RF-CM: Cross-Modal Framework for RF-enabled Few-Shot Human Activity Recognition

XUAN WANG, Northwest University, China
TONG LIU, Northwest University, China
CHAO FENG, Northwest University, China
DINGYI FANG, Northwest University, Shaanxi International Joint Research Centre for the Battery-Free Internet of Things, China
XIAOJIANG CHEN*, Northwest University, China

Radio-Frequency (RF) based human activity recognition (HAR) enables many attractive applications such as smart home, health monitoring, and virtual reality (VR). Among multiple RF sensors, mmWave radar is emerging as a new trend due to its fine-grained sensing capability. However, laborious data collection and labeling processes are required when employing a radar-based sensing system in a new environment. To this end, we propose RF-CM, a general cross-modal human activity recognition framework. The key enabler is to leverage the knowledge learned from a massive WiFi dataset to build a radar-based HAR system with limited radar samples. It can significantly reduce the overhead of training data collection. In addition, RF-CM can work well regardless of the deployment setups of WiFi and mmWave radar, such as performing environments, users' characteristics, and device deployment. RF-CM achieves this by first capturing the activity-related variation patterns through data processing schemes. It then employs a convolution neural network-based feature extraction module to extract the high-dimensional features to be fed into the activity recognition module. Finally, RF-CM takes the generalization knowledge from WiFi networks as guide labels to supervise the training of the radar model, thus enabling a few-shot radar-based HAR system. We evaluate RF-CM by applying it to two HAR applications, fine-grained American sign language recognition (WiFi-cross-radar) and coarse-grained gesture recognition (WiFi-cross-RFID). The accuracy improvement of over 10% in both applications demonstrates the effectiveness of RF-CM. This cross-modal ability allows RF-CM to support more cross-modal applications.

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

Additional Key Words and Phrases: Cross-modal, Human Activity Recognition, Knowledge Transfer

ACM Reference Format:
Xuan Wang, Tong Liu, Chao Feng, Dingyi Fang, and Xiaojiang Chen. 2023. RF-CM: Cross-Modal Framework for RF-enabled Few-Shot Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 35 (March 2023), 28 pages. https://doi.org/10.1145/3580859

*Corresponding author

Authors' addresses: Xuan Wang, Northwest University, China, xwang@stumail.nwu.edu.cn; Tong Liu, Northwest University, China, liutong98@stumail.nwu.edu.cn; Chao Feng, Northwest University, China, chaofeng@nwu.edu.cn; Dingyi Fang, Northwest University, Shaanxi International Joint Research Centre for the Battery-Free Internet of Things, China, dyf@nwu.edu.cn; Xiaojiang Chen, Northwest University, China, xjchen@nwu.edu.cn.

2474-9567/2023/3-ART35 $15.00
https://doi.org/10.1145/3580859

1 INTRODUCTION

Human activity recognition built upon Radio-Frequency (RF) sensors (e.g., RFID, WiFi, and millimeter wave (mmWave) radar) promises many exciting application scenarios, such as gesture recognition [7, 10, 20, 22, 27, 43, 49], pose tracking [15, 32, 33], and vital sign detection [6, 11, 26, 44, 50]. RF-based recognition systems that provide non-intrusive [14] and continuous sensing over the air have attracted widespread attention in both academia and industry.

Machine learning is a viable means to construct recognition models for RF-enabled HAR systems [8, 21, 29, 43]. Despite promising, building machine-learning-based recognition models in real life is accompanied by time-consuming and labor-exhausting processes [55]. For example, CrossSense [53] took 277 days to collect 1.2 million WiFi data for learning a WiFi-based decision model, even when each data only took 20 seconds. In addition, as we build recognition models for different RF sensors, the collection process will multiply significantly. The conspicuous reason is that the recognition models of different sensors are isolated and independent from each other. For instance, a recognition model trained on WiFi data can not be directly used to predict the label of radar data. This is because different sensors have different operating frequency bands, bandwidths, and modulation schemes, resulting in various data distributions between different sensors. When multiple RF sensors are deployed on a large scale, the aforementioned drawback leads to an expensive and repetitive data collection process. While collecting data from different sensors in a home may be feasible (only a few people, e.g., 3-5 users), requiring each employee or visitor in a smart company to provide training samples for various sensors is impractical.

Recent research has attempted to reduce the cost of training data collection by utilizing video to generate RF signals [1, 4]. Although these approaches work well for identifying simple whole-body activities (such as gym sports), their performance will degrade for complex and fine-grained gesture recognition. The reasons are as follows: 1) The accuracy of generated or transferred "pseudo" RF signals are unguaranteed since the video does not contain complex and rich channel information (i.e., environmental information and dynamic variation caused by human activities), but "real" RF signals collected from the real world do. 2) The diversity of generated or transferred "pseudo" RF signals is limited, which can not cover the data distribution collected from the real world, but "real" RF signals collected from the real world do.

In this paper, we ask the following question: Can we leverage the existing massive wireless signal data to establish a new recognition model for other sensors with a small number of training samples? We propose an affirmative answer through RF-CM, a general cross-modal human activity recognition framework. We investigate the viability of cross-modal fine-grained HAR by taking WiFi and radar as examples. Since prior works in WiFi-based HAR have extensively collected WiFi data for extensive activities and made it publicly available [51, 55], WiFi thus is a good candidate as source data. Furthermore, as a dedicated RF sensor, the emerging millimeter wave radar has larger bandwidth and higher resolution than WiFi. There is excellent potential for future deployment in families to identify the fine-grained subtle human activities (e.g., hand gesture activities). By leveraging the excellent knowledge from public WiFi datasets, RF-CM achieves a few-shot radar-based fine-grained HAR system. The cross-modal capability of RF-CM significantly reduces the cost of radar data collection. As illustrated in Fig. 1, RF-CM can work well regardless of the deployment setups of WiFi and mmWave radar, such as performing environments, users' characteristics, and device deployment. RF-CM opens up new possibilities for applications across other sensors and provides a new view for deploying other new sensors rapidly in the future.

To realize our idea into practice, multiple challenges need to be addressed:

Establishing a high-accuracy radar sensing model with a small number of radar training samples is a challenge in our system. Specifically, a few training samples will over-fit [48] the radar model since the feature space distributions extracted from insufficient samples have a significant deviation. We attempt to break through the over-fitting bottleneck by transferring the knowledge learned from massive WiFi datasets to the radar model. However, WiFi and radar have different frequency bands, bandwidths, and modulation schemes, leading to the

(a) WiFi (b) Radar

Fig. 1. Deployment settings of the WiFi and the mmWave radar.

various dimensions and distribution of WiFi data and radar data. Therefore, a fine-tuning strategy of transfer learning [34] is ineffective since the fine-tuning radar samples and the WiFi data have the same dimension. One potential way is to obtain the transformation matrix to scale the radar and WiFi data dimensions to ensure consistency. However, searching for an effective transformation matrix is difficult as wireless transmission channels are complicated.

Fortunately, implied information between WiFi and radar data provides a new angle to tackle this issue. Specifically, although the differences in the dimensions and distributions of the data received by radar and WiFi, the dynamic trajectory of the same activity is the same, and the relationship between different activities is relatively stable. In other words, the classification probability distributions of WiFi and radar data are similar. Therefore, by minimizing the difference in classification probability distributions between WiFi data and radar data, we can transfer the knowledge from WiFi to radar. Our intuition is that the higher similarity between two classification probability distributions indicates more knowledge radar learns from WiFi and the more accurate the model's prediction will be. To further improve the radar model's generalization performance, we tune WiFi data's classification probability distributions to include more abundant knowledge, such as the relative relationship between categories. When transferring knowledge to the radar model, the variety of knowledge is equivalent to increasing the diversity of radar training set samples. The strategies can significantly decrease the number of samples necessary for radar model training while improving generalization performance.

The second challenge is how to distill activity-related features and dispel activity-agonistic features from different RF signals. RF signals usually carry redundant domain information irrelevant to human activities, such as the surrounding environment, the user's position, and the like. Variations in domain factors may distort wireless signal patterns even with the same activity. In addition, radar data and WiFi data may have been collected under completely different domain factors (i.e., different experimental settings, the environment, the user's characteristics, and the user's location, as shown in Fig. 1). It makes extracting the relationships between two domains' signal patterns more challenging, especially when training samples are insufficient.

We overcome this challenge by designing data processing methods and feature extraction networks for different RF signals. Specifically, interference elimination methods and feature pre-extraction to acquire the activity-related features. These methods effectively suppress reflected signals of dynamic and static objects in the environment, such as tables and walls, people walking around, and even the breathing and heartbeat of the user. Then, by designing a deep-learning-based feature extraction network, we can dispel activity-irrelevant features from different RF signals, thus obtaining activity-related features.

The main contributions of RF-CM can be summarized as follows.

- We propose a general cross-modal framework RF-CM. It enables a few-shot radar-based human activity recognition model by using a large number of public WiFi data. RF-CM can achieve good performance when the deployments and environments of radar and WiFi are different.

Table 1. The public dataset summary for gesture recognition studies

Signal	Paper	Users	Ges	Envir	Pos	Dir	Times /Ges	Total
WiFi	Widar3.0 [55]	16	6	3	5	5	5	12000
	RISE [51]	6	6	2	5	5	375	11700
	Wiar [12]	10	16	3	1	1	30	14400
	SignFi [29]	5	276	2	1	1	10	27600
RFID	RISE [51]	6	6	1	5	5	30	900
mmWave	mHomeGes [24]	25	10	3	12	9	30	22000
	mTransSee [22]	32	5	5	13	face	16	54080

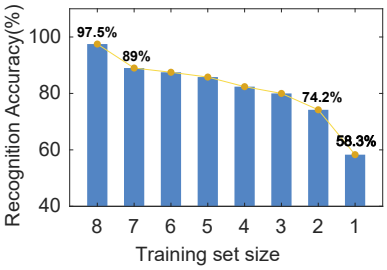


Fig. 2. Recognition accuracy using different number of samples for training.

- We design a cross-modal strategy to effectively teach the radar model to maximize the information extracted from limited training samples, significantly minimizing data collection cost.
- We evaluate the performance of RF-CM using two cross-modal applications, i.e., WiFi cross radar and WiFi cross RFID. Extensive experiments demonstrate the effectiveness of RF-CM. While, the proposed method can also benefit other wireless sensing modalities.

2 BACKGROUND

In this section, we first define the problem we aim to solve. Then we explain our motivation for tackling the issue. Finally, we describe the data formats received by WiFi and mmWave radar.

2.1 Problem Scope

RF-CM mainly concentrates on cross-modality sensing scenarios. We aim to build a few-shot human activity recognition system for mmWave radar with the help of massive WiFi datasets. We here provide a general description of the professional terms we used as follows: 1) *domain:* a specific deployment setup including domain factors like sensing modalities, users, device setup, and deployment environment. A new domain emerges when domain factors change. 2) *cross-modality:* It means that the RF sensors used for sensing in the source domain and the target domain are different. 3) *in-domain:* domain factors in the training and testing sets are the same. 4) *shot:* It refers to the number of labeled samples per gesture and per person. Although other domain factors of locally collected radar data and public WiFi data differ, our work does not target finding better methods to eliminate the influence of domain factors. Many efforts have been made to tackle this problem [20, 55].

2.2 Motivation

The analysis of the following three issues motivates us to explore whether we can utilize the public WiFi dataset to assist in training the radar-based human activity recognition model.

2.2.1 Practical Cross-modal RF Sensing Issue. Radio-Frequency (RF) sensing modalities such as RFID, WiFi, and millimeter wave (mmWave) radar have been widely employed for human activity recognition (HAR) [17, 28]. Among these RF signals, the ability to capture micro-motion dynamics of subtle activities (e.g., hand activities such as brushing, eating, etc.) makes millimeter wave (mmWave) radar a promising sensing approach for human activity recognition. It offers many fine-grained sensing applications. Despite the satisfactory results, radar has no existing large datasets, holding back this otherwise promising sensing modality. To develop a HAR model for radar, a labor-intensive and time-consuming data recollection process is required. Since the existing machine-learned solutions are designed for specific RF modalities. To be more intuitive, we imagine an example. As illustrated in Fig. 1(a), we have already collected rich human activity data by deploying WiFi Access Point (AP) and client in living room S2. When we want to deploy a radar in bedroom S3 for other family members, as in Fig. 1(b), we need

to re-collect a large amount of data in the bedroom to train a sensing model for radar. The HAR model trained on WiFi data in S2 can not support predicting the radar data since the data of WiFi and radar have different dimensions and distributions.

