XRF55: A Radio Frequency Dataset for Human Indoor Action Analysis

FEI WANG, Xi'an Jiaotong University, China
YIZHE LV, Xi'an Jiaotong University, China
MENGDIE ZHU, Xi'an Jiaotong University, China
HAN DING, Xi'an Jiaotong University, China
JINSONG HAN, Zhejiang University, China

XRF55 page: https://aiotgroup.github.io/XRF55 | **XRF55 codes:** https://github.com/aiotgroup/XRF55-repo

Radio frequency (RF) devices such as Wi-Fi transceivers, radio frequency identification tags, and millimeter-wave radars have appeared in large numbers in daily lives. The presence and movement of humans can affect the propagation of RF signals, further, this phenomenon is exploited for human action recognition. Compared to camera solutions, RF approaches exhibit greater resilience to occlusions and lighting conditions, while also raising fewer privacy concerns in indoor scenarios. However, current works have many limitations, including the unavailability of datasets, insufficient training samples, and simple or limited action categories for specific applications, which seriously hinder the growth of RF solutions, presenting a significant obstacle in transitioning RF sensing research from the laboratory to a wide range of everyday life applications. To facilitate the transitioning, in this paper, we introduce and release a large-scale multiple radio frequency dataset, named XRF55, for indoor human action analysis. XRF55 encompasses 42.9K RF samples and 55 action classes of human-object interactions, human-human interactions, fitness, body motions, and human-computer interactions, collected from 39 subjects within 100 days. These actions were meticulously selected from 19 RF sensing papers and 16 video action recognition datasets. Each action is chosen to support various applications with high practical value, such as elderly fall detection, fatigue monitoring, domestic violence detection, etc. Moreover, XRF55 contains 23 RFID tags at 922.38MHz, 9 Wi-Fi links at 5.64GHz, one mmWave radar at 60-64GHz, and one Azure Kinect with RGB+D+IR sensors, covering frequency across decimeter wave, centimeter wave, and millimeter wave. In addition, we apply a mutual learning strategy over XRF55 for the task of action recognition. Unlike simple modality fusion, under mutual learning, three RF modalities are trained collaboratively and then work solely. We find these three RF modalities will promote each other. It is worth mentioning that, with synchronized Kinect, XRF55 also supports the exploration of action detection, action segmentation, pose estimation, human parsing, mesh reconstruction, etc., with RF-only or RF-Vision approaches.

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

Additional Key Words and Phrases: dataset, human action recognition, Wi-Fi, millimeter-wave radar, RFID, mutual learning

ACM Reference Format:
Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. 2024. XRF55: A Radio Frequency Dataset for Human Indoor Action Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 0 (March 2024), 34 pages. https://doi.org/XXXXXXX.XXXXXXX

Authors' addresses: Fei Wang, feynmanw@xjtu.edu.cn, Xi'an Jiaotong University, China; Yizhe Lv, lvyizhe@stu.xjtu.edu.cn, Xi'an Jiaotong University, China; Mengdie Zhu, zhumengdie@stu.xjtu.edu.cn, Xi'an Jiaotong University, China; Han Ding, dinghan@xjtu.edu.cn, Xi'an Jiaotong University, China; Jinsong Han, hanjinsong@zju.edu.cn, Zhejiang University, China.

2474-9567/2024/3-ART0 $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Radio frequency devices such as Wi-Fi transceivers and radio frequency identification (RFID) tags are widely present in people's daily lives and even play important roles without being noticed. For instance, Wi-Fi is commonly recognized in the realm of network communication, and can localize people coarsely-grained in buildings through MAC address. RF-ID is widely tagged in/on products of markets, ID cards, animals in pastures, etc. Although the initial goal of WiFi and RFID is communication, they use electromagnetic waves propagating in the air, leading to information of the physical layer, e.g., signal amplitude and phase, being sensitive to the movements of surrounding humans. Thus researchers have been exploiting the physical layer information of Wi-Fi and RFID for human action recognition [13, 23, 25, 34, 41, 45, 53, 54, 58, 60, 76]. Compared to camera solutions, RF approaches exhibit greater resilience to occlusions and lighting conditions, while also raising fewer privacy concerns in indoor scenarios. These appealing characteristics even motivate researchers to bring millimeter-wave (mmWave) radars from vehicles into indoor scenarios for action recognition [39, 40, 48, 59, 62] in recent years.

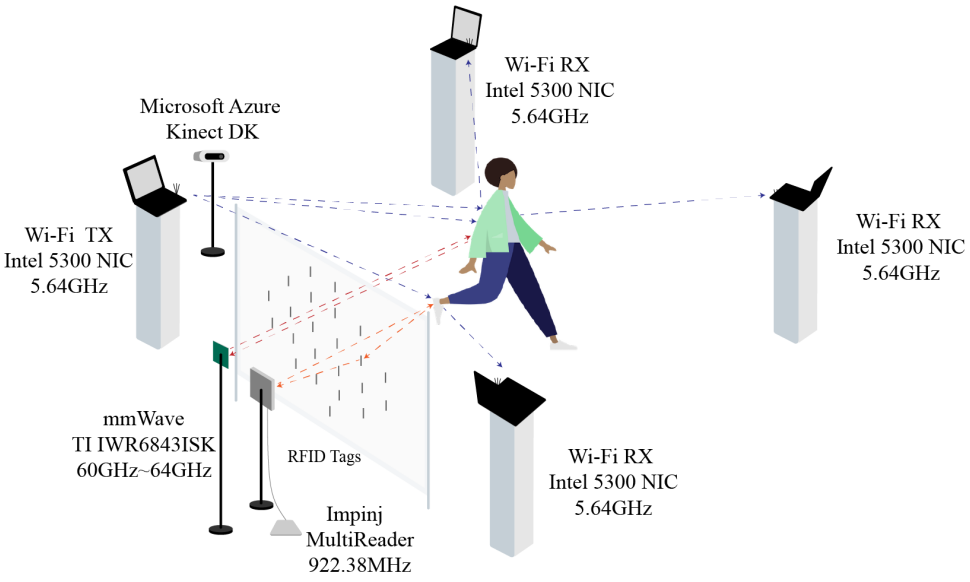


Fig. 1. XRF55 contains 23 RFID tags at 922.38MHz, 9 Wi-Fi links at 5.64GHz, one mmWave radar at 60-64GHz, and one Azure Kinect with RGB+D+IR sensors. XRF55 encompasses 55 action classes, with a total of 42,900 valid data samples, amounting to over 59 hours of effective duration. It is the first RF dataset to possess such a scale of action classes and sample quantity with commercial off-the-shelf devices.

However, current works suffer from several limitations that impede the expansion and development of RF solutions, including public inaccessibility, the lack of training samples, as well as simple or limited action classes that are tailored to specific applications. For example, the majority of existing works have less than 5K samples and 10 action classes, e,g., [13, 25, 48, 59, 65, 68] as listed in Table 1. Several works have more than 10 action classes, e.g., HTHI [3], Widar3.0 [76], RF-finger [54], and mmFit [62], while action classes in these works are designed for specific applications. HTHI includes 12 human-human interaction actions. Widar3.0 includes 6 daily actions such as pushing, pulling, walking, and 10 gestures of drawing 0-9. RF-finger includes gestures of drawing a-z and 10 human-computer interaction gestures. mmFit includes 14 common in-home full-body workouts such

as crunches, leg raise, and push-ups. Though mHomeGe [40] and mTransSee [39] have sufficient samples, while their action classes are very limited, e.g., pull, push, and draw a circle. To our best knowledge, RF-MMD [36] is the first (maybe only) work that has a sufficient variety of action classes as well as samples. However, the hardware used in this work is a homemade centimeter-wave (cmWave) radar, which raises a great barrier for others to deploy and optimize. In addition, datasets in RF-MMD are not publicly accessible yet.

Due to the status, our objective is to create an open-sourced RF dataset with sufficient action classes and samples to facilitate works in RF sensing. When designing the types of actions, we have three main considerations. (1) Indoor actions. RF solutions offer significant advantages over camera solutions in indoor environments such as residences, workplaces, and healthcare facilities, due to their ability to mitigate privacy concerns. (2) More healthcare actions. We believe that identifying healthcare behaviors is an important and valuable application area for RF solutions in indoor environments. (3) Multiple RF modalities. We envision that in the future, more RF devices will enter indoor environments, so it is important to study the collaboration of multiple RF modalities in advance. Thus in this paper, we introduce XRF55, a crossed RF dataset for human indoor action analysis, with 23 RFID tags at 922.38MHz, 9 Wi-Fi links at 5.64GHz, and one mmWave radar at 60-64GHz. These devices are all synchronized, layout shown in Fig. 1. XRF55 includes 55 human indoor action classes and 42.9K action samples in 5 types, i.e., Human-Object Interaction, Human-Human Interaction, Fitness, Body Motion, and Human-Computer Interaction, described in Sec. 3.2.

Further, we apply a deep mutual learning strategy [71] to study the collaboration of Wi-Fi, RFID, and mmWave radar on the action classification task. Under the mutual learning setting, RFID, Wi-Fi, and mmWave radar are three students who learn and teach each other simultaneously to solve the task. We train each student with two losses, i.e., a cross-entropy loss and a Kullback Leibler (KL) Divergence distance to align each student. Trained in this way, RFID, Wi-Fi, and mmWave radar all outperform those trained alone as well as trained with feature fusion. Besides, inspired by [24, 72], we leverage a pre-trained BERT [10] to output semantic embedding over the names of action classes, which we further regard as classification anchors. We compute L1 distance between the learned representation and its corresponding semantic anchor as an auxiliary loss to train the deep network. Experimental results show that BERT will promote RF classification accuracy.

One concurrent work, MM-Fi [66], integrates signals of varying resolutions, including high-resolution LiDAR, mid-resolution mmWave radar, and low-resolution Wi-Fi. Conversely, XRF55 emphasizes the cooperative use of RF signals across different frequency bands, targeting indoor activities where wireless sensing shows the most promise. In comparison, XRF55 surpasses MM-Fi in several aspects: (1) Dataset scale. XRF55's dataset is significantly larger, with 2× action classes, 6.6× duration, and 20× frames than MM-Fi. (2) Action diversity. XRF55 covers a much wider range of actions, supporting varied applications such as domestic violence detection, elderly fall detection, fatigue detection, and human-computer interaction scenarios where hands are preoccupied or unavailable. (3) The hardware layout of XRF55 allows for more flexible and diverse action execution compared to MM-Fi. While MM-Fi's setup limits action space, XRF55's larger, rectangular perceptual region formed by multiple WiFi devices offers greater freedom in action execution.

The contributions of this paper are as follows:

(1) We have created and will release a dataset for human indoor action recognition using RF technology. Our dataset comprises of 23 RFID tags operating at 922.38MHz, 9 Wi-Fi links operating at 5.64GHz, one mmWave radar operating at 60-64GHz, and one Azure Kinect equipped with RGB+D+IR sensors. XRF55 encompasses 55 action classes, with a total of 42,900 valid data samples, amounting to over 59 hours. It is the first RF dataset to possess such a scale of action classes and sample quantity.

(2) XRF55 includes 15 human-object interaction actions, 7 human-human interaction actions, 8 fitness actions, 14 body motion actions, and 11 human-computer instruction actions. These actions were meticulously selected from 19 RF sensing papers and 16 video action recognition datasets. Each action is chosen to support various applications of high practical value, such as elderly fall detection, fatigue monitoring, domestic violence detection,

Dataset	#Actions	#Subjects	#Samples	Modality	Year	Accessible
WiAR [68]	6	6	557	Wi-Fi	2017	Yes
EI [25]	6	11	4h41min	Wi-Fi	2018	Yes
ARIL [55]	6	1	1394	Wi-Fi	2019	Yes
HTHI [3]	12	40	4800	Wi-Fi	2020	Yes
Widar3.0 [76]	16	16	17000	Wi-Fi	2021	Yes
NTU-FI [65]	6	20	2400	Wi-Fi	2022	Yes
FEMO [13]	10	15	1500	RFID	2015	Not yet
RF-finger [54]	36	14	3000	RFID	2018	Not yet
EUIGR [69]	8	15	5400	RFID	2019	Not yet
AAL [51]	11	6	7h12min	RFID	2020	Yes
mHomeGe [40]	10	25	22000	mmWave radar	2020	Yes
m-Activity [59]	5	9	1350	mmWave radar	2021	Not yet
mTransSee [39]	5	32	54080	mmWave radar	2022	Yes
RadarAE [48]	7	8	1094	mmWave radar	2022	Yes
mmFit [62]	14	11	7000	mmWave radar	2022	Not yet
RF-MMD [36]	35	30	25h	cmWave Radar	2019	Not yet
MM-FI [66]	27	40	9h	Wi-Fi, mmWave, RGB+D, LiDAR	2023	Yes
XRF55(Ours)	55	39	42900, 59h35min	Wi-Fi, RFID, mmWave, RGB+D+IR	2023	Yes soon

Table 1. Current works have several limitations, including public inaccessibility, the lack of training samples, as well as simple or limited action classes that are tailored to specific applications, which seriously hinder the expansion and development of RF solutions. We create and will release a multiple radio frequency dataset for human indoor action recognition. XRF55 contains 23 RFID tags at 922.38MHz, 9 Wi-Fi links at 5.64GHz, one mmWave radar at 60-64GHz, and one Azure Kinect with RGB+D+IR sensors. It is the first RF dataset to possess such a scale of action classes and sample quantity.

and even accommodating scenarios where one's hands might be preoccupied or unavailable, like individuals with certain disabilities or temporary injuries. These actions are likely to attract substantial attention from the industrial and academic sectors of the RF sensing community.

(3) We have applied a deep mutual learning strategy to investigate the collaborative potential of Wi-Fi, RFID, and mmWave radar for action classification. Through training with this strategy, we have demonstrated that RFID, Wi-Fi, and mmWave radar all exhibit improved performance compared to when trained individually. Our study represents the first exploration into the mutual learning capabilities of these three RF modalities.

