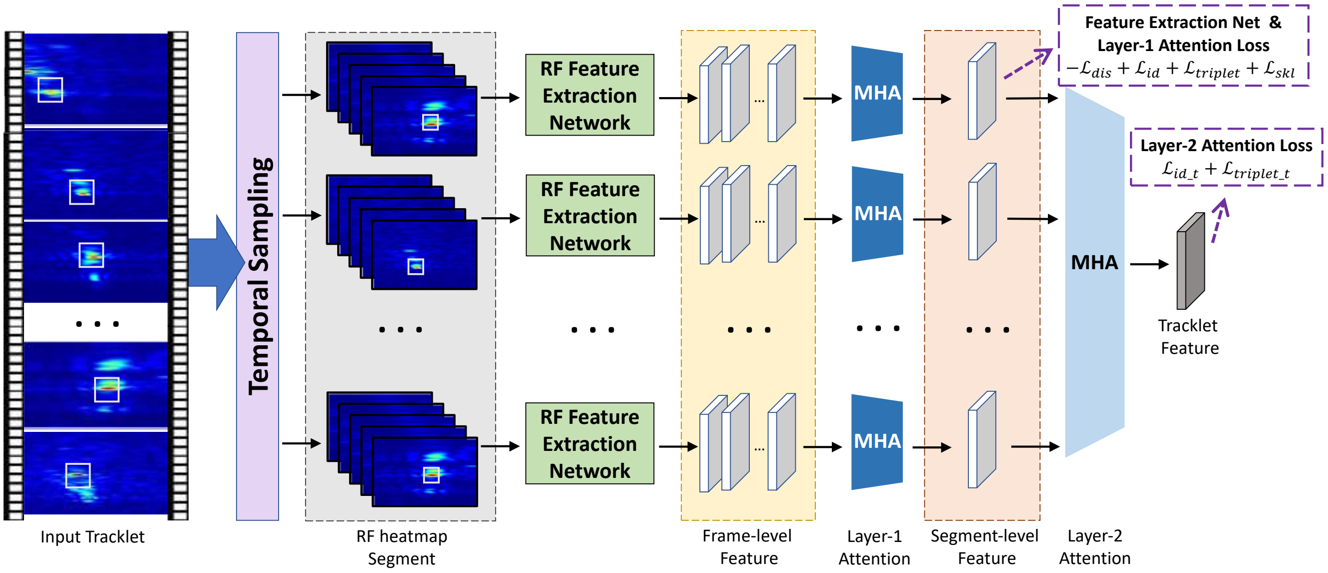

Figure 3. Model architecture. RF-ReID takes an RF tracklet as input. It samples RF segments of 3 seconds (90 frames), and extracts frame-level features with an RF feature extraction network (shown in green). These features are processed by a multi-headed hierarchical attention module (i.e., MHA) with two sub-modules; the first attention sub-module (dark blue) extracts segment-level features and the second sub-module (light blue) extracts tracklet-level features. The supervision for training the RF feature extraction network and the first attention sub-module is added to the segment-level features, and the supervision for the second attention sub-module is added to the tracklet features.

We adopt an architecture similar to [56] for our backbone feature extraction network. The network first uses spatio-temporal convolutions to extract global features from the input RF frames. We then crop out the region of interest around the tracklet trajectory in the feature map. Finally, the cropped features are fed into a sub-network to generate frame-level identifiable features for person ReID.

4.2. Hierarchical Attention Module

The RF feature extraction network generates relevant features from each RF segment, where a segment is a clip of 90 RF frames (3 seconds). As mentioned in Section 3, due to specularity, each RF frame contains information only about some body parts. So we need to aggregate features across frames in the same tracklet. To solve this problem, we propose a learnable two-step hierarchical attention module to aggregate information across each tracklet.

There are two kinds of information in an RF tracklet that is relevant to person identification: shape and walking style. The coarse shape of the person can be obtained by aggregating information from several seconds of RF signals. This is because when a person moves, we can receive signals reflected from different body parts according to their orientation with respect to the radio. Thus, the first attention block is added on frame-level features to aggregate the shape information within each 90-frame segment (3 seconds).