2.2.2 Small Sample Issue. Inadequate training data may cause the trained model to perform well on training samples but poorly generalized on the test dataset. To more intuitively understand the consequences of a small sample set, we conducted an experiment. We invited a participant to perform 6 gestures in front of a mmWave radar, with 10 samples for each gesture. We alter the ratio of training to testing samples. The recognition results in Fig.2 indicate that the accuracy drops dramatically with fewer training samples due to poor generalization performance.

2.2.3 Abundant Publicly Available RF Sensing Datasets. We investigated the literature on deep learning-based contactless sensing using WiFi, RFID, mmWave radar, etc. We found that each study conducted extensive experiments and collected a massive amount of data to confirm the proposed methods' viability. We summarize some public datasets of different modalities in Table 1. Most of these researchers collected tens of thousands of samples and released their datasets on GitHub. HAR systems with WiFi, in particular, provide more activity categories and samples. We appreciate the importance of these efforts, mainly the work done to make the dataset publically available, which has opened an opportunity for academics to explore HAR in the future.

2.3 Primer

2.3.1 WiFi Channel State Information (CSI).. The WiFi 802.11n protocol encodes digital data using OFDM on 64 subcarriers, 52 of which are used to transmit data. CSI depicts the combined impact of, for instance, scattering, fading, and power degradation with distance as different frequency subcarriers travel from the transmitter to the receiver. CSI measurements H represent the Channel Frequency Response (CFR) and are obtained by using the received and transmitted known long training sequence (LTS) in the preamble of WiFi packets, which can be represented as,

$$H = \frac{Y_{LTS}}{X_{LTS}} + Noise \tag{1}$$

where, X_{LTS} and Y_{LTS} are the frequency domain representations of transmitted and received packets, respectively. The measured CSI of each packet is a $N \times M \times P$ three-dimensional matrix of complex values. N and M refer to the number of transmitting and receiving antennas, respectively, and P denotes the number of subcarriers. Abundant physical layer CSI reflects each subcarrier's amplitude and phase variation over the transmission paths, which can be used for human activity recognition.

2.3.2 mmWave Radar. We take the most commonly used FMCW radar as an example for analysis. FMCW radar sends and receives Frequency Modulated Continuous Wave (FMCW) with periodic linear frequency changes. The transmitted signal is reflected by the target and returns to the receiver with a certain delay. In general, the received signal is multiplied by the conjugate of the sent signal to produce an intermediate frequency signal that is easy to process. We represent the IF signal as,

$$S_{IF}(t) = Ae^{-j2\pi\left\{f_0\tau(t)+\mu t\tau(t)-\frac{1}{2}\mu^2\tau^2(t)\right\}}. \tag{2}$$

where $\tau(t) = \frac{2R_0+Vt}{c}$, R_0 is the distance from target to radar. μ is the slope of the FMCW signal. IF data sampled by ADC can be organized to a $M \times N \times K \times Q$ 4D matrix, where M, N, K, Q are the number of chirps per frame, sampling points per chirp, receiving antennas, and frames collected at a time, respectively. To better show the data processing, we take one frame of data (a $M \times N \times K$ cube) as an example and visualize the processing in Fig. 3.

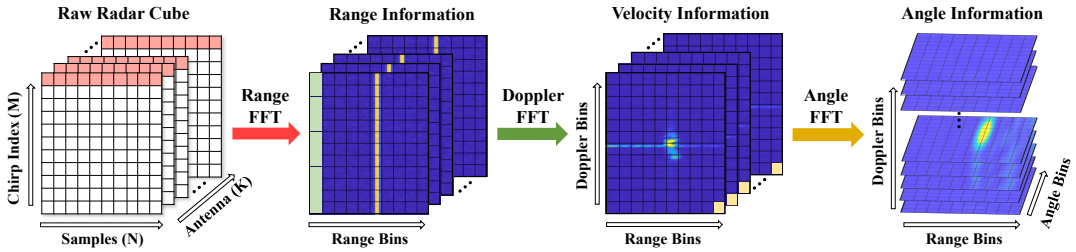


Fig. 3. Processing of one frame of mmWave radar data.

i) *Distance estimation:* When the reflecting object is static, the sampled IF signal has a constant frequency $f = \frac{S \times 2 \times R}{c}$. Therefore, by doing a fast Fourier Transform (FFT) on each chirp of the raw data cube, we can estimate the range R according to the peak of the frequency spectrum. We call this process Range-FFT and denote each FFT bin as a range bin.

ii) *Velocity estimation:* The movement of a human relative to radar will produce Doppler frequency shifts, which results in each chirp having a different phase for the same range bin. Therefore, we can calculate the velocity as $v = \frac{\lambda \omega}{4\pi T_c}$ based on the phase difference ω of the same range bin across two consecutive chirps, where T_c is the interval of two chirps. The specific operation is to perform another FFT (Doppler-FFT) on each range bin of the Range-FFT cube to generate a 2D Range-Doppler matrix for each antenna.

iii) *Angle estimation:* Estimating angle requires at least two receive antennas. The phase difference of two antennas can be exploited to estimate the angle of arrival (AoA), which can be expressed as $\theta = arc \sin\left(\frac{\lambda \Delta\phi}{2\pi d}\right)$. Where d and $\Delta\phi$ are the distance and phase difference between two antennas. It can be obtained by doing another FFT (Angle-FFT) on the Range-Doppler matrix along the antenna dimension. For each Doppler bin, we can obtain a 2D Range-Angle matrix.

3 SYSTEM OVERVIEW

RF-CM is a general cross-modal framework for radar-enabled few-shot human activity recognition. The core idea of RF-CM is to leverage the knowledge from massive WiFi data to teach the radar model to achieve good generalization performance with a few radar training samples. Fig.4 provides an overview of RF-CM, which involves the following modules.

Data Collection. We select 38 American Sign Language (ASL) words from the datasets published by SignFi [29] as our WiFi dataset. We collect radar samples for the 38 words by using a commercial radar sensor (Texas Instruments (TI) IWR1843 [37]) in the real world.

Data Processing and Feature Pre-extraction Artificially. Collected wireless signal data typically contain reflections and noise from other objects (static and dynamic) in the surroundings. Therefore, to remove interferences from raw RF data, we apply a set of signal processing methods for WiFi and radar. Then, we deliberately extract features that can represent the variation patterns induced by human activity according to the properties of each signal. For different RF sensors, the preprocessing and dynamic feature extraction may be different. We describe how they are handled separately in Sec. 5).

Knowledge Transfer Based HAR for mmWave Radar. RF-CM framework consists of two pipelines: a pre-trained WiFi-Net for assistance based on massive high-quality WiFi samples and a new radar-based model trained with a few radar samples. Each pipeline has a Convolutional Neural Network (CNN) based feature extraction network to extract deep representations of activities and a classification module to convert them into probability

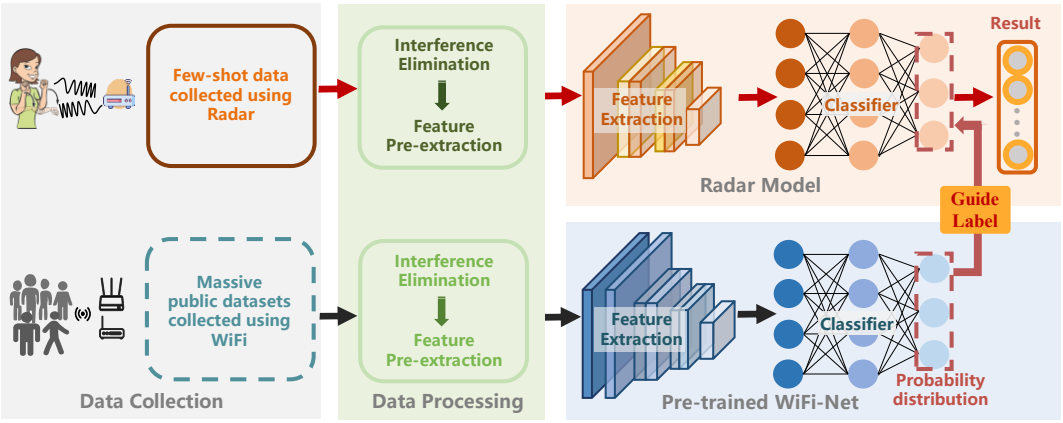


Fig. 4. Overview of RF-CM.

distribution of classes. We separately feed the WiFi and radar samples of the same category into two networks. And we use the soft probability distributions generated by WiFi-Net as 'guide labels' to 'teach' the training of the radar model. By doing so, the radar model's generalization performance can be increased by learning valuable knowledge from WiFi samples.

4 MODELING RF SENSING

In this section, we build theoretical reflection models for WiFi and mmWave to explore the relationship between these wireless signals and human activities. It is the basis for our knowledge transfer based cross-modal HAR.

In an indoor environment, transmitted signals usually bounce off various objects, and these reflected copies are intertwined at the receiver. The received signal is the superposition of signals from multiple paths, which can be expressed as,

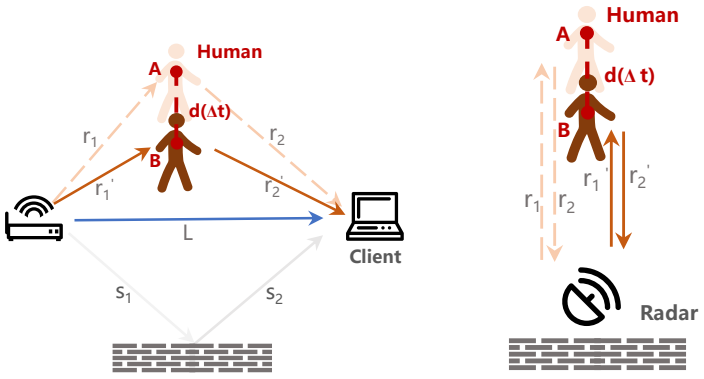
$$s(t) = \sum_{i=1}^{N} a_i p\left(t - \tau_i\right) e^{j\phi} \tag{3}$$

where $p(t)$ is the signal from the transmitter. a_i and τ_i stand for the magnitude and the time of flight (ToF) of the i arriving path, respectively. ϕ is a random phase shift caused by hardware imperfection, which is not a time-varying value. We can get the channel frequency response (CFR) $H(t)$ through $s(t)$ dividing by $p(t)$ in frequency domain. It is the sum of all the static paths' CFR $H_s(t)$ and the dynamic paths' CFR $H_d(t)$.

Fig. 5 shows the general sensing deployments of the WiFi and mmWave radar. In order to make the movement of the human relative to the transceiver consistent with WiFi, we assume that radar is deployed in front of the human. The environment and multipath conditions of the two deployments are the same. At time t_0 the human is in position A, and at time t_1 human moves to B. The displacement of $\Delta t = t1 - t0$ is $d(\Delta t) = d(t_0) - d(t_1)$. The change of signal transmission path length $R(\Delta t) = (r_1 + r_2) - (r_1' + r_2')$ caused by human's movement $d(\Delta t)$. Assuming that there are no other dynamic non-human objects in environments. Therefore, the $H_d(t)$ mainly depends on the dynamic changes caused by human behavior. Since the equipment and deployments for different RF sensors are different, we next analyze their channels separately.

