Contact-free Human Activity Sensing Using Wireless Signals

by **Xinyu Li**

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Andrew Zhang

School of Electrical and Data Engineering

Faculty of Engineering and IT

University of Technology Sydney

January 25, 2023

Certificate of Authorship / Originality

I, Xinyu Li, declare that this thesis is submitted in fulfillment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with Beijing University of Posts and Telecommunications.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: January 25, 2023

Abstract

Human activity sensing has been widely used in various fields such as security, autonomous driving, and human-computer interaction and has essential research significance and application value. Wireless signal-based activity sensing is based on the fact that human activities impact the wireless signal propagations such as reflection, diffraction and scattering, which provides human activity sensing opportunities through analyzing and mapping the variations to the received signals with a specific activity. Compared with other pathways, human sensing with wireless signals has unique advantages: device-free sensing pattern, robustness to environmental factors (e.g., weather, light, and temperature); the ability to penetrate obstacles; and protecting visual privacy.

WiFi and radar are commonly used wireless sensors for human sensing. In this dissertation, we first propose a channel state information (CSI)-based Doppler speed estimation method, which can provide accurate Doppler estimations with phase offset removal for further human activity analysis. However, using the estimated Doppler frequency estimations alone generally cannot obtain satisfactory performance for human activity recognition. By contrast, radar is a natural sensing sensor and can be utilized to estimate activity-related parameters (e.g., the time-varying range and Doppler frequency information) more easily than WiFi signals, we go a step beyond activity parameter estimation and focus on activity recognition with radar signals. Specifically, the main research problems and contributions of this thesis can be summarized as follows.

First, to remove the carrier frequency offset caused by clock asynchronism and attain accurate Doppler speed estimates, we study Doppler frequency estimation using the cross-antenna signal ratio (CASR) method for scenarios with general movement. We first develop a CSI-ratio expression disclosing more insights, and then propose three algorithms for estimating Doppler frequencies: Mobius Transformation-based, signal difference-based, and periodicity-based. These

algorithms exploit different features of the CSI ratio in terms of Doppler frequencies and can be applied to scenarios involving general and/or irregular movement. Using a publically available WiFi CSI dataset *Widar* 2.0, we then validate the efficiency of the proposed Doppler frequency estimation algorithms.

Second, aiming at enhancing the generalization ability of deep learning (DL) methods to human individual differences and improving the HAR performance on different persons' activities, we present an instance-based transfer learning approach *ITL* for cross-target HAR with radar spectrograms. The proposed *ITL* is composed of three interconnected and necessary parts (*MNet* pretraining, CSDS and ACFT) rather than a collection of three distinct pieces. Experiments demonstrate that the proposed approach is more generalized to the data distribution discrepancy and can scale well to recognize different persons' activities.

Third, we propose a supervised few-shot adversarial domain adaptation (*FS-ADA*) method for radar-based HAR. This method does not require a large number of radar data for training when applied to a new environment. Specifically, we adopt the domain adaptation method to learn a common feature space between a pre-existing radar dataset and the newly acquired training data, and present a multitask generative adversarial training mechanism to optimize *FS-ADA*. Experimental results on two few-shot radar-based HAR tasks show that the proposed *FS-ADA* method is effective for few-shot HAR, and outperforms state-of-the-art methods.

Acknowledgements

The dissertation has been completed with the encouragement and help from many people. First, I would like to express my sincerest and deepest gratitude to my principal supervisor Prof. J. Andrew Zhang, for his valuable guidance, persistent patience, and helpful advice. During my UTS Ph.D. life, he guided me through my research work in every detail and sacrificed his time off to help me delve into technical problems. At the same time, he provided me with a free research environment, respecting and taking care of each student's ideas. His truth-seeking and rigorous research attitude have deeply influenced me, prompting me to make progress and improvement. Furthermore, I would also like to sincerely thank my co-supervisor Dr. Kai Wu, Prof. Xiangjian He, and my previous principal supervisor Prof. Guoqiang Mao for their guidance, help and support. I feel incredibly fortunate to be mentored by them during my Ph.D. candidature at UTS. I would like to thank my supervisor Prof. Xiaojun Jing and Prof. Yuan He from Beijing University of Posts and Telecommunications (BUPT) for their guidance and support. Thank Dr. Yuanhao Cui from BUPT, Prof. Fan Liu and Prof. Weijie Yuan from Southern University of Science and Technology for their help, guidance and encouragement. I am also grateful to have met other members at UTS for their encouragement and support. I will always miss this precious time of hard work with you. Finally, but mostly, I wish to express my deepest gratitude to my beloved parents for backing me up and understanding me silently all the time.