The walking style, on the other hand, is a feature that can only be inferred from a longer temporal span. However, within an RF tracklet, there can be many non-walking periods where the person may stop, stand by, sit down, tie their shoes, etc. Those periods cannot be used to infer the walking style. Therefore, we use the second attention block to attend to features from different segments across the tracklet and aggregate them to generate one final feature vector for each tracklet.

4.3. Multi-task Learning for Identifiable Features

To train the RF feature extraction network, we add supervision to the segment-level features (orange box in Figure 3). As shown in Figure 4, we first add two losses widely used in prior works on RGB-based person ReID: identification loss $\mathcal{L}_{id}$ and triplet loss $\mathcal{L}_{triplet}$. For the identification loss, the segment-level features are further passed through another classification network with two fully-connected layers to perform ID classification. This task helps the model learn human identifiable information from RF signals. The triplet loss [14] is computed as

$$\mathcal{L}_{triplet} = \max(d_p - d_n + \alpha, 0),$$

where d_p and d_n are the L_2 distances of segment-level features from the same person and features from different people, respectively. α is the margin of triplet loss (α is set to 0.3). This loss enforces features from different people to be far away from each other and those from the same person to be close.

The first attention layer can be trained end-to-end with the feature extraction network. For the second attention layer, we first generate features for each segment, then we use the second attention layer to aggregate them and train the second attention layer using the ID loss $\mathcal{L}_{id\_t}$ and triplet loss $\mathcal{L}_{triplet\_t}$ on the aggregated feature.

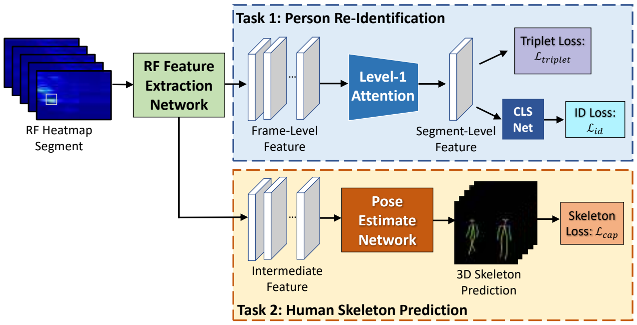
Task 1: Person Re-Identification
RF Heatmap Segment
RF Feature Extraction Network
Frame-Level Feature
Level-1 Attention
Segment-Level Feature
Triplet Loss: $\mathcal{L}_{triplet}$
CLS Net
ID Loss: $\mathcal{L}_{id}$
Intermediate Feature
Pose Estimate Network
3D Skeleton Prediction
Skeleton Loss: $\mathcal{L}_{cap}$
Task 2: Human Skeleton Prediction

Figure 4. Illustration of multitask learning in RF-ReID. The blue box corresponds to the the task of person re-identification, and the yellow box corresponds to the task of 3D skeleton prediction.

Besides these two losses, we force the model to learn to infer the person's skeleton from RF signals and add an additional skeleton loss $\mathcal{L}_{skl}$ for supervision. Specifically, we take the intermediate features in our feature extraction network two layers above the frame-level features and feed them into a pose-estimation sub-network that generates 3D human skeletons. The skeleton loss is a binary cross entropy loss similar to the one used in [56]. This loss forces the features to contain enough information for skeleton generation, which can help the ReID task in capturing person's height and walking style. The skeleton loss also acts as a regularizer on the extracted features to prevent overfitting.

4.4. Environment Discriminator

RF tracklets can have identifiable patterns strongly related to the environment. For example, a person is much more likely to enter his or her own office than other people. As a result, the model can use such environmental features as shortcuts to identify people based on their paths. This will severely harm the model's ability to generalize across different environments or identify people when they do not follow their usual paths. For example, in our Campus Dataset, the model will have difficulty identifying a faculty who is visiting another faculty in their office. In our Home Dataset, the model will learn the specific path that each person walks in their home, and fail to generalize to new unseen homes.