WiFi. As shown in Fig. 5 (a), for one pair of WiFi devices (an Access Point (AP) and a Client), the channel frequency response contains three parts, which can be expressed as,

$$H(t) = H_s(t) + H_d(t) = H_{dp}(t) + H_{sr}(t) + H_{dr}(t) \tag{4}$$



(a) WiFi (b) Radar

Fig. 5. General human activity sensing models of WiFi and mmWave radar.

where $H_{dp}(t)$ is the direct path (AP → Client), $H_{sr}(t)$ is the static reflection path (AP → Walls → Client), and $H_{dr}(t)$ denotes the dynamic reflection path (AP → Human → Client). The CFR at t_0 and t_1 can be expressed as, respectively,

$$H(t_0) = a_0 e^{-j2\pi f \frac{L+(s_1+s_2)+(r_1+r_2)}{c}}, H(t_1) = a_1 e^{-j2\pi f \frac{L+(s_1+s_2)+(r_1'+r_2')}{c}} \tag{5}$$

where f is the frequency of carriers. By making the difference between the channels of two consecutive packets t_0 and t_1, we eliminate the static components and get the dynamic channel variation as, $H(\Delta t) = e^{-j2\pi f \frac{R(\Delta t)}{c}}$

The phase difference can be represented as,

$$\Delta\phi = 2\pi f \frac{R(\Delta t)}{c} \tag{6}$$

mmWave Radar. mmWave radar transmits frequency modulated continuous wave (FMCW) and receives the reflected signals from surroundings. As shown in Fig. 5 (b), the radar will not receive the static reflection path (Radar→ Walls → Radar) since radar usually uses directional antennas to reduce transmission loss and the wall is behind the radar. Therefore, the received signal only has a dynamic path (Radar→ Human → Radar). The received IF signal at times t_0 and t_1 can be simplified as,

$$S_{IF0} = a_0 e^{-j2\pi f_{IF0}(t_0)+\phi(t_0)}, S_{IF1} = a_1 e^{-j2\pi f_{IF1}(t_1)+\phi(t_1)} \tag{7}$$

where $f_{IF0} = \frac{\mu(r_1+r_2)}{c}$ and $f_{IF1} = \frac{\mu(r_1'+r_2')}{c}$, μ is the slope of FMCW. The frequency difference of the IF signal generated by two consecutive frames is,

$$\Delta f = \frac{\mu(r_1 + r_2 - r_1' + r_2')}{c} = \frac{\mu R(\Delta t)}{c} = \frac{\mu \Delta d}{c} \tag{8}$$

Although the signals of the two sensors have different characteristics (e.g., frequencies, frequency bandwidths, and modulation schemes), the $\Delta\phi$ of WiFi (Eq. 6)) and the Δf of mmWave (Eq. 8) are all originate from the human displacement $d(\Delta t)$. In spite of differences in data dimensions and feature distributions, there is a correlation between their signal patterns of the same activity.

5 SYSTEM DESIGN

In this section, we focus on how to realize radar-based human activity recognition with minimal training samples. The lack of training samples prevents the neural network from obtaining the correct representation, which leads to over-fitting. To solve it, our insight is to use the knowledge learned from public WiFi datasets to guide the training of radar model. We first describe the problem formulation. Then we elaborate the system design of RF-CM.

5.1 Problem Formulation

To be concrete, let $D_S = \left\{\left(\mathrm{x}_i^S, y_i^S\right)\right\}_{i=1}^{n_S}$ denote a WiFi dataset containing a large number of training samples, where $x_i^S \in \mathbb{R}^{d_S}$ is the i-th sample with d_S-dimensional feature and $y_i^S \in \{1, \ldots, K\}$ representing the corresponding class label,n_S refers to the size of WiFi dataset. Let $D_T = D_l \cup D_u = \left\{\left(\mathrm{x}_i^l, y_i^l\right)\right\}_{i=1}^{n_l} \cup \left\{\left(\mathrm{x}_i^u\right)\right\}_{i=1}^{n_u} = \left\{\left(\mathrm{x}_i^T, y_i^T\right)\right\}_{i=1}^{n_T}$ denote a radar dataset containing limited training samples D_l and many test samples D_u, where $x_i^T \in \mathbb{R}^{d_T}$ is the i-th sample with d_T-dimensional feature and $y_i^T \in \{1, 2, \ldots K\}$. $n_T = n_l + n_u$ are the number of labeled and unlabeled radar samples. In our problem, D_T and D_S are datasets collected from different sensors, environments, participants, $d_S \neq d_T, n_S \gg n_T$, and $n_u \gg n_l$. We aim to use D_l to build a general model for predicting the results of unlabeled radar data D_u. Learning the knowledge learned from a large WiFi dataset D_S improves the classification accuracy and generalization of the radar model.

Many domain adaptation methods [20, 53] cannot be used since they usually assume that the source and target domain data are of the same dimension. Some methods [9] attempt to share the same network by transforming the feature dimensions of the two domains into a common dimension d or mapping the feature dimension of the target domain into the feature dimension of the source domain. The transformation functions are $m_s : \mathbb{R}^{d_S} \rightarrow \mathbb{R}^d$ and $m_t : \mathbb{R}^{d_T} \rightarrow \mathbb{R}^d$, or $m_{st} : \mathbb{R}^{d_T} \rightarrow \mathbb{R}^{d_S}$. However, these methods all enforce an alignment between domains in a brute-force manner, which may damage the discriminative information of both source and target domain data due to the vast difference across domains. Different from them, we use the probability distributions generated by WiFi samples to supervise the training of radar model. The details of each component will be described in the following sections.

5.2 Signal Processing and Dynamic Feature Pre-extraction

A key challenge of contactless human activities recognition is that the received signal comprises information about both the target's activities and the surrounding environment (such as the reflected signal by the door, wall, table, and other noise signals in the air). As a result, the change in spatial position between the target and transceiver (distance and angle) can significantly alter the reflected signals even with the same activity. In real applications, however, requiring users to complete activities in specific areas is impractical. Therefore, this section aims to extract the activities-related features and suppress activities-irrelevant interference.

5.2.1 Feature Selection. As described in Sec. 2, we can artificially pre-extract the dynamic features caused by activities, such as range, velocity, and angle. Generally, the distance variation of sign language gestures is within 65 cm, the velocity is usually less than 3 m/s, and the horizontal angle variation is about 10°. According to the detailed theoretical analysis of these three patterns provided in [31]), we select the most appropriate feature of the variation patterns for activities. Specifically, the range resolution R_{res} depends on the bandwidth swept by the chirp, which can be expressed as $R_{res} = \frac{c}{2B}$. The velocity resolution of the radar is inversely proportional to the duration of a frame and is given by $\mathbf{v}_{\text{res}} = \frac{\lambda}{2\mathbf{T_f}}$. The angle resolution is $\theta_{\text{res}} = \frac{\lambda}{\mathrm{Nd}\cos(\theta)}$, which is related to the number of antennas N and antenna separation d. For the parameter settings of TI-IWR 1843, the max theoretical resolution of the range and angle are 3.75 cm and 15°, respectively. If we set the duration of a frame to 40ms, the

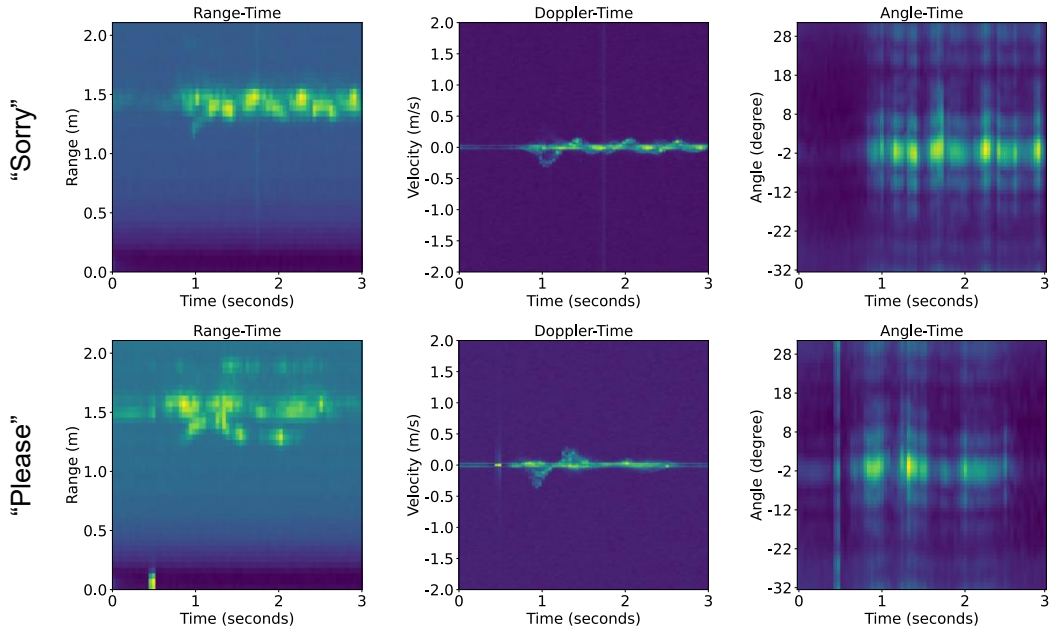


Fig. 6. Range, Velocity and Angle features of Gesture "Sorry" and "Please".

velocity resolution can achieve 0.0487 m/s. We visualize the range, velocity, and angle features of two different activities in Fig. 6. Comparing the three features, we choose the velocity feature in our system since its high resolution can capture the subtle dynamic variation better, even the finger movements. Its superiority has been demonstrated in the [41].

5.2.2 Feature Pre-extraction and Enhancement. We regard that the target's position and hands do not change in a chirp period since the duration of a chirp is usually very short, with a maximum of only tens of microseconds. Therefore, we can obtain a $M \times N \times Q \times K$ 4D matrix by performing the 2D FFT (namely Range-FFT and Doppler-FFT, described in Sec. 2) on the sampled raw IF data cube for each receiver antenna and frame. To obtain the Velocity-Time (VT) feature caused by human activities and enhance signal strength further, we perform incoherent accumulation of the Range-Doppler matrix along multiple receive antennas and range bins. The final Velocity-Time feature of one activity is a 2D matrix of $M \times Q$. This incoherent accumulation makes the power of useful signals stronger and the noise weaker since the noise is random. In addition, inspired by the spectrum features extraction in speech recognition, we perform the logarithmic transformation on the magnitude of VT to obtain XVelocity. This nonlinear transformation enhances the low-intensity but useful dynamic components.

5.2.3 Interference Signal Removal. We take mmWave signals reflected by static objects near the target as static interference, whose velocities are 0 and remain constant over successive frames. To remove these static interferences, we extract the differential Velocity-Time (DVT) feature by taking the difference of the Doppler dimension of two consecutive frames. In addition, we define the reflected signals caused by the breathing and heartbeat of the target as well as the non-target walking around as dynamic interferences, which have various frequencies since their velocity relative to the transceiver is different. Specifically, the frequencies of the reflected signal from breathing and heartbeat and people walking are typically 0.1-0.5 Hz, 0.8-2 Hz, and 1-10 Hz, respectively. Empirically, a person usually performs an ASL gesture at a speed of less than 3m/s. The corresponding frequency

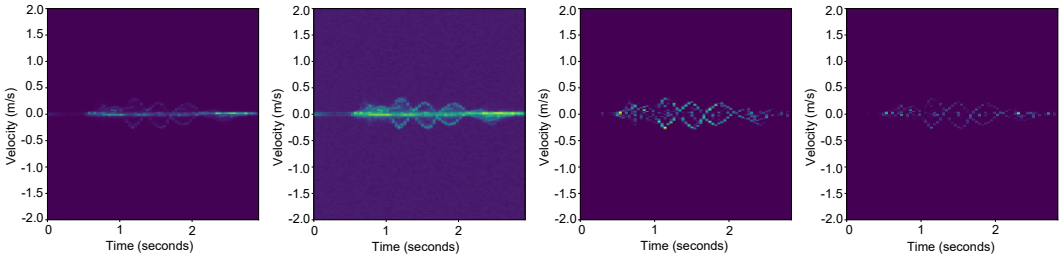


Fig. 7. The impact of signal processing on Velocity-Time feature. (a) Raw Velocity-Time with DC removal, (b) Velocity-Time with log normalization, (c) Velocity-Time with both log normalization and differential interference cancellation, (d) Velocity-Time with Differential interference cancellation but without log normalization.

is 1500 Hz ($3m/s \times 2/\lambda$). As a result, we apply a 10 Hz - 1.5 kHz band-pass filter to filter out these low-frequency interference signals and high-frequency noise.