Xinyu Li

January 25, 2023

Sydney, Australia

Publications

The following is my publication list related to this thesis. In some cases, the journal papers contain material overlapping with the conference publications.

Journal Papers

J-1. **X. Li**, Y. He, F. Fioranelli, A. Yarovoy, and Y. Yang, "Human Motion Recognition with Limited Radar Micro-Doppler Signatures," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 6586-6599, Aug. 2021, doi: 10.1109/TGRS.2020.3028223.

J-2. **X. Li**, Y. He, J. A. Zhang and X. Jing, "Supervised Domain Adaptation for Few-Shot Radar-Based Human Activity Recognition," in *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25880-25890, Nov, 2021, doi: 10.1109/JSEN.2021.3117942.

J-3. **X. Li**, Y. He, and X. Jing, "A Survey of Deep Learning-Based Human Activity Recognition in Radar," in *Remote Sensing*, vol. 11, pp. 1068, May 2019, doi: 10.3390/rs11091068.

J-4. **X. Li**, Y. Cui, J. A. Zhang, F. Liu, D. Zhang and L. Hanzo, "Integrated Human Activity Sensing and Communications," in *IEEE Communications Magazine*, early access, doi: 10.1109/MCOM.002.2200391.

J-5. **X. Li**, J. A. Zhang, K. Wu, Y. Cui and X. Jing, "CSI-Ratio-Based Doppler Frequency Estimation in Integrated Sensing and Communications", in *IEEE Sensors Journal*, vol. 22, no. 21, pp. 20886-20895, 1 Nov.1, 2022, doi: 10.1109/JSEN.2022.3208272.

Conference Papers

C-1. **X. Li**, J. A. Zhang, X. Jing and Y. He, "Few-shot Human Activity Recognition with Radar Micro-Doppler Spectrograms", in *2021 CIE International Radar Conference*, pp. 1–2, Dec. 2021.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this dissertation, we first overview the research background of contact-free human activity sensing using wireless signals, and present several main concerns in the wireless signal-based human sensing field. Based on this, we further introduce our contributions from two perspectives: WiFi-based activity parameter estimation and radar-based human activity recognition (HAR), aiming to solve the difficulties in the field and promote the development of wireless human sensing.

1.1 Research Background

Human target sensing has a wide range of applications in many fields, such as security, autonomous driving, human-computer interaction, etc., and therefore has increasingly received more attention. Most of the current human detection and activity sensing research is based on optical sensors, whose performance can be significantly affected by lighting and weather conditions. Meanwhile, human sensing based on acoustic sensors mainly uses low-cost ultrasonic equipment; however, a delay can be incurred when being employed to detect long-distance targets. In contrast, human target sensing with wireless radio frequency (RF) signals has many unique strengths: first, RF signal is robust to environmental factors such as weather, light, temperature, etc., and can be used in a broad range of scenarios for the sensing purpose; second, it is able to penetrate obstacles such as walls, and can detect people behind obstacles. Furthermore, using wireless signals for sensing can protect visual privacy, avoiding privacy breaches. Specifically, different from the optical sensor that perceives the target by capturing

the visual shape of the target, the returned RF signal modulated by the human target contains rich information, such as the human target's speed, orientation, and distance. Last but not least, the RF sensing system is contact-free and does not require the human body to carry any equipment, which introduces no discomfort and thus reduces the human burden.

Human target detection based on wireless signals has tremendous application potential, such as environment sensing for autopilot, health monitoring, survivors rescue in fire or earthquake, and terrorist detection. For instance, in autonomous driving, mainstream autonomous vehicle manufacturers (such as Tesla, Bayerische Motoren Werke, etc.) have begun to widely use microwave radar as an in-vehicle sensor to identify pedestrians and vehicles. In disaster search and rescue, RF-based human detection has both research value and crucial social significance.