To solve this problem, we need to eliminate the environment-dependent factors during the training process. Thus, we consider signals from each radio location as one environment, and train a discriminator to predict the environment of the signal. The discriminator is trained in an adversarial way so that eventually the model will eliminate the features that are environment dependent. The discriminator operates on segment-level features as shown in Figure 3. A cross entropy loss is used to train the discriminator to predict the environment. The discriminator loss is subtracted from the loss of multi-task training in the feature extraction network. Denoting the RF feature extraction network as F and the environment discriminator as D, the discriminator loss is:

$$\mathcal{L}_{dis} = -\sum_{c=1}^{M} y_c \log(D(F(x))_c),$$

where x is the input tracklet, M is the total number of environment and y_c is the binary indicator of which environment the current tracklet belongs to. The optimization target is:

$$\min_F \max_D V(F, D) = -\mathcal{L}_{dis} + \mathcal{L}_{id} + \mathcal{L}_{triplet} + \mathcal{L}_{skl}$$

5. Dataset

We use the following two datasets.

5.1. RRD-Campus

RRD-Campus refers to an RF-ReID Dataset collected on our Campus. It contains RF signals and synchronized RGB video data. The video data is used both for ground truth labeling and to compare with RGB-based video and image based ReID methods.

The data is collected by deploying 5 radios in 5 different locations across our campus and collecting data for 15 days. Each radio is colocated with an RGB video camera. We synchronize the video data and the RF signals using the NTP protocol, with a maximum time difference of 10 ms.

Labeling: We deploy our data collection system at places that tend to be revisited by the same people, such as the lounge area in front of a lab or a set of offices. We collect video with 960 × 720 resolution and 15 FPS to assist in the labeling process. We then use the video to label all people who show up repeatedly, even if they have different clothes. We further ask the people in the dataset to double-check their own data and make sure that all RF tracklets with their identity are labeled correctly.

Statistics: The dataset contains 100 identities in total. On average, each identity has 8.63 RF tracklets, and each tracklet spans over 11.0 seconds. People in the dataset may change clothes across different days, and our data collection system does not interfere with people's normal activities.

5.2. RRD-Home

RRD-Home is based on the dataset in previous work on Parkinson's Disease analysis with RF signals [20]. The dataset is collected by deploying RF devices in 19 different homes to obtain RF tracklets for moving people, where each home is inhibited by a Parkinson patient and a normal person (typically the spouse). The Parkinson patient is asked to wear an accelerometer to collect the acceleration data. The collected dataset contains the RF tracklets for the people in each home, and the corresponding accelerometer data for the patients. Note that the acceleration data is only used for labeling and is not an input to our model.

Labeling: Accelerometers are widely used for in-home behavior analysis [17]. In RRD-Home, We use accelerometer data to assist in the labeling process. In each home, we first calculate the location, moving distance and speed for each RF tracklet, and then associate the patient's accelerometer data with the corresponding RF tracklets based on the similarity of movement properties. RF tracklets whose motion is synced with the acceleration data from the wearable on the patient are labeled with the patient's ID, whereas the tracklets whose motion does not match the acceleration data are labeled with the ID of the other person in the home. (We ignore RF tacklets from periods during which there is no acceleration data.)

Statistics: The dataset contains 38 different identities in 19 different homes. The data spans 127 days with an average of one week per home. Each identity has 165.91 RF tracklets on average. Each RF tracklet spans over 9.24 seconds. People in the dataset can change clothes across days, and the data collection system does not interfere with people's normal activities.

5.3. RF-Based Skeletons

As mentioned in Section 4.3, to assist training our ReID model we force it to learn features useful for inferring a person's skeleton from RF signals. This is done by leveraging the datasets from past work on 3D RF-based pose and action estimation [24, 54].

6. Experiments

We evaluate the effectiveness and practicality of RF-ReID and compare it with image and video based ReID.