Algorithm 1 describes the detailed data processing and feature pre-extraction. To illustrate the effect of signal processing, we visualize the velocity feature before and after performing our processing strategy. The result is shown in Fig. 7. Fig. 7(a) and 7(b) indicate the log transformation is helpful to signal enhancement. From Fig. 7(b) to 7(c), we can see that our interference cancellation method can remove the interfering signals, including the target's body with a large reflection area.

ALGORITHM 1: mmWave radar data processing

Input: Radar cube $xradar = M \times N \times Q \times K$ (M:chirps; N:samples in one chirp; Q:frames; K:antennas)
Output: pre-processed differential Velocity-Time x_{DVT}

for $t = 1$ *to* Q **do**
  for $i = 1$ *to* K **do**
    for $j = 1$ *to* M **do**
      Conduct Range-FFT with Hanning window on N samples within one chirp → XRange;
    end
    for $n = 1$ *to* N **do**
      Remove DC → XRange (n) = XRange (n)-mean(XRange (n));
      Conduct Doppler-FFT with Hanning window on M chirps of XRange (n) → Range-Doppler (RD);
    end
    Perform IIR Bandpass [10 Hz,1500 Hz] Filter on RD;
  end
  Incoherent accumulate the power of the RD → VT (t) = $\sum_i^K \sum_n^N |$ RD $|$;
  Feature enhancement → XVelocity (t) = log_{10} (VT (t));
  Extract the differential feature → DVT (t) = XVelocity (t+1) - XVelocity (t);
end
x_{DVT} = Normalization (DVT).

.

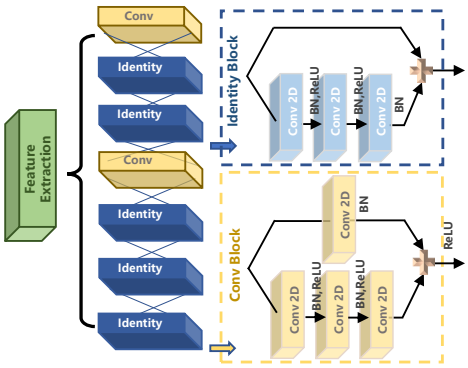


Fig. 8. Feature Extraction Network FE-Net.

5.3 Base Model

The base model contains two modules, a deep learning-based pattern extraction module and a classification module.

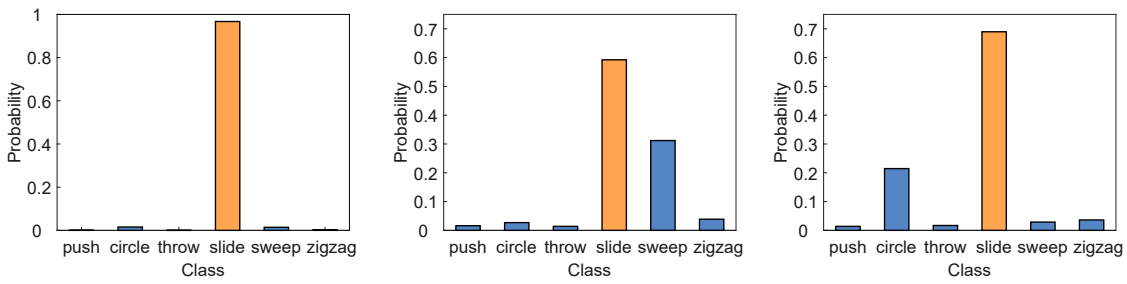
5.3.1 Deep-learning Based Feature Extraction. To learn the high-dimensional features of activities and reduce gesture-irrelevant information, we design a residual network (ResNet) based feature extraction module (FE-Net). As shown in Fig. 8, FE-Net is a stack of two types of residual blocks, Identity Block and Conv Block. Each block adopts a residual architecture. The main path of each block contains 3 convolutional layers (Conv 2D). To accelerate network training, batch normalization (BN) and ReLU activation layers are added between Conv 2D layers. In Identity Block, the output is the superimposition of the input x and the output of the main path. This block not only can learn the new features of the input but also the depth of the block is ignored, which ensures that the network's performance is not degraded. In Conv Block, the additional path is a Conv 2D layer and a BN layer. Then add it to the main path and rectify it through the ReLU activation layer to get the output of the Conv Block. We define the output of the radar's feature extraction module as $FE_T(x_i^T)$.

The reason for choosing ResNet architecture is that it can deepen the feature extraction network to obtain richer information and feature distribution. Meanwhile, the shortcut network protects the integrity of the information, which can guide the network to learn the difference between them, thereby removing activity-irrelevant information. Adopting multiple residual blocks is to derive deep-level representation concealed in the input features.

5.3.2 Base Activity Classification Module. FE-Net captures high-dimensional representations of samples that can be used to classify. classification module contains fully connected (FC) and Dropout layers. We use the FC layers to learn a linear or non-linear combination function of these high-dimensional representations. It is usually followed by a softmax layer for getting the probability distribution of an input belonging to a particular class. The softmax layer converts the inexplicable logits z_i of each class output by FC layers into probabilities,

$$P_{hard} = softmax(C(FE_T(x))) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{9}$$

where $C(\cdot)$ represents the output of the FC layers. We call P_{hard} hard probability distribution. Higher logit scores get higher probability values, and other low values give a minimal probability value close to 0. The class with the highest probability is the predicted result for the input sample. The base classification loss is the cross entropy



(a) Hard prob distribution. (b) Soft prob distribution 1. (c) Soft prob distribution 2.

Fig. 9. Probability distribution of "slide" generated by classifier.

(CE) between true labels y_i and hard probability distribution as follows,

$$L_{hard} = CE(P_{hard}, y) \tag{10}$$

5.4 Knowledge Transfer Principle

However, base model training with small samples easily leads to the model over-fitting. The essence is that the sample diversity in the training set is insufficient. The deep neural network cannot learn the deep correlation characteristics of the samples. Thus, we leverage the knowledge from massive WiFi samples to increase the generalization of the radar model. Section. 4 analyzes that the dynamics information implied in the signal originates from human activities. It motivates us to leverage the intuition that the probability distributions generated by a deep learning network for radar and WiFi samples of the same class are correlated. Therefore, we take the probability distribution produced by the WiFi model's *softmax* as the 'guide labels' for training the radar model. Our basic method is to teach the radar model how to generalize, enabling the radar model to have greater generalization performance with fewer samples.

Nonetheless, directly transferring general knowledge using the WiFi-generated hard probability distributions as guide labels is ineffective. The reason is that the non-linear transformation *softmax* amplifies the difference between logits, resulting in a loss of relevant information between classes in the probability distribution. We take an example of recognizing 6 gestures, including "push", "draw a circle", "throw", "slide", "sweep" and "draw zigzag". Fig. 9(a) shows the WiFi model generated hard probability distribution with a WiFi sample input. It has the highest probability in the category "slide," and the probability of other categories are close to 0. The prediction result is undoubted "slide". However, the information that the gesture "slide" is similar to "sweep" and least similar to "push" is not reflected in hard probability distributions.

To tackle this issue, we leverage the idea of knowledge distillation (KD) [13] that was first proposed to compress a large neural net into a small net. The effectiveness of KD has been demonstrated in computer vision [19, 47] and text natural language processing (NLP) [38, 52]. We use the idea of distillation learning to extract more generalization knowledge from WiFi data. We divide the original logits by a factor ε to get smoothed logits. We call ε smooth factor. Then we perform the *softmax* on the smoothed logits output by FC layers, which can be expressed as,

$$P_{soft} = \frac{\exp\left(z_i/\varepsilon\right)}{\sum_j \exp\left(z_j/\varepsilon\right)} \tag{11}$$

We denote the acquired probability distribution as a soft probability distribution. Compared with a hard probability distribution, it contains more information. For example, Fig. 9(b) and Fig. 9(c) are the soft probability

distributions generated by classification module for two different WiFi samples for the same gesture, respectively. Although the probability of "slide" has dropped, it is still the highest, indicating that the prediction is correct. Furthermore, we see that there is an increase in the probability of the gesture "sweep" and gesture "draw circle", which indicates that the similarities of the gesture "slide" with "sweep" and "draw circle" are higher than other gestures. Thus, using these diverse soft probability distributions generated by WiFi as guidance labels is more helpful for the radar model to learn how to generalize.

Specifically, we use the 'guide labels' produced by a pre-trained WiFi model to supervise the training of the radar's classification module. We denote the soft probability distributions for WiFi and radar samples as,

$$P_{soft}^{wifi} = softmax(C(FE_S(x_i^S))), P_{soft}^{Radar} = softmax(C(FE_T(x_i^l))) \tag{12}$$

where x_i^S and x_i^l are the WiFi and radar samples with the same true label. We adopt the similarity measure function to estimate the difference between the $P_{soft}^{wifi}(x_i^S)$ and $P_{soft}^{Radar}(x_i^l)$. We finally chose the mean square error (MSE) based on our multiple experimental comparisons with other metrics, such as Maximum Mean Discrepancy (MMD), cosine similarity, and KL divergence. The corresponding soft loss can be defined as,

$$L_{soft} = MSE(P_{soft}^{Radar}(x_i^l), P_{soft}^{wifi}(x_i^S)). \tag{13}$$

We aim to minimize the difference between the two distributions so that the knowledge of the WiFi is transferred to the radar model.

5.5 Pre-trained WiFi-Net

5.5.1 Data Processing. One raw CSI trace in SignFi [29] is a complex matrix of $200 \times 30 \times 3$, where "200" refers to the number of sampling points, "30" refers to the number of subcarriers, and "3" refers to the three receiving antennas. The sampling clock and the carrier frequency of the transmitter and receiver are not synchronized, which results in sampling time offset (STO) and sampling frequency offset (SFO). Random phase shifts introduced by STO and SFO make dynamic feature extraction difficult. To estimate and eliminate the phase offset, SignFi [29] adopts the method of optimization to minimize the linear fitting error. After that, the pre-processed CSI phases are unwrapped to recover the lost information. In addition, we extract the differential signal by calculating the difference between two consecutive samples to remove the static clutter in the environment and further highlight the dynamic patterns caused by the activities. Finally, we rearrange the complex matrix of $200 \times 30 \times 3$ to a real matrix of $200 \times 60 \times 3$, where the first 30 values and the last 30 values are the magnitude and phase of the complex values, respectively. Actually, we do not specify the feature used for training WiFi-Net. Any feature, such as phase, amplitude, and doppler frequency shit (DFS), can be used in RF-CM. Therefore, even if we don't know all the specifics of a work's processing methods, we may still use its published preprocessed datasets.

5.5.2 Feature Extraction Module. In our method, an excellent WiFi feature extraction network is necessary since it determines what valuable knowledge WiFi-Net learns from WiFi datasets. We do not specify the feature extraction network for WiFi. It could be the CNN network presented by SignFi or other designed deep neural networks as long as they can extract good feature representations and have high accuracy. In our system, WiFi uses the same network structure FE-Net with different hyperparameters as radar. We denote the output of the WiFi feature extraction module as $FE_S(x_n^S)$, where x_n^S is one WiFi sample.