2 RELATED WORK

2.1 RF Sensing

The communication ability of RF signals such as Wi-Fi and RFID is already well-known to the public, while the sensing ability has been exploited in the networking community for decades. For example, RADAR leveraged Wi-Fi for user indoor location and tracking in 2000 [5]. LANDMARC utilized active RFID for indoor localization in 2003 [44]. There are some surveys that describe the booming status [23, 34, 41, 53]. Compared to camera

solutions, RF approaches exhibit greater resilience to occlusions and lighting conditions, while also raising fewer privacy concerns in indoor scenarios. This appealing characteristic also motivates researchers to deploy mmWave radar in indoor scenarios for human sensing [39, 40, 48, 59, 62], to design specific mmWave radar chip such as Soli project [38], and to design novel cmWave radar to track and capture persons such as WiTrack [2] and RF-Capture [1]. In recent years, RF sensing works also appear in CV/AI venues, e.g., ICCV [36, 57, 74], CVPR [30, 73], ICML [75], AAAI [33], MM [35]. The research content of these works includes hand tracking [30], pose estimation [57, 73], mesh reconstruction [74], and action recognition [33, 35, 36]. We believe that with the awakening of privacy awareness, RF solutions will play an increasingly important role in human sensing in indoor scenarios.

2.2 RF Sensing Datasets

Though RF sensing gains wide attention, open-sourced RF sensing datasets are very limited, which seriously hinders the development of RF solutions. Taking the action recognition task for instance, as we surveyed and as listed in Table 1, there is only a very small number of RF datasets that are publicly accessible. In addition, current open-sourced datasets suffer from several limitations, including the lack of training samples, as well as simple or limited action classes that are tailored to specific applications. For example, WiAR [68], EI [25], ARIL [55], NTU-FI [65], mHomeGe [40], mTransSee [39], and RadarAE [48] include simple and limited action classes, e.g., hand up, walking, and running. HTHI [3], AAL [51], and mmFit [62] include more than 10 action classes, which are specially proposed for human-human interactions, daily activities, and fitness, respectively. Widar3.0 [76] includes 16 action classes and sufficient samples, however, its actions are simple such as walking, pulling, pushing, and drawing 0-9. RF-MMD [36] has a sufficient variety of action classes as well as samples. However, the hardware used in this work is a homemade centimeter-wave (cmWave) radar, which raises a great barrier for others to deploy and optimize. Besides, RF-MMD dataset is not publicly available yet. To facilitate the research in RF action recognition, We create and will release a multiple RF dataset for human indoor action analysis, which includes modalities of RFID, Wi-Fi, and mmWave radar.

3 DATASET ACQUISITION AND DESCRIPTION

3.1 Hardware Setups

RFID devices: We use an Impinj Speedway R420 RFID reader with an RFMax S9028PCLJ directional antenna to broadcast the QUERY command 30 times per second at the frequency of 928.33MHz. We deploy 23 passive RFID tags branded Hansense to backscatter their Electronic Product Code (EPC) to the reader. The reader then can obtain the phase of each tag with the backscattered EPC, leading to the phase series with the size of $(30t) \times 23$, where t is for the recording time in seconds.

Wi-Fi transceivers: We use one Thinkpad X201 laptop with one Intel 5300 wireless network card as the Wi-Fi transmitter, and use *three* other sets as the Wi-Fi receivers. The laptops are positioned at the four corners of a rectangle as shown in Fig. 2, inspired by the placement strategy outlined in Widar3.0 [76]. This arrangement creates a larger rectangular sensing area. Their height is set at 1.2 meters, based on observations from ARIL [55], which is found to be effective for recognizing full-body actions. The transmitter is set to broadcast packets with a speed of 200 packets per second through *one* antenna under High Throughput (IEEE 802.11n) bitrates at channel 128 (5.64GHz). Every receiver monitors this channel with *three* antennas, thus there are 9 Wi-Fi links in total. We install a Wi-Fi tool [19] in the transceivers to conduct channel estimation and obtain the channel state information (CSI) of 30 orthogonal frequency-division multiplexing (OFDM) subcarriers, leading to the recorded CSI with the size of $(200t) \times 1 \times 3 \times 3 \times 30$, where t is for the recording time in seconds.

Millimeter-wave radar: We use a TI IWR6843ISK radar to generate frequency-modulated continuous wave (FMCW) signals at the frequency of 60GHz-64GHz. The radar has *three* transmitting antennas and *four* receiving

(a) Scene 1 (b) Scene 2

(c) Scene 3 (d) Scene 4

Fig. 2. XRF55 dataset is collected from four indoor scenes, where hardware devices maintain relative placements and orientations. The relative placements are marked in scene 3.

antennas. The transmitting parameter is set to 20 frames per second and 64 chirps in each frame. The ADC sampling rate is 256. Meanwhile, we mount the radar on a TI DCA1000EVM board to record raw radar data in real time, leading to the raw radar data with the size of $(20t) \times 3 \times 4 \times 256 \times 64$. Then we apply Doppler Fast Fourier Transform (FFT) and Angle FFT over the raw data to obtain Range-Doppler Heatmaps and Range-Angle Heatmaps with the size of $(20t) \times 256 \times 64$. After that, we concatenate these heatmaps along the range dimension to $(20t) \times 256 \times 128$.

Azure Kinect: The Kinect records RGB, depth, and infrared images with 30 frames per second in 720P.

All devices are carefully synchronized before data recording.

We collect XRF55 in four different scenes, as shown in Fig. 2. In scene 1, we record action samples of RFID, Wi-Fi, mmWave radar, and Kinect clips from 30 subjects. Besides, we record action clips from 3 subjects in the other three scenes respectively.

Further explanation for RFID setup. Our RFID setup has some details that differ from existing work, which we will explain next.

• Explanation for RFID position. Our inspiration to use an array of RFID tags for action recognition is derived from RFIPad [12] and RF-finger [54]. RFIPad and RF-finger employ RFID tag arrays to sense hand movements, requiring the tag array to be in close proximity to the hand. Our approach, however, applies RFID tag arrays to the broader context of indoor action recognition, where the space for human movement is significantly larger.

To explore the capability of RFID in sensing movement at longer distances and to determine the most effective placement for optimal sensitivity to human actions, we conducted several preliminary experiments.

Initially, influenced by the designs of RFIPad and RF-finger, we arranged the movement space between the RFID array and the reader's antenna as shown in Fig. 3(a). However, in this configuration, we noticed that the RFID back-scattering signals were minimally responsive to human movement. Suspecting that the array's obstruction by the human body might be causing this, we reconfigured the tags as shown in Fig. 3(b) to reduce obstruction. Despite this rearrangement, the sensitivity of the RFID array in this new setup remained low. This led us to experiment with the placement shown in Fig. 3(c), where we observed a noticeable response of the array to human movement. Following this discovery, we fine-tuned the distance between the tag array and the antenna, ultimately settling on a distance of 40 cm.

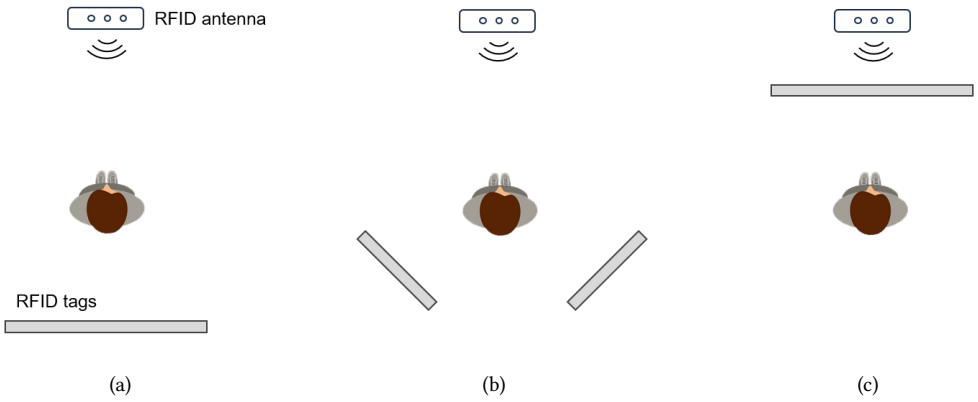


Fig. 3. Preliminary experiment to determine the RFID position. We chose the placement of (C), where we observed a noticeable response of the array to human movement.

• Explanation for RFID attachment. We attached RFID tags to a thin fabric, as shown in Fig. 2. The reason behind this design was our belief that for future applications involving indoor action recognition, RFID tag arrays could be integrated into existing household items such as room screens or curtains.

• Explanation for the number of RFID tags. We initially tested a 6×6 tag array, but for three reasons, we settled on a total of 23 tags. (1) Reducing tag collisions and stabilizing data rates: when the number of tags reached 36, tag collisions prevented the reader from obtaining feedback from each tag at a stable frequency, thus we need to reduce tag number. (2) Eliminating low-sensitivity tags: we observed that some tags on the edges, especially those at the four corners, were less sensitive to human movements, so we removed these tags, bringing the total down to 24. (3) Excluding anomalous tags: upon examining the RFID, we found that the readings from one of the tags were consistently aberrant, thus we excluded that particular tag. Consequently, the final count of RFID tags used was 23, and their relative positions can be seen in the device deployment diagram in Fig. 2.

3.2 Action Selection

Our objective is to create a dataset for human indoor action recognition, utilizing RF solutions that offer significant advantages over camera solutions in indoor environments such as residences, workplaces, and healthcare facilities, due to their ability to mitigate privacy concerns. Our criteria for selecting actions are twofold: first, the actions must be performed in indoor scenarios, and second, they must have sufficient practical utility. Our action

selection includes three stages: (1) selecting indoor actions from large-scale video dataset and RF sensing papers, (2) brainstorming action utilities and excluding actions that are inappropriate or of low practical value, (3) Categorizing actions.

Selecting indoor actions from large-scale video datasets and RF sensing papers. Action recognition is widely studied in the computer vision and RF sensing community. The datasets or papers in these fields contain a large number of action classes, providing us with an excellent action pool for selection. For example, the Kinetics [8] and Moments-in-Time [43] dataset include 700 and 339 action classes, respectively. To select indoor actions, we reference 16 video action datasets, i.e., Kinetics [8], UCF101 [52], HMDB51 [31], Something-Something [17], Charades [49], Moments-in-Time [43], HVU [11], Jester [42], ActivityNet [15], THUMOS14 [27], AVA [18], MMAct [29], Eye in the sky [50], ANUBIS [46], PPMI [67], and FineGym [47]. We also 19 RF sensing papers for action selection, i.e., WiAR [68], EI [25], ARIL [55], HTHI [3], Widar3.0 [76], NTU-FI [65], FEMO [13], RF-finger [54], EUIGR [69], AAL [51], mHomeGe [40], m-Activity [59], mTransSee [39], RadarAE [48], mmFit [62], RF-MMD [36], E-eyes [60], WiSee [45], and CARM [58].

Brainstorming action utilities and excluding actions that are inappropriate or of low practical value. We aim for our selected indoor actions to be fairly general, ideally appearing multiple times across different datasets. Additionally, we hope that the actions hold good practical value, frequently occur in everyday life, and are suitable for volunteers to perform. With these considerations, we filter out actions of checking watch, flipping bottles, pillow fighting, putting on lipstick, saluting, etc., from Kinetics [8]; action of knocking a virtual table twice from mTransSee [39]; actions of pointing to someone, wiping face, etc., from RF-MMD [36]; action of opening refrigerator from CRAM [58].

Categorizing actions. We first refer to the action categorization method in UCF101 [52] and HMDB51 [31], dividing actions into Body Motion Actions, Human-Object Interaction Actions, and Human-Human Interaction Actions. Subsequently, considering the potential application areas of fitness tracking and human-computer interaction in the realm of indoor action recognition, we further delineated categories specifically for Fitness Actions and Human-Computer Interaction Actions.

In all, XRF55 includes 55 human indoor action classes which we categorize into 5 types: Human-Object Interaction, Human-Human Interaction, Fitness, Body Motion, and Human-Computer Interaction. The actions and their sources are listed in Table. 2, with examples of these actions demonstrated in Fig. 4, (potentially supported applications listed in the Appendix).