6.1. Experimental Setup

Training & Testing: **(1) RRD-Campus.** We split RRD-Campus into a training set with 60 identities and a test set with the other 40 identities. As common in RGB-based ReID datasets, we randomly select one sample from each identity in the test set to build the query set and group the remaining samples as the gallery set. **(2) RRD-Home.** The training and test sets of RRD-Home contain 13 and 6 different homes, respectively. Each home has 2 different identities corresponding to the two inhabitants. The query set and the gallery set is constructed similarly to RRD-Campus. **(3) RRD.** We also combine RRD-Campus and RRD-Home to form a larger dataset RRD. The training set of RRD is the combination of RRD-Campus training set and RRD-Home training set. We evaluate our model on both the individual datasets and the combined one. We perform 5-fold cross-validation, where each time we randomly assign identities to the training set and test set.

Evaluation Metrics: During testing, the query samples and gallery samples are encoded to feature vectors using RF-ReID. Then we calculate the cosine distance between each query sample's features and each gallery sample's features and rank the distance to retrieve the top-N closest gallery samples for each query sample. We compute the standard evaluation metrics for person ReID based on the ranking results: mean average precision score (mAP) and the cumulative matching curve (CMC) at rank-1 and rank-5.

RGB Baseline Models: To demonstrate the effectiveness of RF-ReID, we compare it with one state-of-the-art image-based person ReId model [30] and one state-of-the-art video-based model [11]. The comparison is performed on RRD-Campus since only RRD-Campus is collected with synchronized RGB videos. We first train the image-based and video-based person ReID models on the commonly used Market1501 [59] and MARS [58], respectively. We then fine-tune them on the RGB video data in our training set. To fine-tune the video model, we use videos snippets that correspond to the RF tracklets in the training set, and to fine tune the image-based model we use the corresponding images in the training set. During testing, we evaluate the performance of RGB-based models on the same query and gallery set we used to evaluate RF-ReID. For video-based ReID, the input is the corresponding RGB video of each tracklet. For image-based ReID, we compute the features for each frame in the video, and average them to get the features of the sample.

6.2. Quantitative Results

We compare RF-ReID with state-of-the-art image-based and video-based person ReID models on RRD-Campus. As shown in Table 1, our RF-ReID model exhibits a significant improvement over both image-based and video-based models. This is mainly because people in RRD-Campus tend to wear different clothes on different days. Traditional RGB-based ReID models focus on extracting features from clothes, and fail when the same person wears different clothes. In contrast, RF-ReID focuses on the shape and walking style of a person, which remain valid over a long period.

We also report the performance of RF-ReID on RRD-Home. Due to privacy reasons, this dataset does not include RGB images or videos and hence we cannot compare with RGB-based baselines. In contrast, since RF signal is privacy preserving, it is used to track people in their homes. The results from RDD-Home in Table 1 show that our model not only achieves high accuracy on RRD-Campus, but also performs well in real-world home scenarios. Furthermore, since humans cannot recognize a person from RF signals, our model can be used to ReID people without collecting personal information like images or videos of people performing private activities in their homes.

The results in Table 1 highlight the following points:

- RF-ReID works well for long term re-identification that spans days, weeks, or longer.
- RF-ReID can re-identify people without collecting or exposing any information that can be used by a human

Methods	Modality	RRD-Campus			RRD-Home		
		mAP	CMC-1	CMC-5	mAP	CMC-1	CMC-5
Luo *et al.* [30]	RGB Image	41.3	61.4	84.3	-	-	-
Gao *et al.* [11]	RGB Video	48.1	69.2	89.1	-	-	-
RF-ReID (separate)	RF Signal	59.5	82.1	95.5	46.4	74.6	89.5
RF-ReID (combined)	RF Signal	**60.7**	**83.6**	**96.5**	**49.4**	**75.8**	**92.5**

Table 1. Comparison between RF-ReID and RGB-based and Video-based ReID on RRD-Campus and RRD-Home. RF-ReID (separate) is trained and tested on RRD-Campus and RRD-Home separately. RF-ReID (combined) is trained on both RRD-Campus and RRD-Home (i.e., the RRD dataset) and tested on both of them.

to recognize people; we refer to this property as privacy conscious ReID system. This property is critical in healthcare applications and clinical research, where one needs to ReID the subjects so that one may track changes in a patient's health over time or as a result of treatment; yet, it is essential to keep the subjects de-identified and avoid collecting or storing data that exposes the subjects' personal information.