5.5.3 Classification Module. WiFi's classification module contains 2 FC layers and a softmax activation layer. Its classifier training loss is the cross entropy (CE) loss function as follows,

$$L_{WiFi} = CE(softmax(C(FE_S(x_n^S))), y_n^S) \tag{14}$$

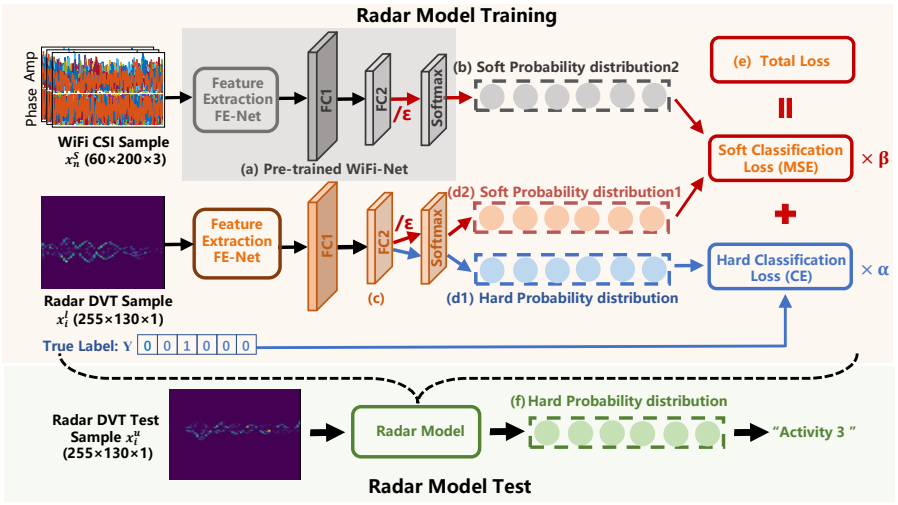


Fig. 10. The training and testing flow of RF-CM. In the training stage, (a) pre-train the WiFi-Net with WiFi datasets. (b) Input a corresponding WiFi sample to WiFi-Net to generate the soft probability distribution 2. (c) Forward propagate the radar samples through FE-Net and two FC layers. (d1) Input the result to softmax layer to generate the hard probability distribution. (d2) Generate soft probability distribution 1 by dividing a factor ε and inputting it into the softmax layer. (e) combine the hard loss and soft loss with weights. In the test stage, (f) input the test sample to obtain the predicted label.

5.6 End-to-end Few-shot Radar Model

To conclude, the total loss of our end-to-end radar model KT-Net is the weighted sum of hard classification loss and soft classification loss, which can be described as,

$$L_T = \alpha L_{hard} + \beta L_{soft} \tag{15}$$

where α and β are trade-off weights to balance the importance between L_{hard} and L_{soft}. We train the end-to-end network by minimizing L_T. In this way, we can transfer valuable knowledge from WiFi to radar by fine-tuning the radar network with guide labels, making the KT-Net perform well with a small number of radar samples.

5.7 Pipeline Summary

We now summarize the entire pipeline of RF-CM as shown in Fig. 10. Initially, we train a good recognition model WiFi-Net on the public labeled WiFi sample set D_S in advance. At the training phase of the radar model, each labeled radar sample x_i^l, is paired with a random WiFi sample x_n^S, having the same activity label Y (One-Hot encoded). The WiFi sample x_n^S passing through a WiFi-Net will generate a soft probability distribution Ps_n^S. We use it as the guide label. In the same way, the corresponding radar sample x_i^l generates a soft probability distribution Ps_i^l and a hard probability distribution Ph_i^l. Finally, KT-Net is trained by minimizing loss function L_T. One is the similarity between Ps_n^S and Ps_i^l, and the other is the cross entropy of Ph_i^l and Y. At the testing stage, only unlabeled radar samples x_i^u need to be input, and the final predicted results are decided on the hard probability distribution.

6 EXPERIMENT

In this section, we evaluate the performance of the proposed RF-CM. We first demonstrate the effectiveness of WiFi cross mmWave radar by realizing a radar-based American Sign Language (ASL) recognition system with the

Table 2. The ASL Words Dataset.

Category	Words
pronoun	who, my, you, what, your
noun	time, drink, food, year, mother, family, book, boy, car, bicycle
verb	help, want, like, need, finish
adjective	cold, old, black, green, black, white, red, gray, nice, hot, sad, sorry, with, without, bad
adverb	where, please, more

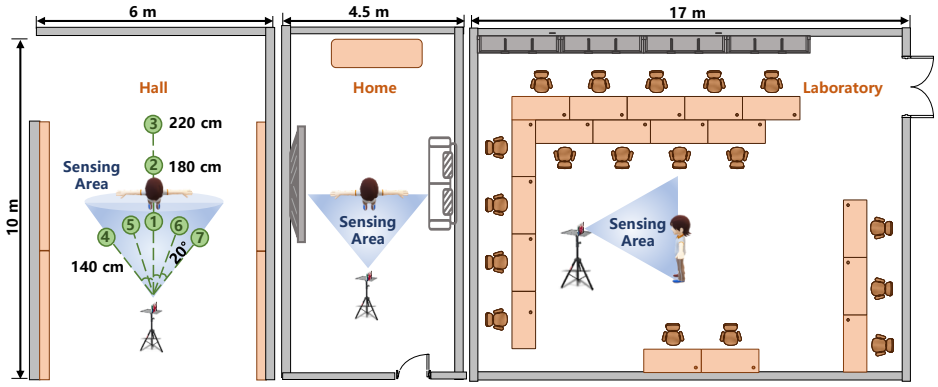


Fig. 11. Experimental environments.

help of public WiFi datasets. In addition, to demonstrate the universality of RF-CM cross to other RF signals, we implement an RFID-based coarse-grained gesture recognition system to verify the feasibility of WiFi cross RFID.

6.1 Implementation

Since it is difficult to gather all words in the ASL vocabulary, we concentrate on a small part of ASL words that are commonly used in each category. We chose specifically 38 ASL words from the pronoun, noun, verb, adjective, and adverb categories, as shown in Table 2.

6.1.1 WiFi Sensing Data Collection and Pre-trained WiFi-Net. The WiFi dataset was chosen based on the assumption that there is only one pair of WiFi AP and client (not specifying device locations). We use the public dataset of SignFi [29] for training an ASL recognition network in advance. The dataset includes CSI traces of 276 sign words. Thereinto, there are 5520 instances collected from the lab and 2760 instances from home. For pre-training WiFi-Net, we selected 760 samples of the 38 words listed in Table 2 from the lab environment dataset (by default). The reason is that the recognition accuracy of mixing the two datasets is lower than that of one of the environments, just like the result reported by SignFi. We used 80% and 20% of WiFi data for the training and testing of WiFi-Net respectively, and obtained an average recognition accuracy of 97.638%.

6.1.2 mmWave Sensing Data Collection. The data collection of mmWave radar is built on a mmWave radar sensor chip (Texas Instruments TI-IWR1843 [37]) with 3 sending antennas and 4 receiving antennas. The sensor is set to transmit an FMCW signal with a bandwidth of 4 GHz (77 GHz-81 GHz) and a 100 MHz/us slope. 130 frames are sent each time, and each frame contains 255 chirps. We set sampling points per chirp to 80. As for the receiver, we set the sample rate to 2.499 MHz. Table 3 summarizes the amount of data collected in three scenarios, namely the hall, the home, and the laboratory, as shown in Fig.11. The impact factors experiments are conducted in the

Table 3. Summary of radar dataset.

Evaluation		Data Samples Collection
Overall Performance (**Dataset 1**)		Default Scenario: Hall, Radar-user 1.4m, Angle 0°, 10 Users × 38 Signs × 10 Instances=3800 Samples
System Robustness	3 Environments (**Dataset 2**)	1. Hall: 2 Users × 9 Signs × 15 Instances = 270 Samples 2. Home: 2 Users × 9 Signs × 15 Instances = 270 Samples 3. Laboratory: 2 Users × 9 Signs × 15 Instances = 270 Samples
	3 Distances (**Dataset 3**)	140cm/180cm/220cm: 3 Distances × 2 Users× 9 Signs × 15 Instances = 810 Samples
	5 Angles (**Dataset 4**)	-40°/-20°/0°/20°/40°: 5 Angles × 2 Users× 9 Signs × 15 Instances = 1350 Samples

hall, where positions 1-3 have different distances of 140 cm, 180 cm, and 220 cm, respectively, and positions 1 and 4-7 have angle intervals of 20°. We use **Dataset 1** for the overall performance and verification of modules. **Dataset 2-4** are used to evaluate robustness.

6.1.3 Platform for Data Processing and Recognition. The data processing coded in Python runs on a computer (Intel i5-10400F CPU @ 2.90 GHz, 32 GB memory). The model training is conducted with Keras library (Python 3.7) on a server with GeForce RTX 2080 Ti GPU.

6.1.4 Hyperparameters Declaration and Sample Selection Strategy. In our training process, we employ 500 iterations (epochs) to train the network using the "rmsprop" optimizer with an initial learning rate of 0.01. We set the batch size to 16. That is, each iteration involves 16 radar and 16 WiFi samples. Specifically, we first loop through each radar sample from the radar training set (76 samples, assumpting 2 samples per gesture are used for training) and each WiFi sample with the same label from the WiFi training set (760 samples). Thus, after 48 iterations, we can traverse all the radar and WiFi samples. Then, to ensure the trained radar model has a good generalization, we randomly choose 16 radar samples from the radar training set and corresponding WiFi samples with the same label from the WiFi training set in each subsequent iteration. Finally, the training process is complete when the number of iterations arrives at 500. The accuracy of the validation data set no longer rises after the 450th epoch usually. This strategy can help the trained radar model has a good generalization.

6.1.5 Baseline. In our comparisons, we take the base model that trains without the help of WiFi as the baseline. It contains the feature extraction module FE-Net and the base classification module with hard classification loss.

6.2 Overall Performance.

We evaluate the overall performance of RF-CM on **Dataset 1**. Specifically, we select 20% and 80% of each participant's radar samples as training and testing sets. Fig. 12 shows the confusion matrix of 38 ASL words across 10 participants. We can see that RF-CM achieves an average recognition accuracy of 83.19% when each participant performs two shots per ASL word for training. Compared with baseline (72.24%), RF-CM has a about 11% recognition accuracy improvement. The results demonstrate that RF-CM can effectively transfer the knowledge from the WiFi dataset to the radar model.

6.3 The Necessity of Model Components

To assess the efficacy of RF-CM's key components, we run the following benchmarks on **Dataset 1**.

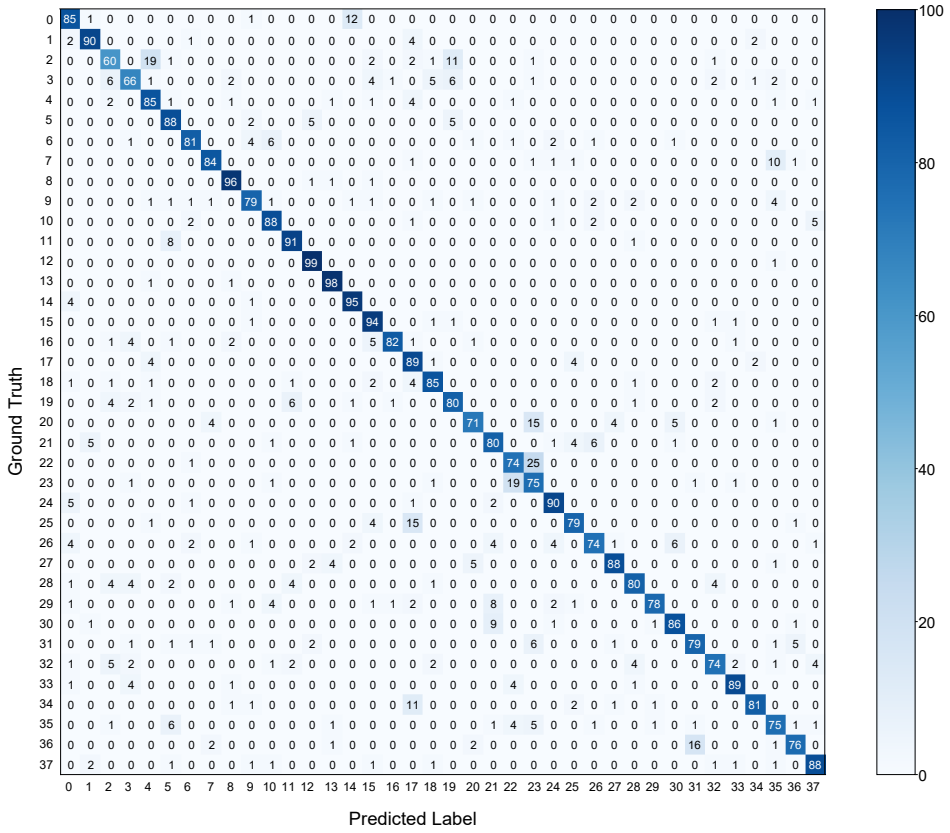
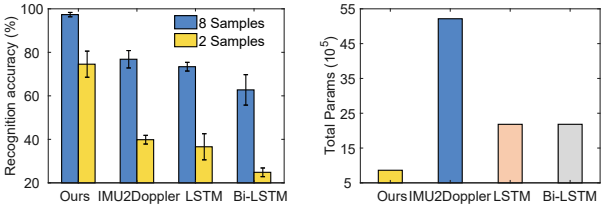


Fig. 12. Confusion matrix of 38 ASL word across 10 participants using only 2-shot per gesture.