WiFi and radar are two commonly used mediums for wireless signal-based sensing. Besides the primary communication purpose, WiFi is widely leveraged for sensing due to its sensitivity to environmental dynamics. The strengths of WiFi-based sensing are as follows. First, WiFi has become pervasive in indoor and outdoor settings with the broad utilization and deployment of wireless communication. Second, WiFi infrastructure is low-cost and easy to be deployed, which makes ubiquitous sensing with WiFi possible. Furthermore, with the Orthogonal Frequency Division Multiplexing technology (OFDM) technique, Channel State Information (CSI) and Received Signal Strength Indicator (RSSI) can be adopted to estimate the properties of the propagation channel and perceive the environment dynamics. The initial purpose of CSI and RSSI is to improve the communication performance based on the communication channel estimations, and has been gradually utilized for environment sensing. When a human target appears in the area illuminated by the transmitted signals, its presence and activity can affect the signal propagation and change the channel state. And the resultant variations of the propagation channel are recorded in the receiving end with CSI/RSSI measurements. In this circumstance, to estimate the human-related parameters such as the human localization, range, orientation, and moving speed and direction, signal processing algorithms are required to quantify the variations of CSI/RSSI measurements, and find out the relationship between the channel state changes and human movements. Based on this insight, a tremendous amount of WiFi-based human sensing approaches have been developed and made significant progress [1]–[3].

Meanwhile, different from WiFi systems that transmit OFDM signals and are mainly used

for communication, radar is a nature sensing tool that mainly uses dedicated signals such as Frequency Modulated Continuous Waves (FMCW) and pulse signals. With the well-designed waveforms, the human target's moving parameters and localization, such as range, angle, radial velocity, and moving direction, can be attained without complex signal processing operations. Therefore, target localization, HAR, tracking, and multiple-targets sensing can be performed with radar, achieving better performance than with the counterpart WiFi. Besides, with the continuous development of electronics and chip technology, miniaturized and portable radars have emerged and are more and more applied to sensing tasks in the civilian field.

1.2 Motivations and Objectives

In this dissertation, we propose a CSI-ratio-based Doppler speed estimation method in Chapter 3, which can provide accurate Doppler frequency estimations, and can be used for recognizing some simple human activities such as walking. However, using the estimated Doppler frequency estimations alone generally cannot obtain satisfactory performance for human activity recognition. By contrast, radar is a natural sensing sensor and can be utilized to estimate activity-related parameters (e.g., the time-varying range and Doppler frequency information) more easily than WiFi signals, we go a step beyond activity parameter estimation and focus on activity recognition with radar signals. Specifically, we present an instance-based transfer learning (TL) approach for cross-target HAR in Chapter 4, and propose a supervised domain adaptation method for few-shot radar-based HAR in Chapter 5, respectively. The motivations and objectives of this thesis can be summarized as follows.

1.2.1 Doppler Speeds Estimation of Moving Human Target with CFO Removal

In WiFi-based sensing, a fundamental issue is to deal with clock asynchronism between transmitters (Tx) and (sensing) receivers (Rx) [4], which are generally geometrically separated. Clock asynchronism causes timing offset (TMO), carrier frequency offset (CFO) and sampling frequency offset (SFO) in the received signals. For sensing, TMO can result in timing ambiguity and hence ranging ambiguity. CFO causes Doppler estimation ambiguity, further degrading the accuracy of Doppler frequency estimation. Furthermore, clock asynchronism also causes time-varying phase shifts across discontinuous CSI measurements, which hinders the coherent

processing of CSI measurements across different timeslots/packets and makes Doppler frequency estimation challenging. Therefore, to obtain accurate Doppler speed estimation of human target using complex signals instead of the power only, removing the phase offsets induced by clock asynchrony is essential.

1.2.2 Cross-Target HAR with Limited Radar Data

Due to the human individual discrepancies, such as the differences in appearances and behaviors, the measurements of the same activity of different persons are generally diverse. When using a trained deep learning (DL) model to recognize the activities from various persons, the HAR performance of this model can be different. Furthermore, when a DL model trained with the activity data of a known person is applied to identify a new person's activity, the performance of this model generally degrades. In this case, to identify the activity of a new person with good performance, a straightforward strategy is to train a specialized model from scratch for each person. However, this solution requires a vast amount of radar data for model optimization, which is difficult and labor-consuming. TL, which utilizes prior knowledge to make a trained model generalize well on new tasks, is one of the potential solutions for cross-target HAR [5]–[7]. However, the existing approaches often suffer from the catastrophic forgetting effect [8], i.e., the tendency of DL models to abruptly forget previously learned tasks after being trained for a new task. As a result, when a DL model is transferred to the HAR task for a new person, its performance in recognizing the previous persons will drop. Therefore, there is an urgent need to deal with the cross-target HAR problem.