- Last, the results indicate that ReID is harder for RDD-Home than RDD-Campus. We believe this is due to two reasons: First, tacklets from homes are shorter (9.2 vs. 11 seconds). Second, the walking style of Parkinson's patients may differ through the day as the impact of medications wears off and they need to take the next dose.

6.3. Ablation Study

We conduct several ablation studies to evaluate the contribution of each component of RF-ReID. All ablation results are for RRD-Campus.

Multi-task Learning: Traditional ReID methods use only triplet loss and ID classification loss. In our RF-ReID model, we have added an additional skeleton loss both for regularization and to learn human-related information. We evaluate the performance of this skeleton loss. Table 2 shows that adding the skeleton loss improves RF-ReID's accuracy. We also report the commonly used metric: Mean Per Joint Position Error (MPJPE) [19, 55], to evaluate the accuracy of the generated skeletons. As shown in the last column of Table 2, the features from RF-ReID contain enough information to generate accurate human skeletons.

Method	mAP	CMC-1	CMC-5	MPJPE
w/o skl loss	57.3	78.2	94.4	-
w/ skl loss	**59.5**	**82.1**	**95.5**	**7.44**

Table 2. Performance of RF-ReID with and without skeleton loss.

Hierarchical Attention Module: RF-ReID has a hierarchical attention module to aggregate features across a tracklet. The attention module has two blocks: the first aggregates shape information within each segment (3 secs), while the second aggregates walking-style information across the whole tracklet. Table 3 demonstrates the effectiveness of each block. If both blocks are replaced by average pooling over the temporal dimension, the performance would drop by 4.1% for mAP and 7.6% for CMC-1. Each block increases the mAP by 3∼4%, and adding them together achieves the highest performance.

Method	mAP	CMC-1	CMC-5
Avg Pool+Avg Pool	55.4	74.5	93.8
Avg Pool+2nd Att.	58.3	80.3	94.4
1st Att.+Avg Pool	58.6	81.0	94.3
1st Att.+2nd Att	**59.5**	**82.1**	**95.5**

Table 3. Performance of RF-ReID with and without the attention module. The first attention layer (1st Att.) denotes the layer operating within each segment (3 sec), and the second attention layer (2nd Att.) denotes the layer operating on the whole tracklet.

Environment Discriminator: RF-ReID use a discriminator to prevent the model from generating environment dependent features. We evaluate RF-ReID's performance with and without the discriminator. As shown in Table 4, adding the discriminator helps the model improve the performance by a large margin.

Figure 6 visualizes the feature space learned by RF-ReID using t-SNE [31]. Each point in the figure corresponds to a feature vector extracted from a tracklet in the test set. There are in total 5 environments in RRD-Campus test set, and we color each point according to its environment. The figure shows that without the discriminator, the feature distribution is strongly correlated with the environment. In contrast, with the discriminator, the features are successfully decoupled from the environment and more uniformly distributed. This result further demonstrates that the proposed environment discriminator can help the model learn identifying features focused on the person rather than the environment.

Method	mAP	CMC-1	CMC-5
w/o discriminator	56.7	74.2	93.3
w/ discriminator	**59.5**	**82.1**	**95.5**

Table 4. Performance of RF-ReID with and without the environment discriminator.

6.4. Qualitative Results

In Figure 5, we show examples from the test set of RRD-Campus. Each example corresponds to a query sample and its closest sample in the gallery. We compare the results generated by RF-ReID to the video-based baseline.