6.3.1 Verification of Feature Extraction Network FE-Net. We compare the recognition performance of our FE-Net with three networks, i) the 3-layer 2D CNN presented by IMU2Doppler (each CNN unit consists of a Convolutional layer, a Batch Normalization layer, and a Max Pooling layer); ii) CNN+LSTM, a 1-layer CNN and 2 Long Short Term Memory (LSTM, Units: [128, 256]); iii) CNN+Bi-LSTM, a 1-layer CNN and a bidirectional LSTM (Bi-LSTM, Units: [128]). In addition, we examine the performance of these networks on two different sample sizes. To make a fair comparison, we use the same training set and test set.

Fig. 13(a) shows the recognition accuracy of these four networks trained with 8 shots and 2 shots per gesture. When we employ 8 shots, our proposed FE-Net outperforms the other three comparison models with an accuracy of over 96.5%. The feature extraction network proposed by IMU2Dopller [3] has an accuracy of 76.31%, which is higher than CNN+LSTM (72.895%) and CNN+Bi-LSTM (62.237%). However, when the training sample decreases to 2 shots, the recognition accuracy of all networks declines dramatically. The reductions of these four networks are 24.26%, 37.50%, 36.81%, and 37.90%, respectively. The reason is that the classification error of samples outside the training set will rise due to over-fitting if the training sample size is too small. Nevertheless, the amount of FE-Net drop is less than others, demonstrating the advantage of FE-Net in the case of a small sample size. Fig. 13(b) shows their respective training parameters are 857510, 5214758, 2181542, and 2182054. FE-Net has significantly fewer parameters than others.



(a) Performance comparison. (b) Total trainable parameters.

Fig. 13. The performance of feature extraction methods.

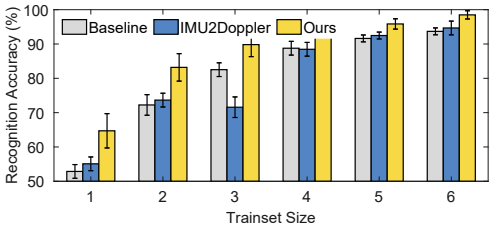


Fig. 14. The performance comparison of knowledge transfer methods.

Table 4. The results for different values of smooth factor ε and loss coefficients (α, β).

	Recognition Accuracy (Improvement)								
(α, β)	ε **= 2**	ε **= 3**	ε **= 4**	ε **= 5**	ε **= 6**	ε **= 7**	ε **= 8**	ε **= 9**	ε **= 10**
(0.5, 1.5)	81.45	**82.93 (10.69)**	79.97	81.88	80.89	81.32	81.12	80.07	80.26
(0.7, 1.3)	82.53	**83.68 (11.44)**	**83.45 (11.21)**	82.53	81.58	80.92	80.86	82.01	82.20
(1, 1)	**82.83 (10.59)**	**82.73 (10.49)**	80.95	82.50	82.24	82.34	80.79	81.61	81.84

6.3.2 Verification of Knowledge Transfer Network KT-Net. We compare our KT-Net with i) baseline and ii) FE-Net with the knowledge transfer method proposed by IMU2doppler [3]. IMU2Doppler adopts domain adaptation approach by decreasing the domain discrepancy between the latent representation of IMU and radar. In particular, IMU2Doppler minimizes the output of the second-last fully connected layer, which has an identical configuration in both the pre-trained IMU model and to be trained radar model. The ratio of the loss function used in IMU2Doppler is the one with the best performance that we have determined through many experiments. The comparison results in Fig. 14 indicate that KT-Net has superior performance than baseline and IMU2Doppler for fewer training samples. For example, when 2 shots per ASL word are used for training, the baseline and IMU2Doppler obtained 72.24% and 73.65% respectively, while KT-Net can achieve an accuracy of 83.19%. Compared with the baseline, the improvement of IMU2Doppler is limited, even decreasing at times. This might be because when the wireless signal is used as the source domain, the latent feature representation learned by the model cannot deliver the correct knowledge to the target model. Furthermore, we can see that compared with the baseline, the improvement of the recognition accuracy of KT-Net and IMU2Doppler becomes smaller as the number of samples rises. For example, the accuracy improvement of KT-Net is 10% with 1 shot for training, while the improvement decreases to 4.5% when model training with 6 shots. This is because that with more samples, the generalization performance of the baseline method will become better.

6.3.3 Dependence on Public WiFi Dataset Deployment. We retrain a new WiFi-Net using another dataset of SignFi collected from the home environment. Its accuracy is 98.01%. And then, we retrain a new radar model using the soft probability distribution generated by the new WiFi-Net on this home dataset as guiding labels. We compare the recognition accuracy of the new radar model with the original model, which uses the same radar dataset. As shown in Fig. 15, the recognition accuracy of the two models is similar, and the gap is about 0.1% to 2.43%. This is in line with the theory that we focused on the accuracy of the WiFi pre-trained model and paid little attention to the deployment of WiFi dataset acquisition. This is because we only require the WiFi data to provide the probability distributions of categories generated through the network. Therefore, the deployment of public WiFi datasets can be varied as long as their feature extraction and classification accuracy are good enough.

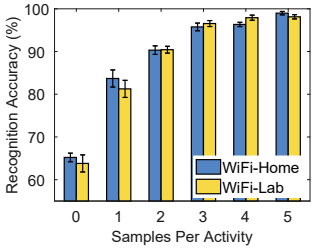


Fig. 15. Recognition performance of radar model trained on various WiFi datasets.

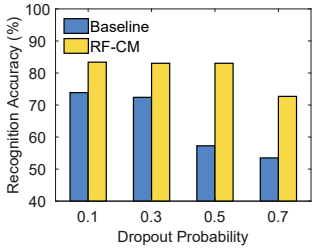


Fig. 16. The impact of Dropout probability.

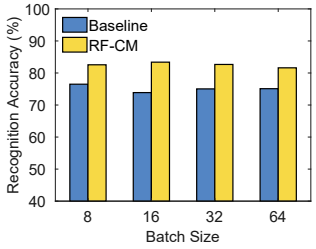


Fig. 17. The impact of batch size.

6.4 Impact of Different Hyperparameters on Generalizability.

6.4.1 Smooth Factor ε and Loss Coefficient. ε affects the diversity of soft probability distributions and the value of β indicates how much WiFi contributes to the radar model. We change ε from 2 to 10 and choose three ratios of hard loss and soft loss (α, β). The result is shown in Table 4. According to the result, we choose the values of (α, β) = (0.7,1.3) and ε = 3 which can produce the best recognition accuracy.

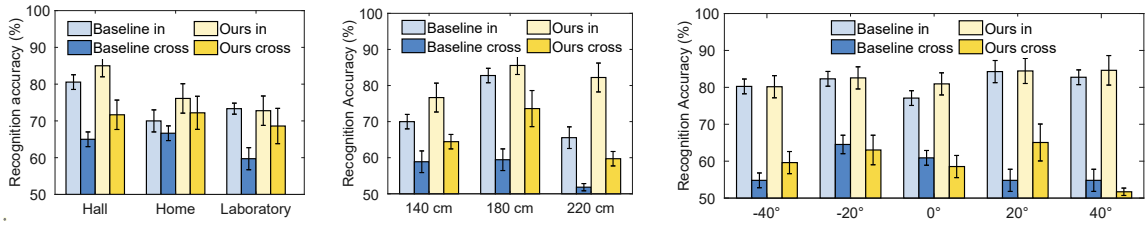
6.4.2 Dropout Probability. The dropout layer is used to prevent overfitting. We set the dropout probability from 0.1 to 0.7 to train and test the radar model and baseline. The comparison results in Fig. 16 show the performance of our system almost does not suffer when the dropout probability is between 0.1 and 0.5, however, the recognition accuracy of the baseline degrades significantly as the dropout probability increases.

6.4.3 Batch Size. We set the batch size to 8/16/32/64 to verify the impact of batch size on the model's generalization performance. Fig. 17 illustrates that the batch size has little effect on the generalization performance of the radar model and baseline.

6.5 Robustness of RF-CM in the Field

To examine the robustness of RF-CM, we use the datasets collected under different domain factors (experimental environments and user positions, described in Sec. 6.1). User position contains two factors: the distance and angle between the user and the radar. Note that we guarantee that all other factors are the same when evaluating each domain factor. We use the leave-one-subject-out cross-validation protocol to evaluate the robustness. Specifically, we choose 2 samples per word and user from each domain for training, 2 samples per word for validation, and the rest for testing. We explain the legend used in Fig. 18 in advance. We denote 'Baseline in' and 'Ours in' as the accuracy given by baseline and RF-CM when the training and testing samples are in the same domain. we denote the corresponding accuracy given by baseline and RF-CM as 'Baseline cross' and ' Ours cross' when training and testing samples are from different domains. For example, we train a recognition network by using the samples collected in the hall environment and then test it with the data from home. We denote this case as 'Baseline/Ours cross'.

6.5.1 Impact of Environment. We use **Dataset 2** to evaluate the performance of RF-CM cross-environment case. The three environments and devices deployment are shown in Fig. 11. The participants stand right in front of the radar with a distance of 140 cm and relative orientation of 0°. We train RF-CM model and base model in one environment (hall) and predict the radar samples collected in another environment (e.g., home and laboratory). Fig. 18 (a) reports the in-domain and cross-environment performance comparison of baseline and RF-CM. When



(a) Different environments. (b) Different distances. (c) Different orientation of users.

Fig. 18. The cross-domain performance of RF-CM.

both training samples and test samples are from hall, the results of baseline and RF-CM are 80.55% and 85%, and when the network trained by the other two environment samples is tested with hall's samples, the accuracy are 65% and 71.66%, respectively, which is a decrease of about 14%. The analysis of the results in other environments is the same as in the hall. Compared with in-domain, RF-CM-Cross accuracy of home and office decreases by about 4%. However, the accuracy of RF-CM-Cross in the hall reduces dramatically. The reason is that performance of the network trained by samples in home and office is relatively poor since the two environments have more complex transmission channels. It is worth noting that although the cross-environment accuracy of RF-CM drops, it is still higher than the baseline.

Fig. 19 shows the cross-environment accuracy of RF-CM decreases compared to the in-domain where the test environment is the same as the training environment (hall). To solve this issue, we can employ the Continual Learning framework [5] to improve the performance. Specifically, when a trained radar model needs to be used in a new environment, we can use a small number of new samples to fine-tune the network to adapt to the new environment. To prove the feasibility of this method, we train a radar model using Hall's data in Table 3, 20% of which was used for training and 80% for testing. When we test this model directly with the home environment data, the accuracy is only 72.93%. Next, we froze the feature extraction layers of the radar model and retrained the model's classifier with a small amount of home environment samples. The result is shown in Fig. 20. We can see that the accuracy increases with the number of samples used for fine-tuning. The accuracy achieves 82.134% when fine-tuning the network with 2-shot. As recognition models are applied in a new environment, more sensing data can be used for model fine-tuning. Therefore, this approach is feasible and worth further exploration.

6.5.2 Impact of Distance Between User and Radar. We use **Dataset 3** to evaluate the performance of RF-CM in the cross-distance case. This experiment was conducted in the hall and the angle between participants and radar remains 0°unchanged. Participants perform gestures at three distances, 140 cm/180 cm/220 cm. Fig. 18 (b) shows the cross-distance accuracy of both baseline and RF-CM are reduced by an average of 15% (compared with in-domain). The reason is that the change in the relative distance between the participants and radar causes the signal strength to become weaker.

6.5.3 Impact of User's Orientation. We use **Dataset 4** to evaluate the performance of RF-CM in a cross-user-orientation case. This experiment was conducted in the hall and the distance between participants and the radar was kept 140 cm unchanged. The participants perform gestures at five positions at 20-degree intervals in the range of $-40°$ to $40°$. The results in Fig. 18 (c) indicate that the cross-angle robustness of baseline and RF-CM is poor as the accuracy drops significantly when testing the model trained on samples from other angles, whether it is the baseline or RF-CM. The reason is features of the same gesture under different orientations of users vary significantly and fail to support successful recognition.