1.2.3 Few-shot Radar-based HAR

Due to its powerful studying ability, DL has been widely applied in radar-based HAR tasks. Most of the existing DL solutions for radar-based HAR are trained with a large volume of labeled data in a supervised manner. However, since cleaning and labeling wireless signals are time-consuming and even infeasible, obtaining a large-scale radar dataset is often tricky. To this extent, data scarcity becomes a bottleneck for the emerging radar-based HAR applications. Prompt solutions need to deal with the few-shot learning issue in the radar-based HAR field. Few-shot learning has been utilized for some applications in radar to deal with the issue of insufficient training data. For instance, Wang et al. [9] proposed a few-shot method based on the hybrid inference network for synthetic aperture radar (SAR) automatic target recognition

(ATR). In [10], a convolutional Bi-LSTM network was proposed for SAR target recognition with few training SAR images. On the other hand, few few-shot learning algorithm has been proposed in the radar-based HAR application.

1.3 Approaches and Contributions

This dissertation studies contact-free human activity sensing with WiFi and radar signals. The research approaches and contributions of this thesis are elaborated as follows:

i) We first present three algorithms for Doppler frequency estimation based on the CSI ratio. These algorithms explore different properties of the CSI ratio, including the circle-preserving property of the Mobius transform, the periodicity of the CSI ratio, and the difference (or correlation) between segments of CSI-ratio signals. We describe these algorithms by referring to human tracking applications in this thesis, but they can be easily adapted to other applications. Using a publically available WiFi CSI dataset *Widar* 2.0, we then validate the efficiency of the proposed Doppler frequency estimation algorithms. Experimental results demonstrate that the proposed algorithms can estimate Doppler frequency accurately, outperforming the commonly used approach based on cross-antenna cross-correlation.

ii) We then introduce an instance-based TL method (*ITL*) for cross-target activity recognition. Unlike the existing instance-based TL approaches, we utilize a different similarity metric to compare the similarity between the source data and the target data. Then, a series of source samples are specially selected for every piece of target data. Furthermore, during the fine-tuning process, the selected source samples are assigned diverse importance by re-weighting their training losses. In this way, the source samples with less domain discrepancy can contribute more to HAR in the target domain. Last but not least, we design a deep CNN model especially for radar-based HAR and use it as the backbone of *ITL*. Experiments show that the proposed *ITL* is more generalized to the data distribution discrepancy and can scale well to recognize different persons' activities.

iii) We also propose a supervised few-shot adversarial domain adaptation (*FS-ADA*) method for HAR, where only a few radar training data are collected from a new application scenario and used for model training. We adopt the domain adaptation method to learn a common

feature space between a pre-existing radar dataset and the newly acquired training data. We also design a multi-class discriminator network, which integrates the category classifier and the binary domain discriminator for model training with limited labeled samples. Then, a multitask generative adversarial training mechanism is proposed to optimize *FS-ADA*. Experimental results for two few-shot radar-based HAR tasks show that the proposed *FS-ADA* method is effective and outperforms state-of-the-art methods.

1.4 Thesis Organization

This dissertation mainly focuses on the research of wireless signal-based human activity sensing, from both the theoretical and the technical perspectives. Chapter 2 summarizes current works, including WiFi-based and radar-based human activity sensing, and DL-based HAR with limited radar samples. Chapters 3, 4, and 5 introduce three published works about wireless human sensing. Chapter 6 summarizes the contributions of this dissertation and discusses several possible research directions for future work. This thesis is organized as follows:

Chapter 2: As a literature review chapter, this chapter first presents a survey about wireless human sensing, including general WiFi-based and radar-based human activity sensing methods. Also, works on DL-empowered HAR with limited radar samples are investigated. Finally, we provide a review of human-related sensing in the context of integrated sensing and communication (ISAC).

Chapter 3: This chapter proposes three algorithms for Doppler frequency estimation based on the CSI ratio. These algorithms explore different properties of the CSI ratio, including the circle-preserving property of the Mobius transform, the periodicity of the CSI ratio, and the difference (or correlation) between segments of CSI-ratio signals.

Chapter 4: It presents *ITL*, an instance-based TL algorithm for cross-target activity recognition with radar data. The proposed *ITL* is more generalized to the data distribution differences and can be employed for identifying different persons' activities. Experiments demonstrate that *ITL* has good performance for recognizing the activities of diverse persons even with limited radar data, outperforming several state-of-the-art HAR methods.