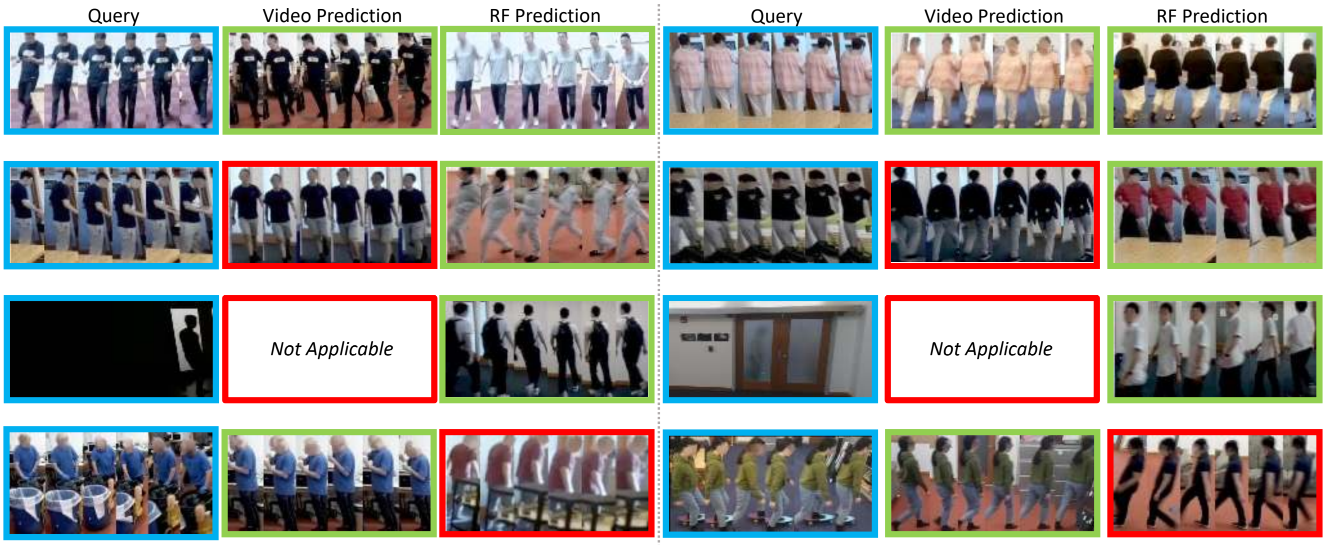

Figure 5. Qualitative results on RRD-Campus test set. Each row shows two examples, separated by the dotted line. The first column of each example is a query sample in RRD-Campus test set. The second column is the top-1 prediction by the video-based ReID model in the gallery set. The third column is the top-1 prediction by RF-ReID in the gallery set. Blue boxes stand for query sample. Green boxes mean the prediction is correct, and red boxes mean the prediction is wrong. The first row shows scenarios where both video-based ReID and RF-ReID succeed matching the correct person. The second row shows scenarios where video-based ReID fails, and matches to the wrong person because he has similar clothes, while RF-ReID provides accurate predictions. The third row shows RGB-based ReID fails under dark and occluded conditions but RF-ReID can still work. The last row shows the limitations of RF-ReID which emphasizes the walking style of the person and can get confused when he drags a bicycle or is skateboarding.

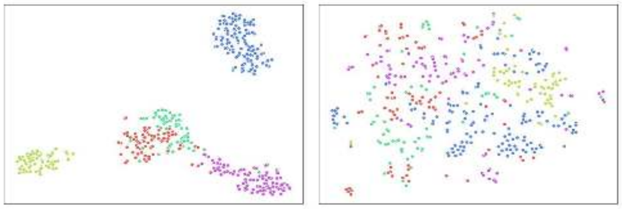

Figure 6. Distribution of features extracted by RF-ReID from different environments. The color indicates which environment a feature belongs to. The sub-figure on the left shows the feature distribution without the environment discriminator, where features from same environment are strongly clustered due to environment-dependent information. The sub-figure on the right shows the feature distribution with the environment discriminator. Here, the features are more uniformly spaced, showing that they are more environment-invariant.

The figure shows that RGB-based ReID is focused on colors and clothes, where as RF-ReID is resilient to changes in colors and clothing items. In particular, in the second row in the figure, RGB-based ReID fails because the gallery contains other people wearing clothes similar to the query sample. In contrast, RF-ReID identifies the correct person even if he/she wears completely different clothes in the gallery. This demonstrates the robustness of RF-ReID against changes in clothes.