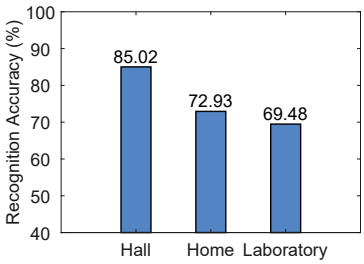


Fig. 19. Test with different environments .

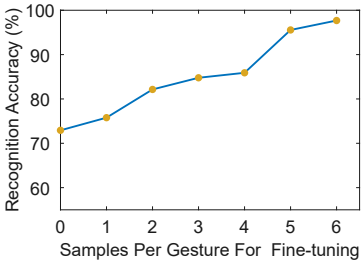


Fig. 20. Fine-tune the trained model with samples from the new environment.

6.6 Extensibility Verification of KT-Net

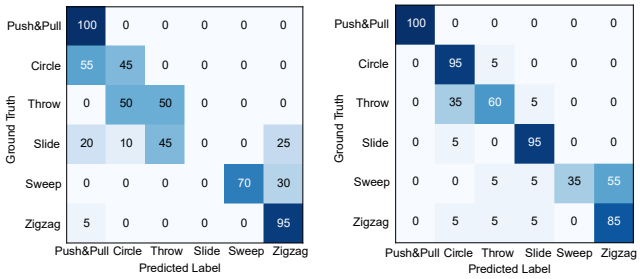
We next assess the generalizability of KT-Net adaption to other sensors-based applications. We realize an RFID-based coarse-grained gesture recognition system with WiFi's help. Neural networks for RFID and WiFi are the same but with different parameters. The feature extraction module contains a Convolutional layer, a Batch Normalisation layer, and a Max Pooling layer. Two fully-connected layers (Units: 128, 6) and a softmax activation were added at the end of the model for classification. We use the cross entropy loss function to optimize the outputs of the final layer with the Adam optimizer (Learning Rate: 0.001, coupled with a learning rate decay of 0.1).

6.6.1 Datasets Description. We used the WiFi and RFID datasets published by RISE [51]. They were collected in a controlled environment using a robotic arm to perform 6 Gestures. With both sets of devices, each gesture was performed 30 times for a total of 180 samples. WiFi devices have 3 transmit antennas and 3 receive antennas respectively. The sample rate was set to 1 kHz, and data collection of each gesture lasts for 8 s. Thus, the size of each original sample is $8000 \times 3 \times 3 \times 30$. The reader and RFID tags are placed at the location of the WiFi transmitter and receiver, respectively. There are 3 tags placed at intervals. For each gesture, the reader receives 40 responses from each tag, corresponding to 40 amplitude and phase values.

6.6.2 WiFi Data Processing. We first intercept the signal segment with dynamic changes through the threshold method, and then adopt the 3-100 Hz Bartworth band-pass filter to filter out high-frequency noise and low-frequency interference signals. Then, since the sample rate is much higher than the rate of gesture change ($\leq$ 100 Hz), so we downsample the signal to reduce the size of one sample while keeping the dynamic characteristics of the gesture. Finally, the differential signal is used to eliminate the reflected signal of the static object so that only the dynamic signal changes caused by the gesture are retained. In this work, we perform the above processing operations on the data from one transmit antenna and extract the amplitude of CSI as the feature of each gesture. The size of each sample after processing is $365 \times 30 \times 3$.

6.6.3 RFID Data Processing. We first smooth the raw phase and amplitude by using a median filter to remove outliers caused by the sudden noise. And then we calculate the difference between two consecutive phase values and amplitude values. Finally, we combine the differential phase and amplitude values sampled during a gesture into a column, namely 80×1. The received data from three tags were combined to build up a sample with a dimension of $80 \times 1 \times 3$.

6.6.4 Overall Performance. Firstly, we trained the feature extractor and classifier using 80% of the WiFi data and obtained an average classification accuracy of 97.2%. Then we train a feature extraction network and classifier with 80% of the RFID samples, which is the baseline shown in the following results. Finally, we use the soft



(a) Baseline (60.3%) (b) RF-CM (78.3%)

Fig. 21. Confusion matrices of baseline and RF-CM when only using one-shot per gesture of RFID dataset.

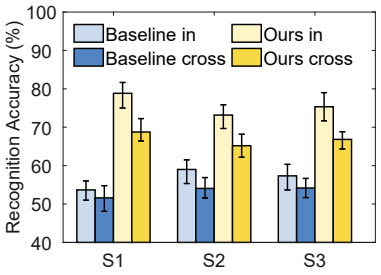


Fig. 22. The recognition accuracy of cross user-position.

Table 5. The 1-shot results for different values of Loss coefficients (α, β) and ε.

	Recognition Accuracy (Improvement)								
Loss	ε = **2**	ε = **3**	ε = **4**	ε = **5**	ε = **6**	ε = **7**	ε = **8**	ε = **9**	ε = **10**
(0.5, 1.5)	70.83 (8.80)	68.33 (6.30)	69.17 (7.13)	68.33 (6.30)	63.17 (1.13)	62.33 (0.30)	**74.83 (12.80)**	71.50 (9.46)	65.50 (3.46)
(0.7, 1.3)	62.83 (0.80)	72.67 (10.63)	63.50 (1.46)	68.50 (6.46)	**72.73 (10.70)**	71.33 (9.30)	**73.33 (11.30)**	68.83 (6.80)	66.50 (4.46)
(1, 1)	67.67 (5.63)	**74.50 (12.46)**	66.33 (4.30)	69.17 (7.13)	66.83 (4.80)	67.83 (5.80)	**75.00 (12.96)**	70.00 (7.96)	69.67 (7.63)

probability distribution obtained from WiFi samples as the label to guide the training of further refinement of the RFID model. Fig. 21 shows the confusion matrices of the baseline and RF-CM. Compared to the baseline, the average accuracy improved by 14% with the help of WiFi.

6.6.5 Impact of the Position of Robotic Arm. S1-S3 in the dataset [51] refer to the locations where the gestures were performed by a robotic arm. We use them to verify the robustness of RF-CM cross-user-position. The network was trained on the premise that one shot per gesture was used as training set. As shown in Fig. 22, when the training samples and the test samples are both from S1, the in-domain accuracy of baseline and RF-CM are 53.67% and 78.83%. When the network trained in S1 was tested using the other two environmental samples, the results of baseline and RF-CM cross-domain are 51.60% and 68.75%, respectively. The analysis in S2 and S3 is the same as that in S1. Although the recognition accuracy of cross-position RF-CM decreases, it is still higher than the baseline.

6.6.6 Impact of Different Parameters. To examine the optimal parameters, we run lots of experiments on various parameter settings (ε and loss ratio (α, β)). We change the smooth factor ε from 2 to 10 and choose three ratios of hard loss and soft loss (α, β). Table 5 shows that the (α, β) = (1,1) and ε = 8 produce the best-performing recognizer. In addition, these two parameters also need to be adjusted when using different training set sizes. It is worth noting that the best values chosen in this case study are not necessarily valid in sec. 6.2. We take designing a method of dynamically changing these values as our future work.

7 DISCUSSION

RF-CM is the first attempt to investigate the viability of leveraging freely accessible WiFi data to support other novel sensors to construct HAR systems. Naturally, there is still much space for future work and further improvement. We discuss a few points here.

Limited Recognition Accuracy for Real Applications. Our results show that knowledge transfer can assist develop a radar model with limited training samples. Towards two-shot training, our method recognizes 38 sign language words with an average accuracy of 83.19% (compared to 72.24% with baseline). This precision is unsatisfactory right out of the box. However, we believe that alternative measures, such as data augmentation and adaption approaches, can be employed to increase accuracy.

Limited Robustness of Cross-domain. We evaluate the system's robustness by changing the environment, user position, and angle. Although the result is superior to the baseline, the recognition accuracy drops significantly when the domain factor is altered. Especially in the cross-user-orientation case, the accuracy plummets when we use data from a specific angle to test a model trained on data from other angles. Nevertheless, we can manually analyze the original data and extract features unrelated to these domain factors like widar3.0 [55], improving the system's robustness across domains in the future.

Strongly Reliant on the WiFi. A fundamental limitation of our approach is that our work is data- and network-dependent on WiFi. Such approaches typically require a 'good enough' WiFi model that can provide the target model with accurate knowledge. Therefore, designing an excellent feature extraction network and classifier for WiFi data is critical. Additionally, the quality of WiFi data significantly impacts knowledge transfer performance. The recognition accuracy of pre-trained WiFi-Net in our system is greater than 97%. In the future, we may construct the network with dynamic parameters ε and loss ratio (α, β), allowing us to find acceptable values for altering the guide label's contribution dynamically in the target model according to the WiFi sample quality.

The Limitation of Low-resolution WiFi Data. Due to the lower resolution, the WiFi data cannot fully reveal the details captured by mmWave radar data. Thus, there exists a possible degradation of data quality (and thus model quality) when transferring from WiFi to RF. However, our work aims to reduce the collection cost of radar samples required by radar model training. The key to enabling our goal is to transfer the sample diversity information of the massive WiFi datasets to the radar model to achieve a generalization performance comparable to training the model with lots of radar samples. In the future, we can use the incremental learning framework [2] to update the pre-trained network with newly collected mmWave samples to improve the radar model's accuracy.

The Applicability of the Methodology to Other Cross-modality Transfer. Although we only focus on the cross-modality transfer between different RF technologies, such as WiFi cross Radar and WiFi cross RFID, in this work, we believe our method can be applied to more sensors, e.g., IMU cross mmWave. This is because the key enabler of our methodology is to use knowledge distillation from one modality to help another. In addition, our methodology can be used in different domains, such as medical imaging applications. For example, we can employ our methodology to distill knowledge from a mass of imaging data obtained by a low-resolution technology and transfer the distilled knowledge to help high-resolution technology quickly identify images with a lower collecting effort.

8 RELATED WORK

In this section, we will discuss the related studies in RF-enabled human activity recognition and cross-domain research.

8.1 RF-based Human Activity Recognition.

RF sensing is an attractive solution for HAR as it neither requires the user to carry the device nor does it have the privacy concerns of cameras and sound.

8.1.1 WiFi-based HAR. WiFi has become a powerful technique for wireless sensing.[29, 33, 39, 42, 54] were designed to extract channel state information (CSI) from WiFi packets and recognize motions of the human body. Wisign [54] recognizes continuous sentence-level ASL rather than separate words via WiFi signals. FingerDraw [42] realized the first sub-wavelength level finger motion tracking system using commodity WiFi devices without attaching any sensor to fingers. GoPose [33] presented a 3D skeleton-based human pose estimation system. WifiU [39] captured fine-grained gait patterns to recognize humans.

8.1.2 Millimeter wave Based HAR. Millimeter wave radar has been widely used in wireless sensing applications [16, 18, 35, 36, 40] due to its wide bandwidth and high resolution. [23] and [30] use point cloud information extracted by commercial millimeter-wave radar to recognize gestures. [25] uses millimeter-wave radar to realize a real-time user-independent gesture recognition system. It extracts the unique characteristics of every gesture and proposes a neural network to build a compact model to inhibit individual differences between different users, which does not require additional radar data collection and retraining. mHomeGes [24] proposed a real-time mmWave arm gesture recognition system for practical smart home usage. m3Track [15] realized a mmWave-based multi-user 3D posture tracking system.