• **15 Human-Object Interaction Actions. Whole-home daily:** *carrying weight, mopping the floor, using a phone, throwing something, picking something, putting something on the table;* **Kitchen:** *cutting something*; **Dress:** *wearing a hat, putting on clothing*; **Bathroom:** *blowing dry hair, combing hair, brushing teeth*; **Healthcare:** *drinking, eating, and smoking.*

• **7 Human-Human Interaction Actions. Social actions:** *shaking hands, hugging, handing something to someone*; **Violence actions for applications of domestic violence and invasion detection:** *kicking someone, hitting someone with something, choking someone's neck, and pushing someone.*

• **8 Fitness Actions. With equipment:** *hula hooping, weightlifting, jumping rope*; **Without equipment:** *body weight squats, Tai Chi, boxing, jumping jack, and high leg lifting.*

• **14 Body Motion Actions. Whole-home daily:** *waving, clapping hands, jumping, walking, turning, running, sitting down, standing up*; **Healthcare:** *falling on the floor, stretching, patting on the shoulder*; **Musical instruments:** *playing Er-Hu, playing Ukulele, playing drum.*

• **11 Human-Computer Interaction Actions. Hand gestures:** *pushing, pulling, swiping left, swiping right, swiping up, swiping down, drawing a circle, drawing a cross*; **When hands are not free:** *foot stamping, shaking head, and nodding.*

Index	Classes	Video datasets	RF datasets
		Human-Object Interaction	
1	Carrying Weight	Kinetics, AVA, MMAct	EI
2	Mopping the Floor	Kinetics, UCF101	
3	Cutting	Kinetics, UCF101, AVA, ANUBIS	
4	Wearing Hat	Kinetics	
5	Using a Phone	Kinetics, Charades, MMAct	RF-MMD
6	Throw Something	HMDB51, Charades, AVA	RF-MMD
7	Put Something on the Table	Something-Something, Charades, AVA	
8	Put on Clothing	Charades, AVA	AAL
9	Picking	Kinetics, HMDB51, Something-Something, Charades, AVA, MMAct	RF-MMD
10	Drinking	Kinetics, HMDB51, Charades, ActivityNet, ANUBIS	AAL, RF-MMD
11	Smoking	HMDB51, Moments in Time, ActivityNet	
12	Eating	HMDB51, Charades	AAL, E-eyes
13	Brushing Teeth	UCF101, ActivityNet, AVA, ANUBIS	AAL, E-eyes, CARM
14	Blow Dry Hair	Kinetics, UCF101	
15	Brush Hair	HMDB51, Charades, ActivityNet	AAL
		Human-Human Interaction	
16	Shake Hands	Kinetics, HMDB51, AVA, ANUBIS	HTHI, RF-MMD
17	Hugging	Kinetics, HMDB51, Moments in Time	HTHI
18	Hand Something to Someone	Something-Something, AVA, ANUBIS	RF-MMD
19	Kick Someone	HMDB51, AVA, Eye in the Sky	HTHI, RF-MMD, WiSee
20	Hit Someone with Something	HMDB51, Something-Something, AVA, Eye in the Sky, ANUBIS	
21	Choke Someone's Neck	Eye in the Sky, ANUBIS	
22	Push Someone	AVA, ANUBIS	HTHI
		Fitness	
23	Body Weight Squats	UCF101, HVU, ANUBIS	FEMO, m-Activity, mmFit
24	Tai Chi	UCF101	
25	Boxing	UCF101, HMDB51, Moments in Time	NTU-FI, m-Activity, RadarAE, mmFit, CARM
26	Weightlifting	Moments in Time	FEMO
27	Hula Hooping	UCF101, HVU, THUMOS14	
28	Jump Rope	UCF101, THUMOS14	
29	Jumping Jack	UCF101, THUMOS14	m-Activity
30	High Leg Lift	Kinetics	mmFit
		Body Motion	
31	Waving	Kinetics, HMDB51, Moments in Time, AVA, ANUBIS	EUIGR, RF-MMD
32	Clap Hands	Kinetics, HMDB51, Moments in Time, HVU, AVA, ANUBIS	Widar3.0, mHomeGe, RF-MMD
33	Fall on the Floor	HMDB51, AVA, ANUBIS, MMAct	WiAR, NTU-FI, CARM
34	Jumping	Kinetics, UCF101, HMDB51, Moments in Time, THUMOS14, AVA, MMAct	m-Activity
35	Running	HMDB51, Charades, AVA	WiAR, NTU-FI, CARM
36	Sitting Down	HMDB51, Charades, MMAct	WiAR, EI, RF-MMD, CARM
37	Standing Up	HMDB51, Charades	WiAR, EI
38	Turning	HMDB51, Moments in Time, AVA	EUIGR, RadarAE
39	Walking	HMDB51, AVA, ANUBIS	WiAR, EI, NTU-FI, m-Activity, CARM
40	Stretch Oneself	Kinetics	
41	Pat on Shoulder	ANUBIS	RF-MMD
42	Playing Erhu	PPMI	
43	Playing Ukulele	Kinetics, ActivityNet, PPMI	
44	Playing Drum	Kinetics, ActivityNet	
		Human-Computer Interaction	
45	Stomping	Moments in Time	
46	Shaking Head	HVU	
47	Nodding	HVU, ANUBIS	
48	Draw Circles		ARIL, Widar3.0, NTU-FI, mHomeGe, mTransSee, WiSee
49	Draw a Cross		ARIL
50	Pushing	HMDB51, AVA	RF-finger, mHomeGe, mTransSee, WiSee, CARM
51	Pulling	AVA	RF-finger, mHomeGe, mTransSee, WiSee
52	Swipe Left	Jester	ARIL, Widar3.0, RF-finger
53	Swipe Right	Jester	ARIL, Widar3.0, RF-finger
54	Swipe Up		ARIL
55	Swipe Down		ARIL

Table 2. Action classes of XRF55. We select actions from 16 video datasets and 19 RF sensing papers.

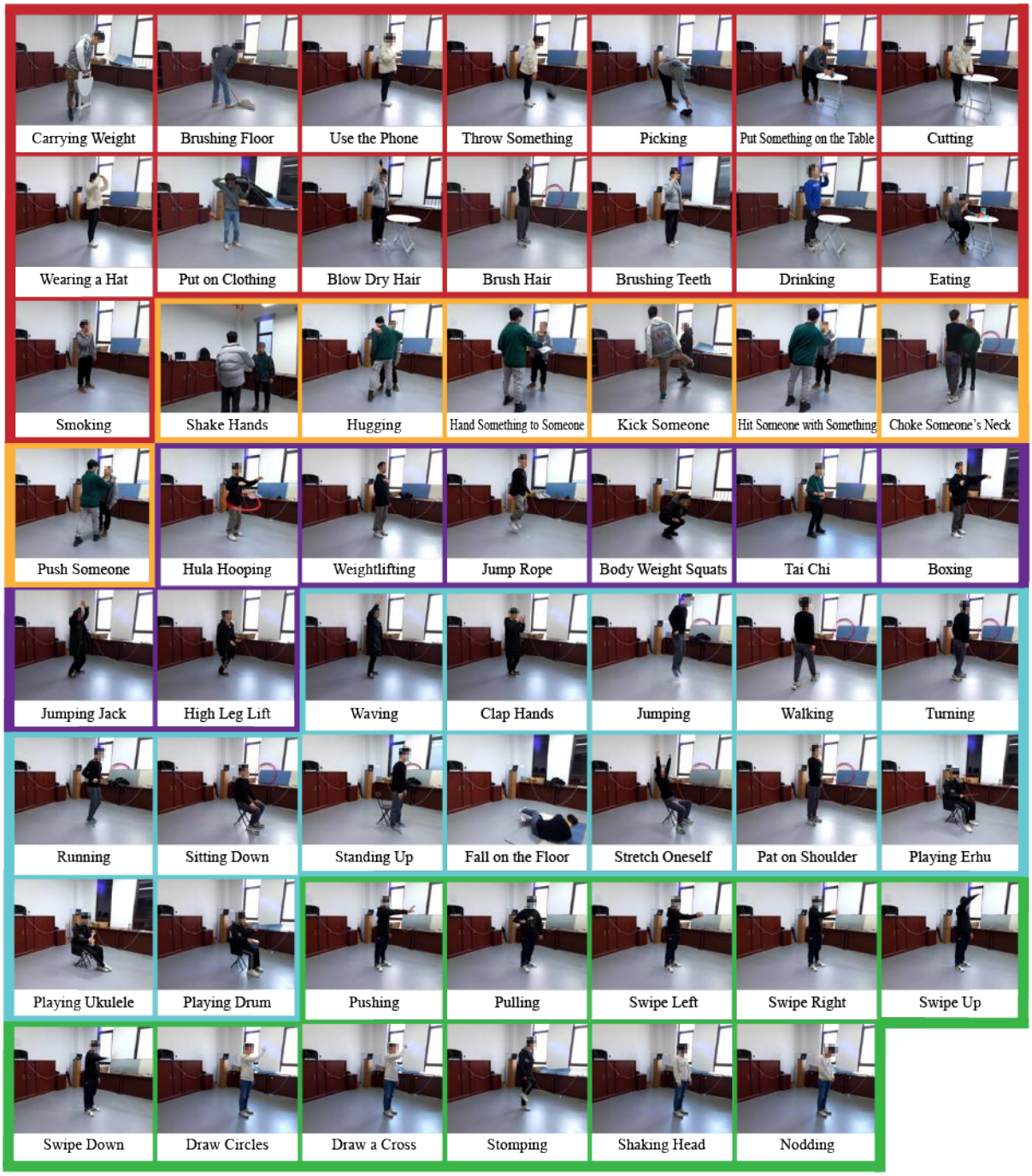


Fig. 4. XRF55 includes 55 human indoor action classes which we categorize into 5 types: Human-Object Interaction, Human-Human Interaction, Fitness, Body Motion, and Human-Computer Interaction.

3.3 Volunteer Coordination

This project received IRB approval from our institute. We recruited volunteers through group chats, social media, and flyers. Those interested joined a group chat we created, where we shared a read-only online spreadsheet. Volunteers checked the spreadsheet and communicated their available time slots in the group chat, which we then updated in the spreadsheet. The data collection schedule was arranged accordingly.

Before data collection, we provided volunteers with information about the research's purpose, steps involved, and potential risks, and reminded them of their right to withdraw at any time. Volunteers were given time to watch our pre-recorded action video examples and communicate any questions or concerns about action execution with us. Each volunteer spent about 30 minutes getting familiar with the actions. During this process, they discussed the action execution details with us. Once the volunteers were ready, we began data collection.

We reorganized the action sequence for two considerations. First, we aimed to use certain props continuously in two actions. For instance, a table was used as a prop for both cutting something and putting something on the table, so we arranged these actions consecutively. Second, we followed physically demanding actions with more relaxed ones. For example, after the physically taxing action of falling, we scheduled less strenuous tasks like putting on clothing. In the fitness category, after intense exercises like weightlifting, body weight squats, jumping jacks, and high leg lifting, we interspersed lighter activities such as picking, waving, clapping hands, and playing Er-Hu. Volunteers could take breaks at any time, and we paused for meals or rest. Everyone performed actions with the same sequences, listed in Table. 3.

1. Shaking hands	20. Playing Ukulele	39. Standing up
2. Hugging	21. Playing drum	40. Sitting down
3. Handing something to someone	22. Jumping	41. Blowing dry hair
4. Hitting someone with something	23. Running	42. Cutting something
5. Choking someone's neck	24. Turning	43. Drinking
6. Pushing someone	25. Walking	44. Putting something on the table
7. Kicking someone	26. Patting on the shoulder	45. Eating
8. Mopping the floor	27. Foot stamping	46. Carrying weight
9. Combing hair	28. Shaking head	47. Brushing teeth
10. Weightlifting	29. Nodding	48. Using a phone
11. Throwing something	30. Drawing a circle	49. Hula hooping
12. Picking something	31. Drawing a cross	50. Jumping rope
13. Body weight squats	32. Pushing	51. Falling on the floor
14. Waving	33. Pulling	52. Putting on clothing
15. Boxing	34. Swiping left	53. Tai Chi
16. Jumping jack	35. Swiping right	54. Stretching
17. Clapping hands	36. Swiping up	55. Smoking
18. High leg lifting	37. Swiping down	
19. Playing Er-Hu	38. Wearing a hat	

Table 3. Action execution sequence. Everyone performed actions with the same sequences.

We recruited 39 volunteers aged 23 to 35, with heights ranging from 161cm to 182cm, weights between 44kg and 73kg, and BMIs from 16.33 to 25. Data collection for each volunteer lasted approximately 4-5 hours. As compensation, we provided each with a 64GB USB 3.0 flash drive, valued approximately equivalent to twice the

local minimum wage for 5 hours. Due to the impact of the COVID-19 pandemic, the data collection process was interrupted multiple times, and it took us nearly 100 days in total to complete the data collection.

	orientation	position	categories	actions
1	significant	significant	Human-Object Interaction	mopping the floor, carrying weight, throwing something, picking something
			Human-Human Interaction	none
			Fitness	none
			Body Motion	walking, running
			Human-Computer Interaction	none
2	significant	slight	Human-Object Interaction	brushing teeth, blowing dry hair, combining hair
			Human-Human Interaction	shaking hands, hugging, handing something to someone, kicking someone, hitting someone with something, pushing someone
			Fitness	none
			Body Motion	turning, waving, clapping hands
			Human-Computer Interaction	none
3	slight	slight	Human-Object Interaction	cutting, wearing hat, using a phone, putting something on the table, putting on clothing, drinking, smoking
			Human-Human Interaction	none
			Fitness	body weight squats, Tai Chi, boxing, weightlifting, hula hooping, jumping rope, jumping jack,
			Body Motion	fall on the floor, jumping, pat on shoulder
			Human-Computer Interaction	none
4	slight	no	Human-Object Interaction	none
			Human-Human Interaction	none
			Fitness	none
			Body Motion	none
			Human-Computer Interaction	foot stamping, shaking head, nodding, drawing a circle, drawing a cross, pushing, pulling, swiping left, swiping right, swiping up, swiping down
5	little	little	Human-Object Interaction	eating
			Human-Human Interaction	none
			Fitness	none
			Body Motion	standing up, sitting down, playing ukulele, playing drum, playing Er-Hu, stretch oneself
			Human-Computer Interaction	none

Table 4. We roughly categorize these actions into five groups based on differences in action execution orientation and location.

3.4 Action Execution

In the data collection phase, we requested volunteers to perform actions within a rectangular space formed by four Wi-Fi devices at their discretion. This setup granted them a degree of freedom in how they performed their actions. Despite this intended flexibility, there were inherent limitations regarding the direction and positioning for certain actions during the data collection process. Based on differences in execution orientation and location, we can roughly categorize these actions into five groups, as listed in Table. 4.

(1) Significant variations in both orientation and position. For actions of walking, running, mopping the floor, carrying weight, throwing something, and picking something, volunteers may start these actions at a different location with a new orientation.

(2) Significant variations in orientation, and slight changes in position. For instance, in the case of turning, volunteers mostly tended to turn by 90 degrees or 180 degrees. After performing a turn, their orientation would change by 90 or 180 degrees, and there would also be some variation in their position. In addition to turning, during actions like waving, clapping hands, brushing teeth, blowing dry hair, combining hair, shaking hands, hugging, handing something to someone, kicking someone, hitting someone with something, choking someone's neck, pushing someone, volunteers were also reminded to frequently change the direction of their movements.