Further, the third row in the figure shows that RGB-based ReID fails when faced with poor lighting or occlusions. In the example on the left, the light is turned off, and as a result the RGB-based ReID model fails to detect the person altogether, while the RF-based model works accurately without being affected by poor lighting. Additionally, in the example on the right, the person is behind a door so the camera can only see a vague shadow. RF-ReID can still work in this scenario because RF signals naturally traverse walls and occlusions.

We also observe that RF-based ReID can fail under some circumstances, as shown in the last row. In the example on the left, the person in the query is walking with a bicycle. The bicycle changes the person's walking style; it also disturb the RF reflections due to its metallic frame, leading to inaccurate prediction. Another failure case is when the person is on a skateboard. Since RF-ReID focuses on the person's walking style, it fails to identify this person correctly.

7. Conclusion

We have proposed RF-ReID, a novel approach for person ReID using RF signals. Our approach can extract longterm identifying features from RF signals and thus enables person re-identification across days, weeks, etc. This is in contrast to RGB-based ReID methods which tend to focus on short-lived features such as clothes, bags, hair style, etc. Also, unlike cameras, RF signals do not reveal private or personal information. Thus, our ReID method can be used in healthcare applications and clinical studies where it is important to track the health of each subject over time, while keeping the data de-identified. Finally, RF signals work in the presence of occlusions and poor lighting conditions, allowing us to ReID people in such scenarios. We believe this work paves the way for many new applications of person ReID ,where it is desirable to track people for a relatively long period and without having to collect their personal information.

References

[1] Fadel Adib, Zach Kabelac, Dina Katabi, and Robert C Miller. 3d tracking via body radio reflections. In *11th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 14)*, pages 317–329, 2014. 3

[2] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 837–846. ACM, 2015. 3

[3] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3908–3916, 2015. 1, 3

[4] Petr Beckmann and Andre Spizzichino. The scattering of electromagnetic waves from rough surfaces. *Norwood, MA, Artech House, Inc., 1987, 511 p.*, 1987. 2, 3

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[6] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018. 1, 3

[7] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016. 1

[8] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2017. 1, 3

[9] Kevin Chetty, Qingchao Chen, Matthew Ritchie, and Karl Woodbridge. A low-cost through-the-wall fmcw radar for stand-off operation and activity detection. In *Radar Sensor Technology XXI*, volume 10188, page 1018808. International Society for Optics and Photonics, 2017. 3

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 3

[11] Jiysang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018. 1, 2, 3, 6, 7

[12] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors. *Person Re-Identification*. Advances in Computer Vision and Pattern Recognition. Springer, 2014. 2

[13] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2005. 2

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 3, 4

[15] Feng Hong, Xiang Wang, Yanni Yang, Yuan Zong, Yuliang Zhang, and Zhongwen Guo. Wfid: Passive device-free human identification using wifi signal. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 47–56. ACM, 2016. 3

[16] Chen-Yu Hsu, Aayush Ahuja, Shichao Yue, Rumen Hristov, Zachary Kabelac, and Dina Katabi. Zero-effort in-home sleep and insomnia monitoring using radio signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):59, 2017. 3

[17] Chen-Yu Hsu, Rumen Hristov, Guang-He Lee, Mingmin Zhao, and Dina Katabi. Enabling identification and behavioral sensing in homes using radio reflections. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 548. ACM, 2019. 3, 5

[18] Chen-Yu Hsu, Yuchen Liu, Zachary Kabelac, Rumen Hristov, Dina Katabi, and Christine Liu. Extracting gait velocity and stride length from surrounding radio signals. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2116–2126. ACM, 2017. 3

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 7

[20] Zachary Kabelac, Christopher G Tarolli, Christopher Snyder, Blake Feldman, Alistair Glidden, Chen-Yu Hsu, Rumen Hristov, E Ray Dorsey, and Dina Katabi. Passive monitoring at home: A pilot study in parkinson disease. *Digital Biomarkers*, 3(1):22–30, 2019. 5