8.2 Cross-domain Human Activity Recognition

CrossGR [20] extracts user-agnostic but gesture-related WiFi signal features to realize cross-target gesture recognition system. [10] designs a position-independent feature MNP which captures the pattern of moving direction changes of hand to achieve cross-position gesture recognition. Widar3.0 [55] achieves a zero-effort cross-domain gesture recognition system by using the unique velocity profile of gestures. CrossSense [53] adopts transfer learning to achieve cross-site sensing. These systems are all supposed to employ the same RF sensor and only investigate changing one or more domain factors in the user, location, and environment, but none of them consider sensing scenarios across various sensing modalities. TARF [45] designs a technology-agnostic network for RF-based HAR by proposing a novel data generation technique to reduce differences in measurements from different RF devices. IMU2Doppler [3] leverages the enormous IMU public dataset to assist mmWave in developing an activity recognition model. Nevertheless, the IMU-assisted RF training model has lower recognition accuracy than our method. This is because the sampling rate of the IMU is usually lower than that of the RF sensors. Based on the last work [46], a signal with a lower sampling rate can degrade the sensing performance. Therefore, it could lose some information when using the IMU data to cross RF data. [4] converts already-available online videos to RF data for training RF-enabled HAR, which significantly reduces the cost of RF data acquisition. However, the method may be suitable for large-amplitude activities like sports, but for gesture recognition, the recognition accuracy will be decreased since complex interference factors in actual wireless transmission are removed from the transform model. Different from the above-related works, in this paper, RF-CM aims to build a few-shot HAR model for a new sensor by leveraging the large sample from other sensors. It is a general cross-modal framework for human activity recognition and can be extended to multiple sensing modalities. RF-CM can use more public datasets as the source domain regardless of data features, which benefits from that it has no restrictions on the processing methods and feature extraction models of both sensors.

9 CONCLUSION

We have presented RF-CM, a novel cross-modal human activity recognition system. RF-CM aims to significantly reduce the overhead of collecting training data across modalities. RF-CM applies a series of signal processing schemes to dispel the activity-agnostic features. It then uses a neural network to further extract the activity-related features for WiFi and radar data, respectively. RF-CM leverages the generalization knowledge from massive WiFi data to teach the radar model to maximize the information extracted from limited radar training samples during deployment. We demonstrate the effectiveness of RF-CM by applying it to 2 representative human activity recognition systems (i.e., WiFi-cross-radar and WiFi-cross-RFID). Our extensive evaluation illustrates that RF-CM has good expansibility for supporting fine-grained (i.e., American sign language) and coarse-grained (i.e., gesture) cross-modal applications. The result demonstrates that RF-CM is a new low-cost approach for cross-modal human activity recognition with high accuracy.

ACKNOWLEDGMENTS

This work is supported by NSFC A3 Foresight Program Grant 62061146001. This work is also partially supported by the National Natural Science Foundation of China under Grant Nos.61972316 and 62002291.

REFERENCES

[1] Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. 2021. Vid2Doppler: synthesizing Doppler radar data from videos for training privacy-preserving activity recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–10.

[2] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. 2021. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks* 135 (2021), 38–54.

[3] Sejal Bhalla, Mayank Goel, and Rushil Khurana. 2021. IMU2Doppler: Cross-Modal Domain Adaptation for Doppler-based Activity Recognition Using IMU Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–20.

[4] Hong Cai, Belal Korany, Chitra R Karanam, and Yasamin Mostofi. 2020. Teaching rf to sense without rf training measurements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[5] Yifan Chang, Wenbo Li, Jian Peng, Bo Tang, Yu Kang, Yinjie Lei, Yuanmiao Gui, Qing Zhu, Yu Liu, and Haifeng Li. 2021. Reviewing continual learning from the perspective of human-level intelligence. *arXiv preprint arXiv:2111.11964* (2021).

[6] Zhe Chen, Tianyue Zheng, Chao Cai, and Jun Luo. 2021. MoVi-Fi: Motion-robust vital signs waveform recovery via deep interpreted RF sensing. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 392–405.

[7] Cao Dian, Dong Wang, Qian Zhang, Run Zhao, and Yinggang Yu. 2020. Towards domain-independent complex and fine-grained gesture recognition with RFID. *Proceedings of the ACM on Human-Computer Interaction* 4, ISS (2020), 1–22.

[8] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. RF-net: A unified meta-learning framework for RF-enabled one-shot human activity recognition. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 517–530.

[9] Lixin Duan, Dong Xu, and Ivor Tsang. 2012. Learning with augmented features for heterogeneous domain adaptation. *arXiv preprint arXiv:1206.4660* (2012).

[10] Ruiyang Gao, Mi Zhang, Jie Zhang, Yang Li, Enze Yi, Dan Wu, Leye Wang, and Daqing Zhang. 2021. Towards position-independent sensing for gesture recognition with Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–28.

[11] Jian Gong, Xinyu Zhang, Kaixin Lin, Ju Ren, Yaoxue Zhang, and Wenxun Qiu. 2021. RF Vital Sign Sensing Under Free Body Movement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–22.

[12] Linlin Guo, Lei Wang, Chuang Lin, Jialin Liu, Bingxian Lu, Jian Fang, Zhonghao Liu, Zeyang Shan, Jingwen Yang, and Silu Guo. 2019. Wiar: A public dataset for wifi-based activity recognition. *IEEE Access* 7 (2019), 154935–154945.

[13] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).

[14] Kun Jin, Si Fang, Chunyi Peng, Zhiyang Teng, Xufei Mao, Lan Zhang, and Xiangyang Li. 2017. Vivisnoop: Someone is snooping your typing without seeing it!. In *2017 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 1–9.

[15] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. m3Track: mmwave-based multi-user 3D posture tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 491–503.

[16] Soo Min Kwon, Song Yang, Jian Liu, Xin Yang, Wesam Saleh, Shreya Patel, Christine Mathews, and Yingying Chen. 2019. Hands-free human activity recognition using millimeter-wave sensors. In *2019 IEEE International Symposium on Dynamic Spectrum Access Networks*

(DySPAN). IEEE, 1–2.
[17] Chenning Li Li, Manni Liu, and Zhichao Cao. 2020. WiHF: Gesture and user recognition with WiFi. *IEEE Transactions on Mobile Computing* (2020).
[18] Guangzheng Li, Ze Zhang, Hanmei Yang, Jin Pan, Dayin Chen, and Jin Zhang. 2020. Capturing human pose using mmWave radar. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 1–6.
[19] Shuang Li, Binhui Xie, Jiashu Wu, Ying Zhao, Chi Harold Liu, and Zhengming Ding. 2020. Simultaneous semantic alignment network for heterogeneous domain adaptation. In *Proceedings of the 28th ACM international conference on multimedia*. 3866–3874.
[20] Xinyi Li, Liqiong Chang, Fangfang Song, Ju Wang, Xiaojiang Chen, Zhanyong Tang, and Zheng Wang. 2021. Crossgr: accurate and low-cost cross-target gesture recognition using Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–23.
[21] Chi Lin, Tingting Xu, Jie Xiong, Fenglong Ma, Lei Wang, and Guowei Wu. 2020. WiWrite: An accurate device-free handwriting recognition system with COTS WiFi. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 700–709.
[22] Haipeng Liu, Kening Cui, Kaiyuan Hu, Yuheng Wang, Anfu Zhou, Liang Liu, and Huadong Ma. 2022. mTransSee: Enabling Environment-Independent mmWave Sensing Based Gesture Recognition via Transfer Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–28.
[23] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, 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.
[24] 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.
[25] Haipeng Liu, Anfu Zhou, Zihe Dong, Yuyang Sun, Jiahe Zhang, Liang Liu, Huadong Ma, Jianhua Liu, and Ning Yang. 2021. M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar. *IEEE Internet of Things Journal* (2021).
[26] Jinyi Liu, Youwei Zeng, Tao Gu, Leye Wang, and Daqing Zhang. 2021. WiPhone: smartphone-based respiration monitoring using ambient reflected WiFi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–19.
[27] Xiulong Liu, Dongdong Liu, Jiuwu Zhang, Tao Gu, and Keqiu Li. 2021. RFID and camera fusion for recognition of human-object interactions. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 296–308.
[28] Chengwen Luo, Zhongru Yang, Xingyu Feng, Jin Zhang, Hong Jia, Jianqiang Li, Jiawei Wu, and Wen Hu. 2021. RFaceID: Towards RFID-Based Facial Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–21.
[29] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. SignFi: Sign Language Recognition Using WiFi. *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–21.
[30] Sameera Palipana, Dariush Salami, Luis A. Leiva, and Stephan Sigg. 2021. Pantomime: Mid-Air Gesture Recognition with Sparse Millimeter-Wave Radar Point Clouds. *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies* 5, 1 (2021), 27.
[31] Sandeep Rao. 2017. Introduction to mmWave sensing: FMCW radars. *Texas Instruments (TI) mmWave Training Series* (2017).
[32] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2021. Winect: 3d human pose tracking for free-form activity using commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–29.
[33] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2021. 3D Human Pose Estimation Using WiFi Signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 363–364.
[34] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30.
[35] Arindam Sengupta, Feng Jin, and Siyang Cao. 2020. NLP based skeletal pose estimation using mmWave radar point-cloud: A simulation approach. In *2020 IEEE Radar Conference (RadarConf20)*. IEEE, 1–6.
[36] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. 2020. mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs. *IEEE Sensors Journal* 20, 17 (2020), 10032–10044.
[37] TI. [n. d.]. IWR1843. Website. https://www.ti.com/product/AWR1843.
[38] Lidan Wang, Vishwanath Sindagi, and Vishal Patel. 2018. High-quality facial photo-sketch synthesis using multi-adversarial networks. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 83–90.
[39] Wei Wang, Alex X Liu, and Muhammad Shahzad. 2016. Gait recognition using wifi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 363–373.
[40] Yong Wang, Wen Wang, Mu Zhou, Aihu Ren, and Zengshan Tian. 2020. Remote monitoring of human vital signs based on 77-GHz mm-wave FMCW radar. *Sensors* 20, 10 (2020), 2999.
[41] Haowen Wei, Ziheng Li, Alexander D Galvan, Zhuoran Su, Xiao Zhang, Kaveh Pahlavan, and Erin T Solovey. 2022. IndexPen: Two-Finger Text Input with Millimeter-Wave Radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–39.

[42] Dan Wu, Ruiyang Gao, Youwei Zeng, Jinyi Liu, Leye Wang, Tao Gu, and Daqing Zhang. 2020. FingerDraw: Sub-wavelength level finger motion tracking with WiFi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–27.
[43] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. 2021. OneFi: One-Shot Recognition for Unseen Gesture via COTS WiFi. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 206–219.
[44] Chenhan Xu, Huining Li, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Xingyu Chen, Kun Wang, Ming-chun Huang, and Wenyao Xu. 2021. CardiacWave: A mmWave-based Scheme of Non-Contact and High-Definition Heart Activity Computing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–26.
[45] Chao Yang, Xuyu Wang, and Shiwen Mao. 2022. TARF: Technology-agnostic RF Sensing for Human Activity Recognition. *IEEE Journal of Biomedical and Health Informatics* (2022).
[46] Kun Yang, Xiaolong Zheng, Jie Xiong, Liang Liu, and Huadong Ma. 2022. WiImg: Pushing the Limit of WiFi Sensing with Low Transmission Rates. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.
[47] Yuan Yao, Yu Zhang, Xutao Li, and Yunming Ye. 2019. Heterogeneous domain adaptation via soft transfer network. In *Proceedings of the 27th ACM international conference on multimedia*. 1578–1586.
[48] Xue Ying. 2019. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, Vol. 1168. IOP Publishing, 022022.
[49] Yinggang Yu, Dong Wang, Run Zhao, and Qian Zhang. 2019. RFID based real-time recognition of ongoing gesture with adversarial learning. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 298–310.
[50] Youwei Zeng, Dan Wu, Jie Xiong, Enze Yi, Ruiyang Gao, and Daqing Zhang. 2019. FarSense: Pushing the range limit of WiFi-based respiration sensing with CSI ratio of two antennas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.
[51] Shuangjiao Zhai, Zhanyong Tang, Petteri Nurmi, Dingyi Fang, Xiaojiang Chen, and Zheng Wang. 2021. RISE: Robust wireless sensing using probabilistic and statistical assessments. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 309–322.
[52] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 5907–5915.
[53] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. 2018. CrossSense: Towards cross-site and large-scale WiFi sensing. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 305–320.
[54] Lei Zhang, Yixiang Zhang, and Xiaolong Zheng. 2020. Wisign: Ubiquitous american sign language recognition using commercial wi-fi devices. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 3 (2020), 1–24.
[55] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2021. Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).