(3) Slight variations in both orientation and position. such as cutting, wearing hat, using a phone, putting something on the table, putting on clothing, drinking, smoking, body weight squats, Tai Chi, boxing, weightlifting, hula hooping, jumping rope, jumping jack, high leg lifting, falling on the floor, jumping, patting on the shoulder.

(4) Slight variation in position, no orientation changes. In all human-machine interaction actions, like foot stamping, shaking head, nodding, drawing a circle, drawing a cross, pushing, pulling, swiping left, swiping right, swiping up, swiping down. We assume that the interacting objects are the millimeter-wave radar or RFID tag array. Hence, volunteers consistently face these two devices while performing actions. During each action, over the 20 repeated movements, volunteers slightly change their positions.

(5) Little variation in orientation and position. For example, actions that utilize a chair as a prop, such as eating, standing up, sitting down, playing ukulele, playing drum, playing Er-Hu, or stretching, typically involve a chair with a relatively stable orientation and position.

It is essential to clarify that even though there are constraints on the direction and position for someone performing certain actions, such as actions in case 5, their execution in terms of position and even orientation varies when different individuals execute the same action.

3.5 Dataset Statistics

We recruited 39 subjects and let them repeat each action 20 times at the sensing area. We set the action-conducting window as 5 seconds for subjects to finish every repeat. These settings result in XRF55 with 42.9K samples that last 59h35min. Each sample is a quadruple comprised of Wi-Fi$\in 1000 \times 1 \times 3 \times 3 \times 30$, RFID$\in 150 \times 23$, mmWave radar$\in 100 \times 256 \times 128$, and corresponding synchronized videos from the Kinect. Further, we reshape or downsample the quadruple to reduce the training overhead to dimensions as shown in Table 5. We designate the first 14 trials of each action performed by each subject as training samples, while reserving the last 6 trials for testing purposes. This results in XRF55 having 30.0K training samples and 12.9K test samples, respectively.

	RFID	Wi-Fi	mmWave	Kinect
Before	18GB	183GB	3250GB	8226GB
After	17GB	87GB	78GB	8226GB
Dimension	148×23	1000×270	$17 \times 256 \times 128$	720P
Statistics	scene1	train: 23.1K, 32h5min / test: 9.9K, 13h45min		
	scene2	train: 2.31K, 3h12min / test: 0.99K, 1h22min		
	scene3	train: 2.31K, 3h12min / test: 0.99K, 1h22min		
	scene4	train: 2.31K, 3h12min / test: 0.99K, 1h22min		
Total	train: 30.0K, 41h42min / test: 12.9K, 17h52min			

Table 5. XRF55 includes 30.0K training samples and 12.9K test samples, respectively. Despite releasing data that is processed by us, we will also release the raw dataset for other researchers to involve their own expertise.

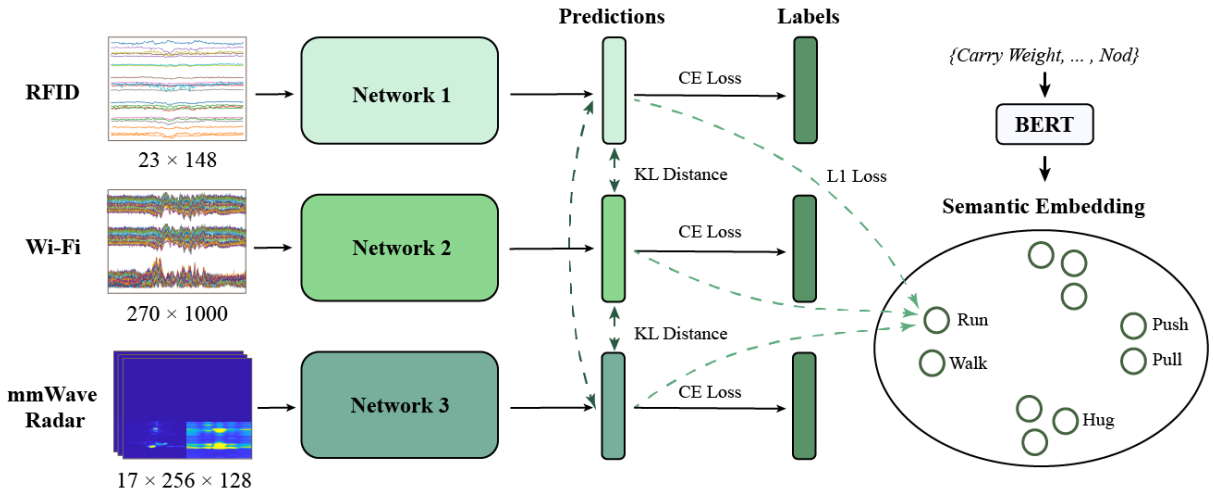


Fig. 5. We apply deep mutual learning over XRF55 datasets. Under DML setting, Wi-Fi, RFID, and mmWave radar are three students who learn and teach each other simultaneously. After training, all students can work alone. We train each network with a supervised cross-entropy loss, two Kullback Leibler (KL) divergence distances to the other two networks, and a BERT semantic loss.

4 MUTUAL LEARNING

XRF55 includes data from 23 RFID tags, 9 Wi-Fi links, and one mmWave radar, covering decimeter waves, centimeter waves, and millimeter waves. We envision that in the future, more RF devices will enter indoor environments, so it is important to study the collaboration of multiple RF modalities in advance. We consider one real requirement that if some RF devices fail to work properly, other devices will not be affected. To meet this requirement, we leverage a deep mutual learning (DML) strategy [71] over our dataset for the task of action recognition. Under DML setting, Wi-Fi, RFID, and mmWave radar are three students who learn and teach each other simultaneously to solve the task. After training, all students can work alone.

4.1 Deep Mutual Learning

We denote N samples of synchronized RFID, Wi-Fi, and mmWave radar from C action classes as $\{x_i^1, x_i^2, x_i^3\}_{i=1}^N$, and the corresponding labels as $\{y_i\}_{i=1}^N$, where $y_i \in \{1, 2, ..., C\}$. DML strategy can be formulated with a cohort of three networks, i.e., $\mathcal{W}_1$, $\mathcal{W}_2$, and $\mathcal{W}_3$, as shown in Fig. 5. Every network is trained with three losses, i.e., a supervised cross-entropy classification loss, and two Kullback Leibler (KL) divergence distances to the other two networks among the learned representation, denoted as Eq. 1.

$$L(\mathcal{W}_u) = L_{CE}(\mathcal{W}_u) + \frac{1}{2} \sum_{v \neq u}^{v \in 1,2,3} D_{KL}(\mathcal{W}_v \| \mathcal{W}_u) \tag{1}$$

where u and v are network indexes.

To compute the loss, we first compute the prediction probability. Give a sample x_i^u, we simplistically denote the output of SoftMax layer of network $\mathcal{W}_u$, i.e., SoftMax($\mathcal{W}_u(x_i^u)$) as z_u. The probability of x_i^u belonging to class c can be computed via Eq. 2.

$$p_u^c(x_i^u) = \frac{exp(z_u^c)}{\sum_{c=1}^{C} exp(z_u^c)}, \tag{2}$$

Thus the supervised cross-entropy loss $L_{CE}(\mathcal{W}_u)$ of network $\mathcal{W}_u$ over N training samples is computed as Eq. 3.

$$L_{CE}(\mathcal{W}_u) = -\sum_{i=1}^{N}\sum_{c=1}^{C} I(y_i, c) log(p_u^c(x_i^u)) \tag{3}$$

where $I(y_i, c)$ is an indicator that outputs 1 if $y_i = c$, otherwise 0.

Meanwhile, KL distance from network $\mathcal{W}_u$ to network $\mathcal{W}_v$ can be computed as Eq. 4.

$$D_{KL}(\mathcal{W}_v \| \mathcal{W}_u) = \sum_{i=1}^{N}\sum_{c=1}^{C} p_v^c(x_i^v) \log \frac{p_v^c(x_i^v)}{p_u^c(x_i^u)}, \tag{4}$$

Though there is no teacher network in DML strategy, our experiments show that all student networks still converge well. This is because all students are directly guided by the supervised cross-entropy loss, which leads them to the right way. Meanwhile, all students learn extra information from their peers through the KL distance, which promotes their own performance (refer to [71] for a more detailed explanation).

4.2 BERT Semantic Embedding

We notice that RF signals exhibit similar patterns in actions such as walking and running, pulling and pushing, etc., much like the similarity in semantic meanings among these words. Inspired by [24, 72] that explore the semantic representation of Wi-Fi signals, we apply BERT [10] to extract the semantic embedding of action names in XRF55. Specifically, we feed the action names of XRF55 into the pre-trained BERT model [61], and take the output of the POOLER layer, which returns the embedding of the classification token, as the semantic embedding of the actions. Further, we take the semantic embedding corresponding to each action class as the classification anchor, and compute the L1 distance between the representation learned by each student network and the anchor as an auxiliary semantic loss to train the network, denoted as Eq. 5.

$$L_{bert}(\mathcal{W}_u) = \sum_{i=1}^{N} L_1(\mathcal{W}_u(x_i^u), BERT(y_i)), \tag{5}$$

Thus the loss in Eq. 1 is expanded to Eq. 6.

$$L(\mathcal{W}_u) = L_{CE}(\mathcal{W}_u) + \frac{1}{2}\sum_{v\neq u}^{v\in 1,2,3} D_{KL}(\mathcal{W}_v \| \mathcal{W}_u) + L_{bert}(\mathcal{W}_u) \tag{6}$$

At last, we use Eq. 7 to train the deep network shown in Fig. 5 collaboratively.

$$L = L(\mathcal{W}_1) + L(\mathcal{W}_2) + L(\mathcal{W}_3) \tag{7}$$

After training, all student networks can work alone.

4.3 Implementation

We implement the network with Pytorch 1.13.1. We train the network with Adam optimizer [28] with a batch size of 64. The initial learning rate is 0.001 with a decay rate of 0.5 at every 40 epochs. We train the network for 200 epochs. Before each epoch, training samples are randomly shuffled.

5 VALIDATION

5.1 Quantitative Results

Metric. We choose action classification accuracy as the metric to evaluate the performance of all methods. Classification accuracy is a straightforward yet powerful indicator, representing the proportion of correctly identified action instances against the total number of instances in the dataset. It is crucial for our study as it

directly reflects the effectiveness of the model in distinguishing between 55 actions, which is essential in the context of action recognition.

Training data and test data. In Sec. 5.1 and Sec. 5.2, we employ the training and test sets from scene 1 for the training and testing sets, respectively. The method for dividing the training and testing data within scene 1 is detailed in Sec. 3.5. Specifically, each volunteer executes each action 20 times, from which the first 14 repetitions are allocated to the training set, while the subsequent 6 repetitions are designated for the test set. Thus, for 55 actions performed by 30 volunteers in scene 1, there are 23,100 samples as the training set and 9,900 samples as the test set.

Backbone network selection.

Recall that the RFID dataset consists of 23 tags and 148 sampling points, resulting in a sample size of 23x148. To process this data, we treat 23 as the channel dimension and 148 as the temporal dimension. We apply long-short-term memory network [22], 1D temporal ViT [14], and 1D temporal ResNet18 [21] on the RFID dataset. Similarly, we apply these networks to the Wi-Fi dataset. For the mmWave radar samples, we apply a 2D spatial ResNet18 and 2D spatial ViT as the mmWave baseline. We select LSTM, ResNet, and ViT as basic networks because they epitomize the three predominant structures in deep learning: RNN, CNN, and Transformer, respectively. RNN, CNN, and Transformer have demonstrated advancements in learning temporal representation for action recognition [4, 20, 37]. Our decision is driven by the desire to explore the intrinsic learning capabilities of these networks on data, rather than pursuing the most advanced algorithms for achieving state-of-the-art (SOTA) results. Therefore, we opt for the most elementary versions of LSTM, ResNet, and ViT to maintain our focus on their core learning abilities. Table. 6 shows that ResNets are strong baselines for temporal RF classification, where DTW means Dynamic Time Warping + 1 nearest neighbor, a very typical time-serial classification approach. Thus we choose ResNet18 as the backbone networks to evaluate XRF55 dataset.

Backbone	RFID	Wi-Fi	mmWave
DTW+1NN	14.2	21.8	9.5
LSTM	58.29	79.41	87.53
ViT	58.34	76.17	54.57
ResNet18	59.22	87.26	87.97

Table 6. We test three principal network types- LSTM (RNN), ResNet (CNN), and ViT (Transformer). Given its outstanding performance on the XRF55 dataset, we choose ResNet as our baseline network architecture.

Network adaptations. The base BERT model generates a semantic embedding of size 1x768. To apply the BERT semantic loss, we propose three basic adaptations to conventional ResNets, as illustrated in Fig. 6. (1) we add a linear layer before the last layer of the ResNet and compute losses in sequence. (2) we add a parallel linear layer along with the last layer and compute the loss in parallel. (3) we set the output channel dimension of the last block of ResNet to 768, apply pooling, and compute losses in sequence. For the large BERT model, we use 1024 instead of 768. In addition, we extend the action phrases to sentences in the form of “I am doing something", such as “I am mopping the floor" and “I am running". These sentences are then fed into a pre-trained BERT model to generate semantic embeddings. In all settings, we compute the cross-entropy loss and mutual learning loss through the 1×55 output with Eq. 2, Eq. 3, and Eq. 4.