[21] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. 2008. 2

[22] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. Spotfi: Decimeter level localization using wifi. In *ACM SIGCOMM computer communication review*, volume 45, pages 269–282. ACM, 2015. 3

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[24] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 872–881, 2019. 2, 3, 6

[25] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 1

[26] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 2

[27] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and

Ivan Poupyrev. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)*, 35(4):142, 2016. 3

[28] Giuseppe Lisanti, Iacopo Masi, Andrew D Bagdanov, and Alberto Del Bimbo. Person re-identification by iterative reweighted sparse ranking. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1629–1642, 2014. 2

[29] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5790–5799, 2017. 3

[30] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 6, 7

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7

[32] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person reidentification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2016. 3

[33] Athira Nambiar, Alexandre Bernardino, and Jacinto C Nascimento. Gait-based person re-identification: A survey. *ACM Computing Surveys (CSUR)*, 52(2):33, 2019. 2

[34] Zhengyu Peng, José-María Muñoz-Ferreras, Roberto Gómez-García, and Changzhi Li. Fmcw radar fall detection based on isar processing utilizing the properties of rcs, range, and doppler. In *2016 IEEE MTT-S International Microwave Symposium (IMS)*, pages 1–3. IEEE, 2016. 3

[35] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*, pages 27–38. ACM, 2013. 3

[36] Tauhidur Rahman, Alexander T Adams, Ruth Vinisha Ravichandran, Mi Zhang, Shwetak N Patel, Julie A Kientz, and Tanzeem Choudhury. Dopplesleep: A contactless unobtrusive sleep sensing system using short-range doppler radar. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 39–50. ACM, 2015. 3

[37] Peter M Roth, Martin Hirzer, Martin Köstinger, Csaba Beleznai, and Horst Bischof. Mahalanobis distance learning for person re-identification. In *Person re-identification*, pages 247–267. Springer, 2014. 2

[38] Cong Shi, Jian Liu, Hongbo Liu, and Yingying Chen. Smart user authentication through actuation of daily activities leveraging wifi-enabled iot. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, page 5. ACM, 2017. 3

[39] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 1

[40] Yonglong Tian, Guang-He Lee, Hao He, Chen-Yu Hsu, and Dina Katabi. Rf-based fall monitoring using convolutional neural networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):137, 2018. 3

[41] Deepak Vasisht, Swarun Kumar, and Dina Katabi. Decimeter-level localization with a single wifi access point. In *13th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 16)*, pages 165–178, 2016. 3

[42] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29, 2013. 2

[43] Wei Wang, Alex X Liu, and Muhammad Shahzad. Gait recognition using wifi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 363–373. ACM, 2016. 3

[44] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st annual international conference on mobile computing and networking*, pages 65–76. ACM, 2015. 3

[45] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures. In *Proceedings of the 20th annual international conference on Mobile computing and networking*, pages 617–628. ACM, 2014. 3

[46] Jie Xiong and Kyle Jamieson. Arraytrack: A fine-grained indoor location system. In *Presented as part of the 10th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 13)*, pages 71–84, 2013. 3

[47] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 4733–4742, 2017. 3

[48] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer, 2016. 3

[49] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6):2860–2871, 2019. 1

[50] Yunze Zeng, Parth H Pathak, and Prasant Mohapatra. Wi-who: wifi-based person identification in smart spaces. In *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*, page 4. IEEE Press, 2016. 3

[51] Zhenyuan Zhang, Zengshan Tian, and Mu Zhou. Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor. *IEEE Sensors Journal*, 18(8):3278–3289, 2018. 3

[52] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person reidentification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3219–3228, 2017. 1

[53] Mingmin Zhao, Fadel Adib, and Dina Katabi. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 95–108. ACM, 2016. 3

[54] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018. 2, 6

[55] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human mesh recovery using radio signals. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10113–10122, 2019. 2, 3, 7

[56] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. Rf-based 3d skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 267–281. ACM, 2018. 2, 3, 4, 5

[57] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4100–4109. JMLR. org, 2017. 3

[58] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 6

[59] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 6

[60] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018. 1