Chapter 5: This chapter presents *FS-ADA* for few-shot radar-based HAR. This method does not require much radar data for training when applied to a new environment. We provide

extensive experiments to verify the performance of our proposed method. The results show that *FS-ADA* outperforms the state-of-the-art benchmarks on two few-shot learning tasks.

Chapter 6 : A summary of this thesis is given in this chapter. Several research directions for future work on wireless sensing are also discussed.

Chapter 2

Literature Review

This chapter is devoted to reviewing the related works on contact-free human activity sensing using wireless signals, including WiFi-based and radar-based activity sensing methods, and the DL-based HAR approaches with limited radar training samples. Furthermore, we provide a review of integrating human activity sensing with communications (IHASC). Based on geographical deployments, we categorize current IHASC into three classical configurations, namely, monostatic, bistatic and distributed deployments, and discuss their properties, critical research problems and solutions.

2.1 WiFi-based Human Activity Sensing

The ubiquitous deployment and wide coverage have enabled WiFi to provide both network communication and the ability of sensing surrounding environments. When transmitted to the physical space, the WiFi signals will interact with the surrounding objects and experience reflection, scattering, absorption, polarization and other multi-path effects. Then, a rich set of information about the environment is contained in the received signals, such as the objects' location and moving status. Furthermore, since the off-the-shelf commodity WiFi is low cost and can be used for sensing without any additional modification, WiFi-based sensing is promising to be deployed and applied at scale.

RSSI and CSI are two commonly used measurements for WiFi-based sensing. Compared with RSSI representing the aggregative power of the entire signal bandwidth, CSI provides the fine-grained channel responses of different subcarriers, and has been applied more widely due to

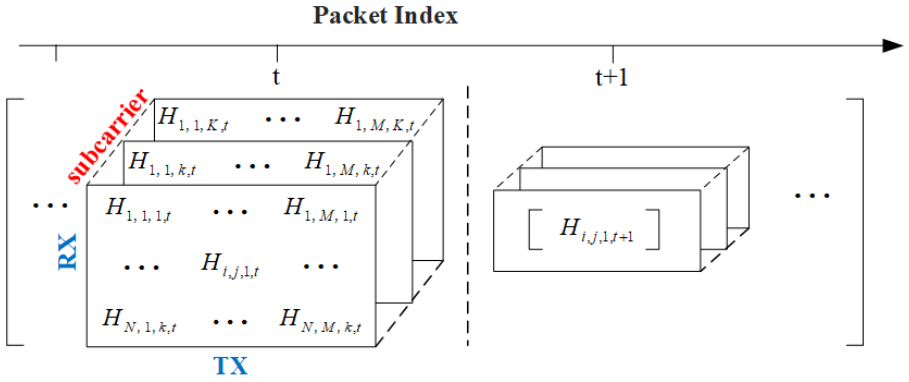


Figure 2.1: A time series of 3D CSI measurements from a MIMO-OFDM WiFi system. Adopted from [11].

its better sensing performance. In this section, we review the state-of-the-art WiFi CSI-based HAR approaches and the phase offset removal issue in bi-static WiFi systems.

2.1.1 Human Activity Recognition with WiFi CSI

WiFi CSI is a metric that depicts the channel properties between the transmitter and the receiver along multiple paths. It was initially introduced to estimate the quality of wireless channels for effective and reliable transmission of communication data and recently has been utilized for wireless sensing. Furthermore, with the OFDM technique, CSI from each transmitter-to-receiver link at each carrier frequency can be obtained, providing sufficient frequency information for sensing. For each subcarrier, the WiFi channel in the frequency domain can be modeled by

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{N}, \tag{2.1}$$

where X is the pre-defined transmitted signal, Y is the received signal, and N is the noise. With X and Y, the CSI matrix H can be estimated at the receiving end after received signal processing such as deinterleaving, demapping and demodulation. Then, for a MIMO-OFDM WiFi system with M transmitting antennas, N receiving antennas, and K subcarriers, the CSI measurements can be represented as a 3-dimensional (3D) matrix of complex values, as shown in Figure 2.1.

It can be seen from Figure 2.1, a time series of CSI measurements characterize MIMO channel variations in time (packet), frequency (subcarrier) and spatial (antenna) domains. Information

from different domains can be extracted from the time-varying CSI matrix and integrated for diverse sensing purposes, e.g., human target detection, activity recognition, and human localization.