Results. Table 7 shows the action classification accuracy on the XRF55 dataset. When RFID, Wi-Fi, and mmWave radar samples are trained with baseline networks solely, the corresponding accuracies are 59.22%, 87.26%, and 87.97%, respectively. This suggests that Wi-Fi and mmWave radar have similar performance in action recognition accuracy, and outperform RFID by a large margin. Furthermore, by incorporating network adaptation

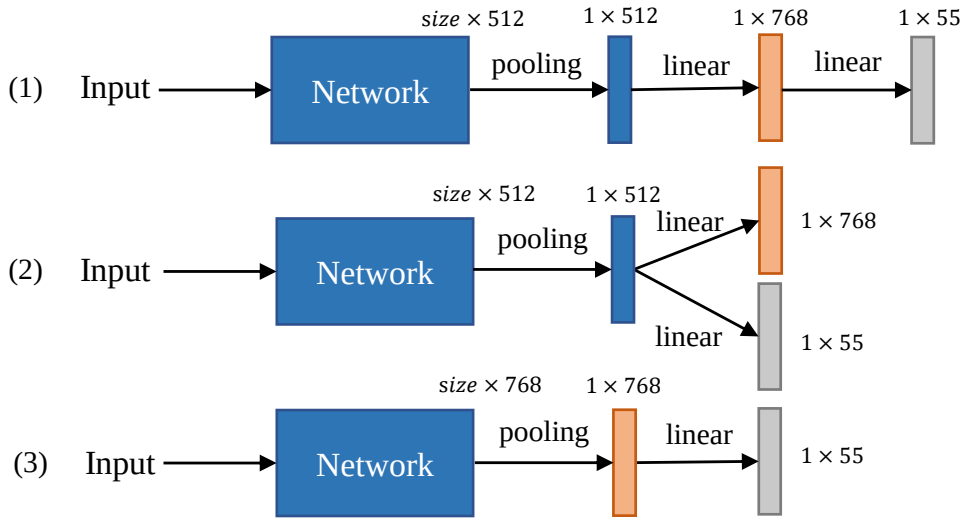


Fig. 6. Three network adaptations to apply semantic embedding loss for the base BERT model. For the large BERT model, 768 is set to 1024.

Network	BERT		RFID	Wi-Fi	mmWave
baseline	/		59.22	87.26	87.97
Fig. 6(1)	base	phrase	58.79	88.52	89.29
Fig. 6(1)	base	sentence	59.89	88.22	89.79
Fig. 6(1)	large	phrase	57.70	87.80	89.51
Fig. 6(1)	large	sentence	57.54	88.27	89.82
Fig. 6(2)	base	phrase	62.90	89.65	90.22
Fig. 6(2)	base	sentence	60.91	89.60	90.61
Fig. 6(2)	large	phrase	60.75	89.15	90.56
Fig. 6(2)	large	sentence	60.59	89.35	90.58
Fig. 6(3)	base	phrase	63.23	89.74	89.93
Fig. 6(3)	base	sentence	62.21	89.78	90.10
Fig. 6(3)	large	phrase	63.64	89.90	90.10
Fig. 6(3)	**large**	**sentence**	64.78	89.73	89.93

Table 7. When trained collaboratively with mutual learning and semantic loss. RFID, Wi-Fi, and mmWave radar gain a noteworthy improvement in action classification accuracy, by as much as 5%, 2%, and 2%, respectively.

in Fig. 6(1) and mutual learning, we observe a boost in accuracy for Wi-Fi and mmWave radar, but a decline in accuracy for RFID. However, by utilizing the network in Fig. 6(2) and engaging in mutual learning, all modalities, including RFID, Wi-Fi, and mmWave radar, exhibit an improvement in accuracy. Moreover, if we train samples from RFID, Wi-Fi, and mmWave radar using the network in Fig. 6(3) collaboratively, we can achieve a noteworthy enhancement in recognition accuracy for RFID, by as much as 3%-5%. Table. 7 indicates that, on XRF55, when the

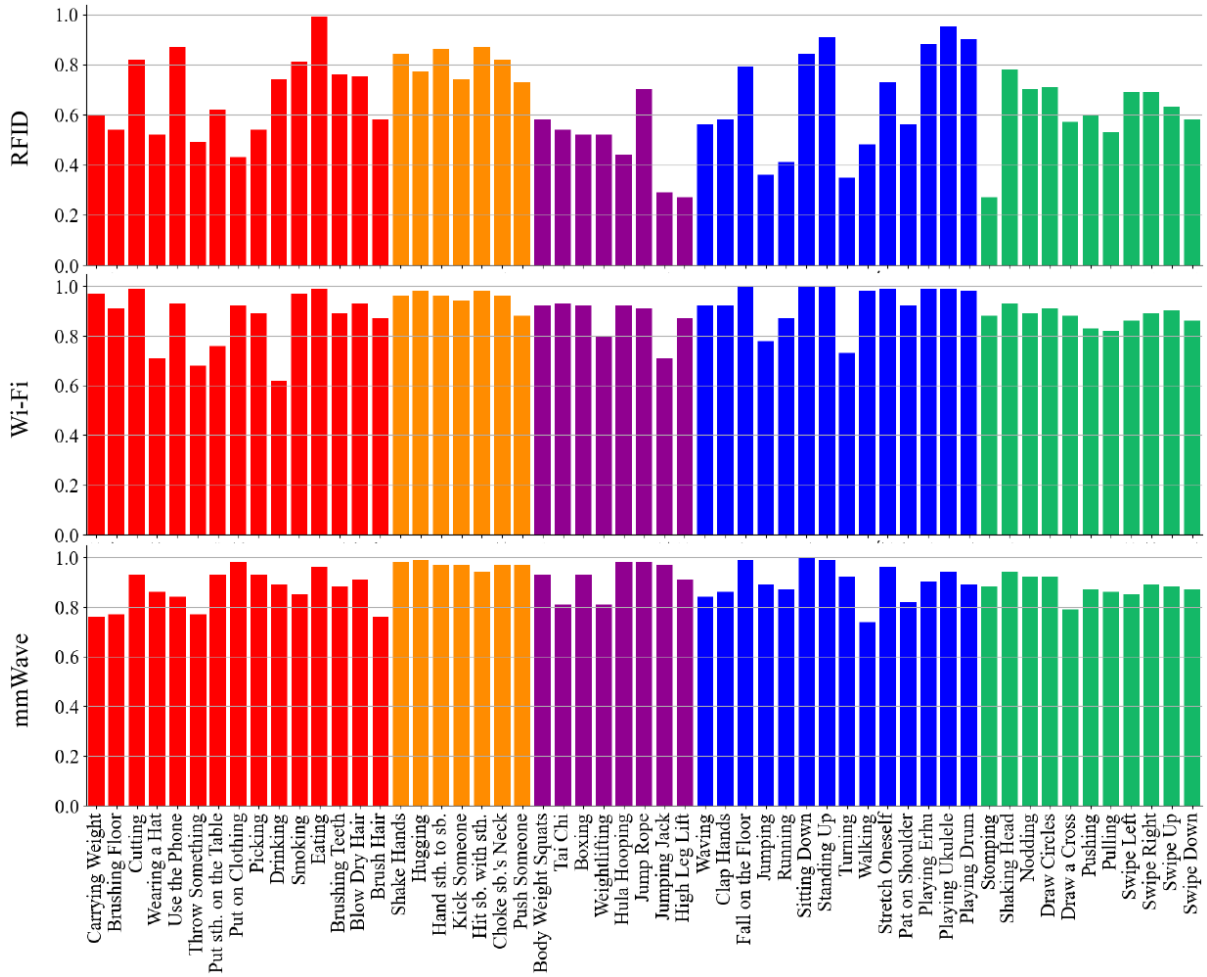


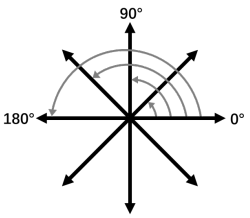
Fig. 7. Action recognition accuracy on XRF55. Wi-Fi and mmWave perform better than RFID. Wi-Fi achieving comparable performance to mmWave technology is unexpected, given its significantly lower cost.

network structure is fixed, whether BERT is configured as 'base' or 'large', and whether it uses phrase embedding or sentence embedding, the impact on results is minimal. For example, in the four settings of Fig. 6(1) in the table, the variation in the accuracy is within 1%. The critical factor influencing outcomes is the choice of network structure. As shown in Table. 7, when Fig. 6(2) is applied instead of Fig. 6(1), there is a 2% increase in RFID accuracy. Furthermore, using Fig. 6(3) leads to an additional 2-4% increase in RFID accuracy over the results of Fig. 6(2).

We opt for the setting in the last row of Table. 7 as it outperforms the others in RFID, while also exhibiting comparable performance in Wi-Fi and mmWave radar. Upon this setting, Fig. 7 shows the action recognition results of RFID, Wi-Fi, and mmWave radar over each action class. We observe that RFID performs the worst in some lower limb actions such as foot stamping, jumping jack, and high-leg lifting, while performing well in several hand movements such as eating, playing Er-hu, and playing the drum. This may be because RFID tags are

more sensitive to movements that are similar in height to their deployment, and less sensitive to movements with a large height difference. An additional noteworthy finding is that the recognition accuracy for human-human interaction actions (indicated in orange) tends to be higher than those involving just a single person. This could be because, in our experiment, we do not distinguish between a target user and other users. Rather, we consider the interactions between two people as a unified entity, leading to larger signal variations and being more discriminant for action classification.

Moreover, we conclude several rules to find out actions that can be easily recognized from Fig. 7. **(1) Consistency.** Actions that are performed in a consistent manner across different sessions. For example, actions that utilize a chair as a prop, such as standing up, sitting down, playing the ukulele, playing the drum, playing the erhu, and stretching, typically involve a user sitting on a chair with a relatively stable orientation and position. These characteristics make such actions more easily recognizable by models. **(2) High Contrast Actions.** Actions that have distinct, easily distinguishable features from other activities. For example, In the action of falling, the body makes contact with the floor, whereas in other actions, there is typically no contact with the ground. Therefore, the action of falling on the floor can be easily recognized by models. **(3) Limited Variability.** Actions that have limited variability in how different individuals perform them are generally easier to classify. For example, human-computer interaction actions of a hand swiping leftward/rightward/upward/downward have similar patterns across different people, making it easier for models to recognize these actions.



An encouraging observation from Fig. 7 is that actions with significant changes in orientation and position, like carrying weight, mopping the floor, walking, running, and turning, yielded favorable classification accuracy with both Wi-Fi and mmWave radar, which was once considered a challenging problem in RF model generalization. We attribute these favorable results to the substantial volume of our training samples. For example. in the action of turning, volunteers mostly tended to turn by 90 degrees or 180 degrees. The most common eight directions for performing actions are illustrated in the right figure. Volunteers will return to their initial orientation in a maximum of four to eight turns. Since each volunteer performs each action 20 times, with the first 14 repetitions used for training, there is a high likelihood that our training samples have covered all directions of the turning action. Even in extreme cases where the direction of an action performed by a volunteer is not represented in the training set. In scene 1, we have a total of 30 volunteers. This means that there are 30×14 = 420 instances of the turning action in the training set, very likely encompassing most directions. For the walking action, similar to the action of turning, we also have 420 walking samples for training. Within the rectangular area enclosed by four Wi-Fi units, these 420 walking samples are likely to cover a wide variety of walking actions. In conclusion, we believe that the path to generalization is through diversity. Expanding the volume of training samples has been proven effective across multiple domains, such as in image classification and ChatGPT, and this belief was the primary motivation behind our data collection effort. Our goal is to provide the RF community with a dataset that has a substantial volume of training samples, thereby accelerating research into the generalization of RF sensing.

5.2 Ablation Study

In the ablation study, we aim to figure out four unsolved questions in the experiments.

(1) How do deep mutual learning and BERT work? Since deep mutual learning and BERT are two plugins in our framework, we alternately apply only one of them to train the network. Table. 8 shows that deep mutual learning is a promising method that enables accuracy improvement over RFID, Wi-Fi, and mmWave radar, especially the mmWave radar. Instead, BERT can largely increase RFID accuracy. We think this may be that when we use L1 distance to compute the BERT semantic loss via Eq. 5, the loss serves as a supervised regression

Mutual	BERT	RFID	Wi-Fi	mmWave
/	/	59.22	87.26	87.97
✓	/	59.98	88.73	90.08
/	✓	63.26	88.97	87.08
✓	✓	64.78	89.73	89.93

Table 8. Deep mutual learning is promising for all modalities, and BERT works well in RFID improvement. BERT can largely increase RFID accuracy.

guidance, together with the supervised cross-entropy loss, provides enhanced guidance for the convergence of the suboptimal RFID.

(2) Which RF modality contributes the most to the improvement of RFID? We cyclically choose two out of three modalities and apply the deep mutual learning strategy to train the network. Adapting mutual learning from three modalities to two modalities is straightforward, i.e., eliminating one of the KL distance losses between modalities in Eq. 1. Table. 9 shows that Wi-Fi can largely promote the RFID accuracy from the baseline 59.22% to 68.16%, while mmWave radar decreases the RFID accuracy to 57.71%. The reason why Wi-Fi contributes the most to RFID accuracy improvement may be their readings are both in 1D temporal information as shown in Fig. 5.

Mutual Modality			Accuracy (%)		
RFID	Wi-Fi	mmWave	RFID	Wi-Fi	mmWave
✓	✓	/	68.16	89.91	/
✓	/	✓	57.71	/	89.77
/	✓	✓	/	89.38	90.46
✓	✓	✓	64.78	89.73	89.93

Table 9. Wi-Fi can impressively promote the RFID accuracy from the baseline 59.22% to 68.16%.

(3) Does feature fusion among different modalities work? When applying deep mutual learning, one of our key considerations is ensuring that the RF devices continue to function even if some of them fail. If leaving this consideration, we can apply ADD fusion at the final layers of the baseline networks of RFID, Wi-Fi, and mmWave radar, and train the networks together. The results, as shown in Table. 10, indicate that simple late fusion is not an effective approach to collaboration between these RF modalities and may result in decreased action recognition performance.

RFID	Wi-Fi	mmWave	Accuracy (%)
✓	✓	/	82.53
/	✓	✓	86.43
✓	/	✓	86.44
✓	✓	✓	85.91

Table 10. Simple fusion is not an effective approach to collaboration between these modalities.

(4) Which mutual learning loss performs better? Deep mutual learning is to extract information from peer networks by aligning learned representation from different modalities, thus we apply three other losses, i.e., maximum mean discrepancy (MMD) distance [6], Hinge distance, and cosine similarity, to train the network. Table. 11 shows that KLD distance performs the best in general, which is also the default loss in [71]. MMD distance is especially good for Wi-Fi.

Mutual Loss	RFID	Wi-Fi	mmWave
KLD	64.78	89.73	89.93
MMD	60.97	91.15	88.54
Hinge	61.66	88.94	88.12
Cosine	59.84	88.85	87.84

Table 11. KL distance performs the best in general. MMD distance is especially good for Wi-Fi.

5.3 Cross-domain Validation

Cross-person. To evaluate cross-person performance, we utilize data from scene 1. The models are trained with all samples from the first 21 subjects and subsequently tested with all samples from the 9 remaining subjects who were not seen during training. Specifically, the training set consists of both training and test samples of the first 21 subjects in scene 1, as described in Sec. 3.5. Conversely, the test set includes both training and test samples of the 9 subjects not included in the training phase, also from scene 1. During the test, we use 0/1/2 test samples of each action of each subject to finetune the trained models. Table 12 (Baseline) shows that mmWave maintains a certain (55%) generalization ability to unseen subjects. Though RFID and Wi-Fi are hard to generalize to unseen subjects without any strategy, 1/2-shot finetuning works better. In comparison, if we have no pre-trained models, but use only 1 or 2 test samples to train a model for action recognition, then accuracy drops as shown in Table 12 (Direct). This shows our dataset helps in the model pre-training. We further perform a leave-one-subject-out assessment, where the models are trained on all samples from 29 subjects in scene 1. For testing, we use all samples from the one remaining subject in scene 1. The results of this evaluation are documented in Table 12. Moreover, DML performs best in all comparisons. These results can serve as baselines for future work.

Cross-scene. Our dataset contains action samples of 39 subjects from 4 different scenes, with 30 subjects in scene 1 and 3 subjects in each of the other 3 scenes. As explained in Sec. 3.5, each subject executes each action 20 times, from which the first 14 repetitions are allocated to the training set, while the subsequent 6 repetitions are designated for the test set. For the case of Table. 13 (DML)-scene 2, we use the training samples from scene 2 for DML training, and evaluate the trained model with the test samples from the same scene. Conversely, for the case of Table. 13 (Finetune)-scene 2, we initially train the model using training samples from scene 1, and subsequently finetune it using 0/1/2 samples (per subject per action) from scene 2. We then test the finetuned model with the remaining 20/19/18 samples (per subject per action) in scene 2. Table. 13 (Finetune) indicates the model's limited transferability to new scenes, particularly with RFID and Wi-Fi data. Yet, finetuning with one or two samples from new scenes significantly enhances accuracy. This implies that XRF55 is valuable for pre-training foundational models for cross-scene action recognition. A similar approach for scenes 3 and 4 also yields results in Table 13, echoing the insights from the evaluation on scene 2.

6 POTENTIAL USEFULNESS

(1) Additional dataset for developing practical applications. Since XRF55 provides large-scale training and testing samples, the industry sector could take it as an additional dataset to develop practical applications, such as elderly care and monitoring, smart home automation, healthcare applications, fitness training, VR/AR.

Method	#train-#test	#-Shot	RFID	Wi-Fi	mmWave
Baseline	21-9	zero	10.00	3.62	55.08
		one	15.77	40.78	59.61
		two	19.97	49.80	62.35
	29-1	zero	12.64	5.55	61.55
		one	19.43	48.13	67.46
		two	26.77	51.01	70.20
Direct	0-9	one	6.54	37.45	19.35
		two	10.43	47.19	27.70
	0-1	one	10.14	48.23	11.01
		two	10.96	50.62	24.44
DML	21-9	zero	9.46	3.29	57.85
		one	18.95	45.65	61.89
		two	21.59	53.21	65.31
	29-1	zero	13.72	7.68	61.41
		one	21.17	49.49	68.43
		two	27.73	51.78	70.99

Table 12. Results of tests on unseen subjects.

Method	Scene	#-Shot	RFID	Wi-Fi	mmWave
DML	scene 2	/	58.38	93.94	78.08
	scene 3	/	88.59	95.66	72.93
	scene 4	/	70.71	96.67	74.04
Finetune	scene 2	zero	1.94	2.52	11.36
		one	6.06	49.83	25.33
		two	9.29	57.51	33.77
	scene 3	zero	2.15	2.14	26.42
		one	18.09	50.85	34.51
		two	34.11	63.30	45.96
	scene 4	zero	2.91	2.03	26.58
		one	9.02	50.94	35.25
		two	14.83	61.41	43.60

Table 13. Results of few-shot tests under cross-scene settings.

(2) Real-Time processing and edge computing. Investigating methods to deploy algorithms on edge devices, like smart home routers or IoT devices, for real-time data processing. For example, XRF55 can be used to train large models, which can then be compressed as per requirements for deployment on edge devices. This approach aims to reduce latency and enhance system responsiveness.

(3) Synthesizing RF data. XRF55 provides paired human video data and RF data, enabling researchers to model the relationship between RF signals and the human body. Through generative learning, one can synthesize RF signals corresponding to given human postures and use these synthesized RF signals as training data to train action recognition models as [63, 70]. This approach significantly reduces the burden of data collection and accelerates the generalization of RF sensing.

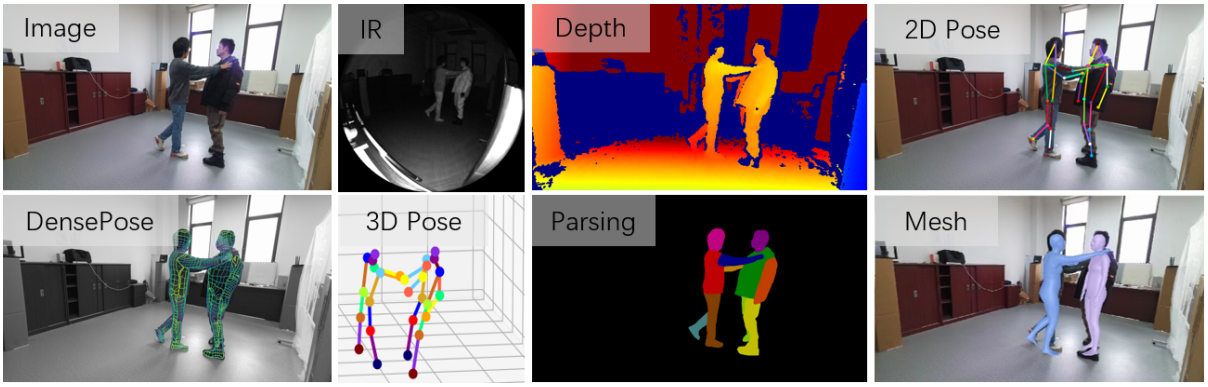


Fig. 8. With synchronized Kinect RGB+D+IR videos, XRF55 also supports fine-grained person perception with RF-only and RF-Vision approaches as [9, 16, 26, 32, 57, 64] have done.

(4) Advanced pre-processing methods. XRF55 provides raw radio frequency data, thus people can take it to propose and validate advanced pre-processing methods on radio frequency signals, enhancing data quality and efficiency in various applications.

(5) Advanced action recognition models. XRF55 can serve as a benchmark for validating the performance of proposed advanced action recognition models for radio frequency signals.

(6) Advanced collaboration methods. XRF55 enables the exploration of advanced collaboration or knowledge distillation methods among Wi-Fi, mmWave radar, and RFID. Additionally, with the synchronized Kinect, XRF55 can be used to investigate advanced collaboration methods between videos and radio frequency signals.

(7) Fine-grained human sensing models. XRF55 supports and facilitates the exploration of pose estimation, human parsing, and mesh reconstruction, as shown in Fig. 8, using either RF-only or RF-Vision solutions, similar to the approaches in previous works [9, 16, 26, 32, 57, 64].

(8) Identification and authentication. The variations in action execution among different individuals, serving as behavioral biometric indicators, can be employed for person identification [56] and person re-identification [7].

(9) Pre-trained radio frequency models. As XRF55 contains the largest number of action classes and action samples, it can be leveraged for pre-training radio frequency models. These pre-trained models can then be applied to other datasets or new environments with few-sample fine-tuning.

(10) Action class incremental learning. In the scenario where a customized action recognition system needs to recognize new actions that have not been previously trained, catastrophic forgetting may occur if the system is trained with only the new-action data and no old-class data. XRF55 can be used to explore class incremental learning approaches specifically designed for radio frequency signals.

7 DATASET AND CODE AVAILABILITY

The dataset and code can be found at the XRF55 website: https://aiotgroup.github.io/XRF55, where we also provide a hardware setup tutorial for those who are interested in taking Wi-Fi, RFID, mmWave radar for human sensing.

In this code repository, *model/resnet1d.py*, *model/resnet1d_rfid.py*, and *model/resnet2d.py* correspond to Wi-Fi, RFID, and mmWave networks, respectively. To begin, please download and unzip the dataset from Kaggle. Next, you can proceed with splitting the dataset and generating the necessary labels using the scripts *split_train_test.py* and *generate_txt.py*. Additionally, you will need to configure the training parameters by modifying the *opts.py* file. To initiate the training process, run the *dml_train.py* script. Finally, evaluate the trained models using *dml_eval.py*.

In the file *XRF_dataset.py*, we established the dataloader, which loads each data instance and retrieves the corresponding *bert_label* from *word2vec/bert_new_sentence_large_uncased.npy*. This label plays a crucial role in the training process as it is used to compute the loss parameter. In the *dml_train* module, we perform joint training using Wi-Fi, RFID, and mmWave radar data. Throughout the training process, we extract the output layer from each model to calculate the loss. This approach enables us to achieve the intended goal of mutual learning.

More details could be found at the project description.

8 CONCLUSION

We have introduced a comprehensive dataset, XRF55, consisting of 23 RFID tags, 9 Wi-Fi links, and one mmWave radar, covering a total of 55 human indoor action classes, with a total of 42,900 valid data samples, amounting to over 59 hours of effective duration. It is the first RF dataset to possess such a scale of action classes and sample quantity with commercial off-the-shelf devices. Our work also marks the pioneering investigation into the mutual learning among multiple RF modalities, to facilitate the deployment of an increasing number of RF devices in indoor environments. Additionally, XRF55 provides large-scale training and testing samples for the industry to develop practical applications, such as elderly care and monitoring, smart home automation, healthcare applications, fitness training, etc. XRF55 also holds significant utility in evaluating advanced action recognition models, enabling advanced radio frequency signals and video collaboration models, supporting fine-grained human sensing tasks (e.g., pose estimation, human parsing, and mesh reconstruction), extracting behavioral biometrics from radio signals, serving as foundation models for pre-training and subsequent fine-tuning in other datasets or new environments, and exploring class incremental learning approaches. XRF55 has an obvious limitation: it is collected in environments where users perform actions without any interference from their surroundings, which may not fully capture the conditions of real-world scenarios where external factors could affect user interactions. Nevertheless, we still envision XRF55 helps in transitioning RF sensing technology from laboratory to everyday life applications.

ACKNOWLEDGEMENTS

This paper was supported by the National Natural Science Foundation of China under grant 62102307, U21A20462, 62372400, 62372365, “Pioneer” and “Leading Goose” R&D Program of Zhejiang under grant No. 2024C03287. Key Research and Development Program of Shaanxi (ProgramNo.2021GXLH-Z-021), and China Postdoctoral Science Foundation under grant 2023T160511 and 2021M692562. We are grateful to anonymous associate editors and reviewers for their valuable comments. We are also grateful to Dr. Yunpeng Song and Dr. Ge Wang for fruitful discussions. We appreciate Dr. Yang Du for providing his office space for our preliminary experiments and equipment debugging. We thank all volunteers for their participation.

REFERENCES

[1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–13.

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

[3] Rami Alazrai, Ali Awad, Alsaify Baha'A, Mohammad Hababeh, and Mohammad I Daoud. 2020. A dataset for Wi-Fi-based human-to-human interaction recognition. *Data in brief* 31 (2020), 105668.

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6836–6846.

[5] Paramvir Bahl and Venkata N Padmanabhan. 2000. RADAR: An in-building RF-based user location and tracking system. In *Proceedings IEEE INFOCOM 2000. Conference on computer communications. Nineteenth annual joint conference of the IEEE computer and communications societies (Cat. No. 00CH37064)*, Vol. 2. Ieee, 775–784.

[6] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22, 14 (2006), e49–e57.

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

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

[9] Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Yingfeng Chen, Bin Fang, Jiming Chen, Yuchi Huo, and Qi Ye. 2022. ImmFusion: Robust mmWave-RGB Fusion for 3D Human Body Reconstruction in All Weather Conditions. *arXiv preprint arXiv:2210.01346* (2022).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. 2020. Large scale holistic video understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 593–610.

[12] Han Ding, Chen Qian, Jinsong Han, Ge Wang, Wei Xi, Kun Zhao, and Jizhong Zhao. 2017. Rfipad: Enabling cost-efficient and device-free in-air handwriting using passive tags. In *2017 IEEE 37Th international conference on distributed computing systems (ICDCS)*. IEEE, 447–457.

[13] Han Ding, Longfei Shangguan, Zheng Yang, Jinsong Han, Zimu Zhou, Panlong Yang, Wei Xi, and Jizhong Zhao. 2015. Femo: A platform for free-weight exercise monitoring with rfids. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 141–154.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[15] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[16] Jiaqi Geng, Dong Huang, and Fernando De la Torre. 2022. DensePose From WiFi. *arXiv preprint arXiv:2301.00250* (2022).

[17] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*. 5842–5850.

[18] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6047–6056.

[19] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 2011. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM computer communication review* 41, 1 (2011), 53–53.

[20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.

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

[22] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[23] Zawar Hussain, Michael Sheng, and Wei Emma Zhang. 2019. Different approaches for human activity recognition: A survey. *arXiv preprint arXiv:1906.05074* (2019).

[24] Md Tamzeed Islam and Shahriar Nirjon. 2020. Wi-fringe: Leveraging text semantics in wifi csi-based device-free named gesture recognition. In *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 35–42.

[25] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th annual international conference on mobile computing and networking*. 289–304.

[26] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D human pose construction using WiFi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[27] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. 2014. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://crcv.ucf.edu/THUMOS14/.

[28] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[29] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. 2019. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8658–8667.

[30] Manikanta Kotaru and Sachin Katti. 2017. Position tracking for virtual reality using commodity WiFi. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 68–78.

[31] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*. IEEE, 2556–2563.
[32] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. 2023. HuPR: A Benchmark for Human Pose Estimation Using Millimeter Wave Radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5715–5724.
[33] Bing Li, Wei Cui, Wei Wang, Le Zhang, Zhenghua Chen, and Min Wu. 2021. Two-stream convolution augmented transformer for human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 286–293.
[34] Chenning Li, Zhichao Cao, and Yunhao Liu. 2021. Deep AI enabled ubiquitous wireless sensing: A survey. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–35.
[35] Diangang Li, Jianquan Liu, Shoji Nishimura, Yuka Hayashi, Jun Suzuki, and Yihong Gong. 2020. Multi-person action recognition in microwave sensors. In *Proceedings of the 28th ACM international conference on multimedia*. 411–420.
[36] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. 2019. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 872–881.
[37] Shujian Liao, Terry Lyons, Weixin Yang, Kevin Schlegel, and Hao Ni. 2021. Logsig-RNN: A novel network for robust and efficient skeleton-based action recognition. *arXiv preprint arXiv:2110.13008* (2021).
[38] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–19.
[39] 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.
[40] 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.
[41] Yongsen Ma, Gang Zhou, and Shuangquan Wang. 2019. WiFi sensing with channel state information: A survey. *ACM Computing Surveys (CSUR)* 52, 3 (2019), 1–36.
[42] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. 2019. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.
[43] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. 2019. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence* 42, 2 (2019), 502–508.
[44] Lionel M Ni, Yunhao Liu, Yiu Cho Lau, and Abhishek P Patil. 2003. LANDMARC: Indoor location sensing using active RFID. In *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003.(PerCom 2003)*. IEEE, 407–415.
[45] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. 2013. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. 27–38.
[46] Zhenyue Qin, Yang Liu, Madhawa Perera, Tom Gedeon, Pan Ji, Dongwoo Kim, and Saeed Anwar. 2022. ANUBIS: Skeleton Action Recognition Dataset, Review, and Benchmark. *arXiv preprint arXiv:2205.02071* (2022).
[47] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2616–2625.
[48] Zhiyao Sheng, Huatao Xu, Qian Zhang, and Dong Wang. 2022. Facilitating Radar-Based Gesture Recognition With Self-Supervised Learning. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 154–162.
[49] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 510–526.
[50] Amarjot Singh, Devendra Patil, and SN Omkar. 2018. Eye in the sky: Real-time drone surveillance system (dss) for violent individuals identification using scatternet hybrid deep learning network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1629–1637.
[51] Ronnie Smith, Yuan Ding, George Goussetis, and Mauro Dragone. 2021. A COTS (UHF) RFID floor for device-free ambient assisted living monitoring. In *Ambient Intelligence–Software and Applications: 11th International Symposium on Ambient Intelligence*. Springer, 127–136.
[52] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01* (2012).
[53] Sheng Tan, Yili Ren, Jie Yang, and Yingying Chen. 2022. Commodity WiFi Sensing in 10 Years: Status, Challenges, and Opportunities. *IEEE Internet of Things Journal* (2022).
[54] Chuyu Wang, Jian Liu, Yingying Chen, Hongbo Liu, Lei Xie, Wei Wang, Bingbing He, and Sanglu Lu. 2018. Multi - Touch in the Air: Device-Free Finger Tracking and Gesture Recognition via COTS RFID. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*. 1691–1699. https://doi.org/10.1109/INFOCOM.2018.8486346

[55] Fei Wang, Jianwei Feng, Yinliang Zhao, Xiaobin Zhang, Shiyuan Zhang, and Jinsong Han. 2019. Joint activity recognition and indoor localization with WiFi fingerprints. *IEEE Access* 7 (2019), 80058–80068.
[56] Fei Wang, Jinsong Han, Feng Lin, and Kui Ren. 2019. Wipin: Operation-free passive person identification using wi-fi signals. In *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6.
[57] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. 2019. Person-in-WiFi: Fine-grained person perception using WiFi. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5452–5461.
[58] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st annual international conference on mobile computing and networking*. 65–76.
[59] Yuheng Wang, Haipeng Liu, Kening Cui, Anfu Zhou, Wensheng Li, and Huadong Ma. 2021. m-activity: Accurate and real-time human activity recognition via millimeter wave radar. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8298–8302.
[60] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. 2014. E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. 617–628.
[61] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.
[62] Yucheng Xie, Ruizhe Jiang, Xiaonan Guo, Yan Wang, Jerry Cheng, and Yingying Chen. 2022. mmFit: Low-Effort Personalized Fitness Monitoring Using Millimeter Wave. In *2022 International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 1–10.
[63] Hongfei Xue, Qiming Cao, Chenglin Miao, Yan Ju, Haochen Hu, Aidong Zhang, and Lu Su. 2023. Towards Generalized mmWave-based Human Pose Estimation through Signal Augmentation. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.
[64] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 269–282.
[65] Jianfei Yang, Xinyan Chen, Han Zou, Dazhuo Wang, Qianwen Xu, and Lihua Xie. 2022. EfficientFi: Toward large-scale lightweight WiFi sensing via CSI compression. *IEEE Internet of Things Journal* 9, 15 (2022), 13086–13095.
[66] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. 2023. MM-Fi: Multi-Modal Non-Intrusive 4D Human Dataset for Versatile Wireless Sensing. *arXiv preprint arXiv:2305.10345* (2023).
[67] Bangpeng Yao and Li Fei-Fei. 2010. Grouplet: A structured image representation for recognizing human and object interactions. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 9–16.
[68] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee. 2017. A survey on behavior recognition using WiFi channel state information. *IEEE Communications Magazine* 55, 10 (2017), 98–104.
[69] 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.
[70] Xiaotong Zhang, Zhenjiang Li, and Jin Zhang. 2022. Synthesized Millimeter-Waves for Human Motion Sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 377–390.
[71] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4320–4328.
[72] Yi Zhang, Zheng Yang, Guidong Zhang, Chenshu Wu, and Li Zhang. 2021. XGest: Enabling Cross-Label gesture recognition with RF signals. *ACM Transactions on Sensor Networks (TOSN)* 17, 4 (2021), 1–23.
[73] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.
[74] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, and Dina Katabi. 2019. Through-wall human mesh recovery using radio signals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10113–10122.
[75] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*. PMLR, 4100–4109.
[76] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. https://doi.org/10.1145/3307334.3326081

A ACTIONS FOR APPLICATIONS

All selected actions support practical applications in indoor scenarios. We list some applications next.

1) Carrying weight. In elderly care or daily health, carrying weight may lead to falls or injuries, thus tracking this action aids in timely assistance.

2) Mopping the floor. In elderly care or rehabilitation, mopping the floor helps in assessing physical functioning, movement capability, and recovery progress in daily life.

3) Cutting something. Tracking cutting action in the kitchen aids in understanding users' cooking habits or enhancing kitchen safety when using a knife.

4) Wearing hat. Wearing hats of elderly individuals or those with special needs may reflect their cognitive status and daily activity capabilities. It can also offer clothing pairing suggestions or fashion tips.

5) Using a phone. Identifying the frequency and duration of smartphone usage can aid in understanding users' behavior patterns and lifestyle habits. Identifying excessive smartphone use can be employed for health tracking and reminding users to take appropriate breaks. It also helps parents with child supervision.

6) Throwing something. In smart homes, tracking this action can be used to automatically adjust the cleaning schedules of robot vacuums or to remind users to clean up. It can also be employed for detecting domestic violence or emotional abnormalities.

7) Putting something on the table. In smart homes, for example, putting a remote control on the table can trigger automatic adjustments like turning on/off devices, and adapting lighting and temperature.

8) Put on clothing. Dressing actions of elderly or physically restricted individuals helps in determining if assistance is needed and notifying family or caregivers when necessary. It can also serve as an indicator to aid remote tracking of their health and lifestyle.

9) Picking. Smart home systems detect picking up action to provide assistance and prevent falls, particularly for the elderly or those with mobility issues, adjust furniture, or offer mechanical aids to reduce strain.

10) Drinking. This helps in tracking water intake frequency and quantity to provide health advice, or to adjust water dispensers based on detected drinking habits.

11) Smoking. This can be used to track and intervene in smoking habits, automatically activate air purifiers for indoor air quality, and enhance fire risk tracking.

12) Eating. In smart homes, tracking the frequency and timing of meals can be used to analyze eating habits, and encourage users to have regular meals, especially for the elderly or disabled.

13) Brushing teeth. This can track and remind users about their tooth brushing frequency and duration, and record brushing habits as part of health data to aid users and dentists in understanding oral health.

14) Blowing dry hair. Detecting hair drying, smart homes can adjust the temperature, track dryer safety to prevent fire and remind users about heat levels for hair and scalp protection.

15) Combing hair. By logging brushing data and analyzing hair brushing habits for personal beauty and health tracking.

16) Shaking hands. After detecting handshakes, smart homes adjust home settings for social ambiance, and remind personal hygiene, like hand washing.

17) Hugging. Similar to 16) shaking hands. Besides, hugging provides insights into users' emotional states.

18) Handing something to someone. In smart homes, recognizing this action can trigger automated responses like adjusting lighting or music and enhancing the ambiance. In particular, detecting the action of handing over a medicine bottle can prompt reminders for medication schedules or alert caregivers, enhancing health management and safety.

19) Kicking someone. In smart homes, automatically detecting this action can quickly alert emergency services of potential violence or conflicts and track homes with children, the elderly, or vulnerable residents to prevent domestic violence or accidents.

20) Hitting someone with something. In smart home, enhance family safety by identifying potential aggressive actions like hitting someone, particularly in homes with children or vulnerable individuals, use this

recognition for behavioral analysis and correction, especially in educating children on proper conduct, and track for indications of family conflicts or psychological health issues, prompting further attention or intervention.

21) Choking someone's neck. Detecting choking alerts emergency services for rapid response to violence, and tracking for potential domestic hazards, especially in homes with children or vulnerable individuals.

22) Pushing someone. Detecting pushing actions triggers quick alerts to emergency services for possible violence or conflicts and helps track homes with children, the elderly, or vulnerable residents to prevent domestic incidents.

23) Body weight squats. Detecting squats enhances workout safety and effectiveness, supports rehabilitation, and logs squat activities in health records to track exercise routines and overall physical health.

24) Tai Chi. Detecting Tai Chi movements guides practice with real-time feedback, aids rehabilitation by tracking movement accuracy, and supports elderly health by improving physical health and flexibility through Tai Chi.

25) Boxing. Detecting boxing movements evaluates performance for skill enhancement and safe training, with real-time feedback and improvement suggestions, and AI-integrated virtual coaching offers personalized fitness plans and guidance.

26) Weightlifting. Recognizing weightlifting movements, smart fitness tools offer performance assessments, immediate safety feedback, and strength and technique analysis for professional coaching, while AI tailors training regimens.

27) Hula hooping. Hula hoop motion detection facilitates effective exercise with instant feedback, targets fitness goals such as fat loss or core strengthening, and aids rehabilitation by evaluating progress, especially in balance and coordination.

28) Jumping rope. Jump rope action assessment boosts skills and training safety with immediate feedback, tracks goals such as stamina or weight management, and analyzes technique and rhythm for customized training advice.

29) Jumping jack. Jumping jack motion recognition enhances skill and efficiency with instant feedback, aids in tracking stamina and weight goals, and provides in-depth technical and rhythm analysis for tailored training guidance.

30) High leg lifting. Intelligent fitness systems assess high leg lifting actions for improved exercise effectiveness and safety with real-time feedback and use motion data for detailed technical analysis and personalized training suggestions.

31) Waving. Waving actions serve as natural user interfaces in smart homes, allowing residents to control devices like lights, TV, or music systems with simple gestures, and in specific scenarios, recognized waving can signal for help in emergencies, especially when someone is unable to speak.

32) Clapping hands. Clapping is utilized in smart homes for gesture-based control of devices like lights, TVs, music systems, and in-home entertainment systems, recognized clapping enhances children's interactive games and educational experiences, such as in music and rhythm learning.

33) Falling on the floor. Fall detection for the elderly or those with mobility impairments triggers alerts to emergency services or family, vital for high-risk scenarios, and in homes with children or vulnerable individuals, it helps track dangers and facilitates prompt responses.

34) Jumping. Jump movement recognition in fitness enhances skills and efficiency with immediate feedback and assists rehabilitation by evaluating recovery and physical abilities, focusing on lower limb function and balance enhancement.

35) Running. Running motion recognition assesses performance for skill and efficiency improvement with real-time feedback, while specialized analysis uses running data for gait, rhythm, and posture evaluation, offering personalized training advice.

36) Sitting down. By auto-adjusting environmental settings like lighting, temperature, or entertainment devices upon detecting sitting down, remind those who sit for extended periods to take breaks for health, and optimize energy use by recognizing room state, automatically turning off lights and electronics.

37) Standing up. Tracking standing actions can auto-adjust settings like lighting and TV on standing up, remind sedentary individuals to move for health, and assist elderly or disabled persons in standing actions for safety.

38) Turning. By automatically adjusting settings, like turning on devices when users face them, and tracking movements of children or the elderly for safety by recognizing turning actions.

39) Walking. By tracking walking actions to ensure the safety of children, the elderly, or those with mobility issues, optimizing energy use by turning off lights and appliances in empty rooms, and analyzing walking patterns like stride and speed for pedestrian identity detection.

40) Stretching. Identify stretching actions to remind sedentary workers to take breaks, analyze user behavior and mood for personalized services, and assess employee stress and workload in offices for environmental optimization.

41) Patting on shoulder. Detecting shoulder patting as a sign of discomfort, prompting stretching or rest, automatically adjusting environmental settings for relaxation, and assisting elderly or mobility-impaired individuals for comfort.

42) Playing Er-Hu. Recognizing playing Er-Hu actions creates interactive music games or simulated instrument experiences, and for long-term players, systems track playing movements to remind musicians to rest and manage wrist or shoulder strain.

43) Playing Ukulele. Detecting the action of playing the ukulele in smart homes can enhance musical learning experiences and provide interactive entertainment, potentially offering real-time feedback for skill improvement and personalized training.

44) Playing drum. Recognizing drumming actions in smart home environments can be applied to interactive music education and entertainment, offering feedback for improving drumming skills and enabling immersive musical experiences.

45) Foot stamping. foot stamping actions are used for accessible interaction control in smart homes, enabling users with limited mobility or hand-use to control devices like lights, and in special scenarios like elderly care, they serve as emergency signals, triggering rapid response protocols.

46) Shaking head. Shaking head is used for accessible interaction in smart homes, allowing users with mobility or hand-use limitations to control devices or respond to services, and smart systems assess user reactions and emotions through this gesture, adjusting services or providing feedback accordingly.

47) Nodding. Nodding in smart homes enables users with limited mobility to control devices and services, assists in assessing user satisfaction for service adjustments, and is used in interactive entertainment systems for game control and content activation.

48) Draw a circle. Offering intuitive gesture control for smart homes and devices, like adjusting volume or navigating menus, provides accessible interaction for those with physical limitations, and in VR and AR, it navigates interfaces, selects objects, and controls virtual environments.

49) Draw a cross. Drawing a cross gesture controls smart devices to cancel or reject actions, offers an easy interaction method for those with physical limitations, and is used in games and entertainment systems for interface control.

50) Pushing. Pushing action detects home entries for security and behavior analysis, offers intuitive gesture control for device navigation, and enables easy interaction for users with mobility or hand-use limitations.

51) Pulling. Single-handed pulling gestures intuitively control smart homes and devices, like zooming screens or activating functions, and offer easy, accessible interactions for users with limited mobility or hand use, such as opening virtual curtains or adjusting sliders.

52) Swiping left. Swipe gestures intuitively navigate smart home interfaces and settings, offer accessible device interaction for users with mobility limitations, and facilitate navigation and interaction in VR and AR environments.

53) Swiping right. Swiping right gestures in smart homes can facilitate intuitive navigation and control of devices, streamlining interactions like selecting options or browsing through menus in a user-friendly manner.

54) Swiping up. Detecting this gesture in smart home settings enables users to interact seamlessly with devices, allowing for efficient control such as scrolling through content or accessing different features in a user-friendly manner.

55) Swiping down. Enhancing user interaction, such as easily navigating menus or closing applications, makes device control more intuitive and efficient.

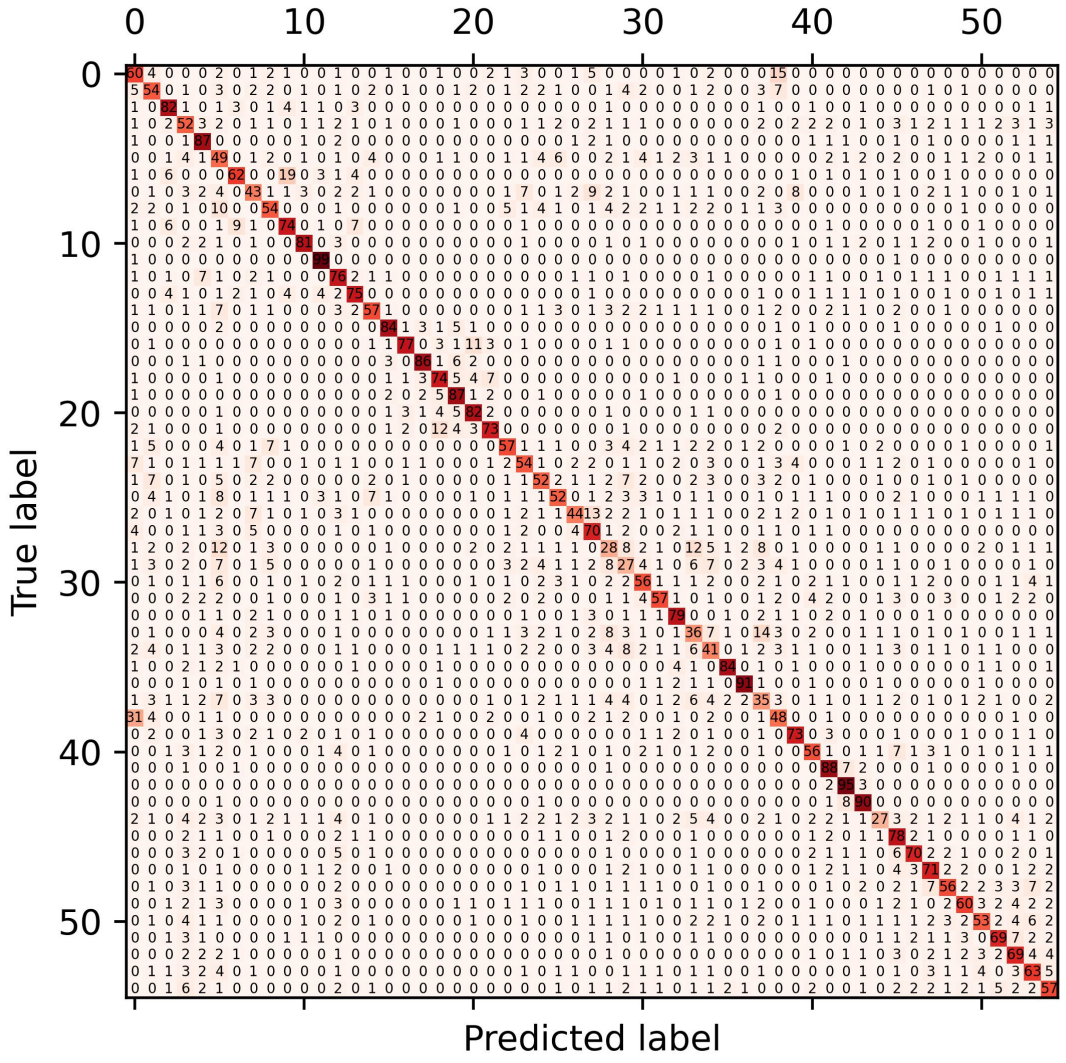


Fig. 9. Confusion matrix of RFID.

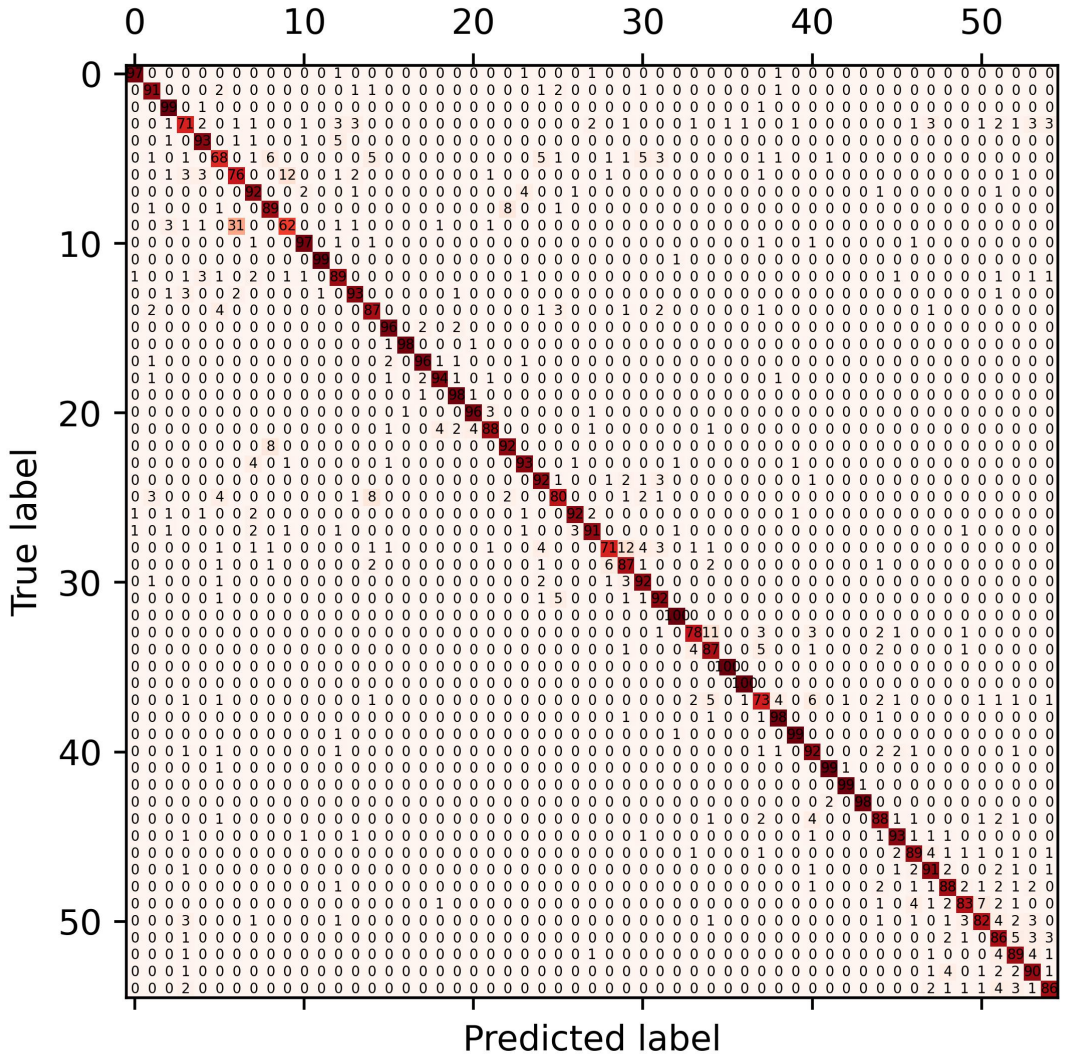


Fig. 10. Confusion matrix of WiFi.

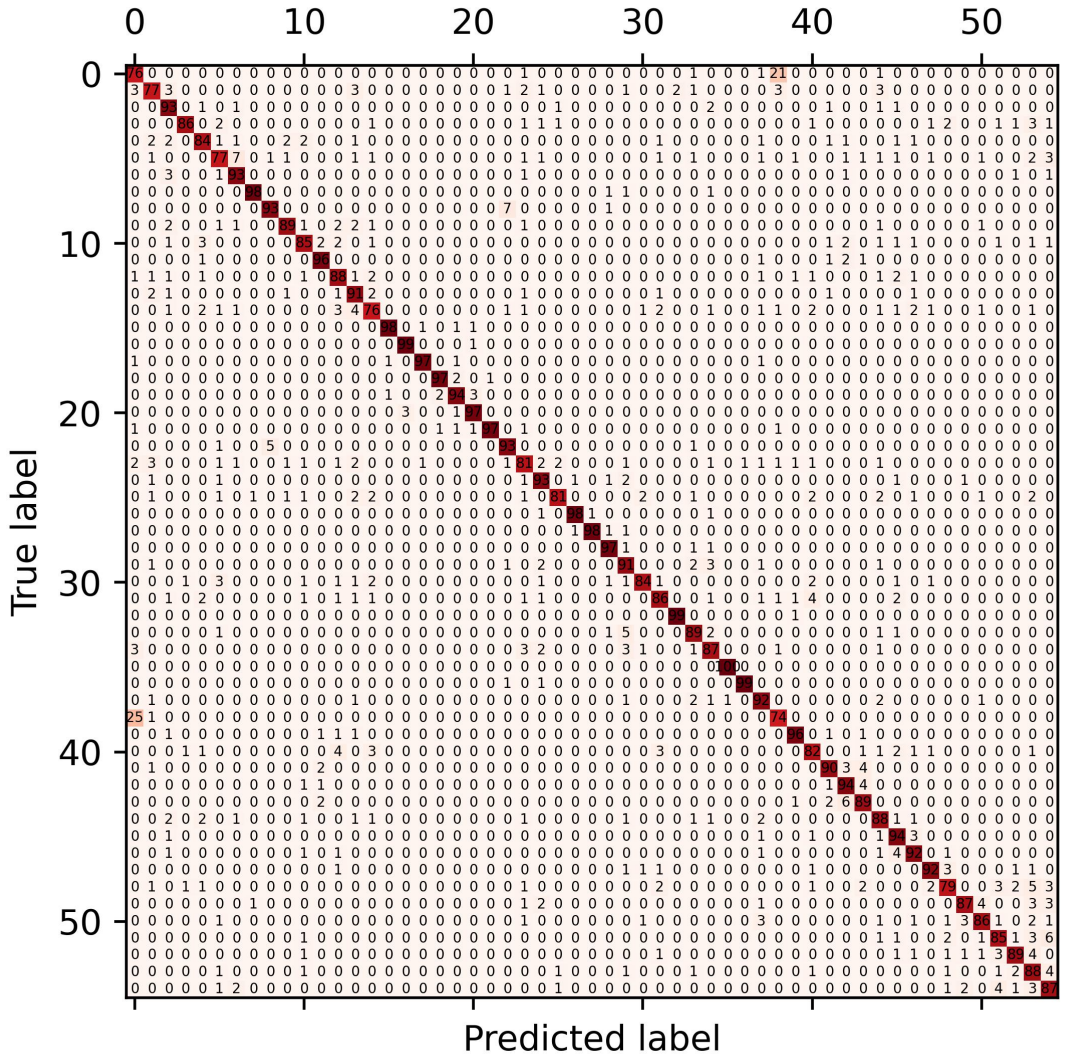


Fig. 11. Confusion matrix of mmWave radar.