There are two categories of CSI-based human sensing approaches, i.e., learning-based and model-based techniques. The learning-based method tries to learn the mapping between the input data and the output label by using a tremendous amount of training samples for model optimization. As a typical and effective learning-based method, DL has been widely used for RF-based human sensing, and has made significant progress. The phase and amplitude of CSI, whose variations can convey the changing patterns of human movements, are the commonly used features in learning-based human sensing. For instance, Chen et al. [1] proposed an attention-based bi-directional long short-term memory (BLSTM) model with WiFi CSI data for the HAR purpose. Alazrai et al. [2] transformed raw WiFi CSI data into a series of 3D "time-frequency- spatial" images and proposed a CNN model to recognize human-to-human interactions. Besides, Shi et al. [12] developed a DL method for HAR with only one-shot training sample. However, DL method requires a considerable amount of training samples to learn the data distribution. And the performance of deep learning approaches generally improves with the labeled training samples increasing, which leads to the burden of data segmentation and annotation. Furthermore, it is difficult for DL to characterize the physical characteristics of echo signals with the data-driven learning pattern. As a result, the reliability and interpretability of DL models may be poor.

Compared with learning-based approaches, the model-based approaches can mathematically model the relationship between CSI dynamics and human movements, and can achieve better performance when used for fine-grained applications [2]. Furthermore, model-based methods are more interpretable and reliable with solid theoretical derivation. In modeling-based approaches, human-related parameters, e.g., range, angle and moving speed, are first estimated with reflected signals and then used for human activity analysis. A series of CSI physical models have been proposed, such as the angle-of-arrival (AoA) model, the Fresnel model, and the CSI-ratio model. Specifically, the AoA model [13] inferred human activities by estimating the angle of incident signals at the receiving node and is generally used for human target localization and tracking tasks. Meanwhile, the distance of the human target can be estimated with the time-of-flight (ToF) of transmitted signals. For instance, SpotFi [14] proposed a ToF sanitization algorithm and used the estimated ToF for likelihood estimates. However, such an estimation

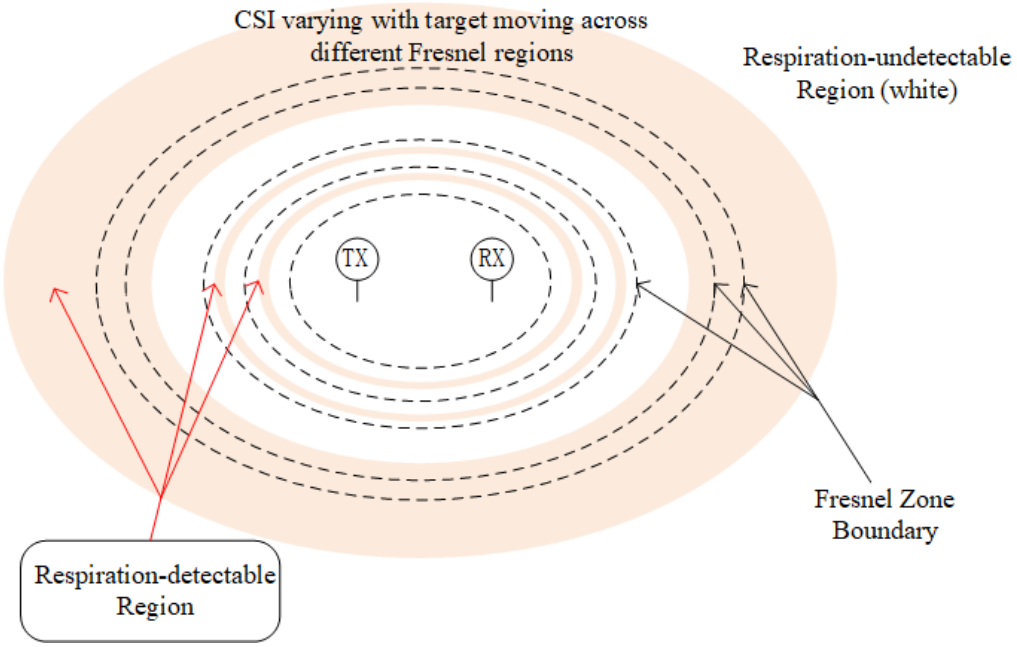


Figure 2.2: Multi-subcarrier Fresnel zones for respiration detection. Adopted from [15].

can not be used for determining range due to the lack of precise synchronization across nodes or access points (APs). Besides, Zhang et al. [15] proposed the Fresnel model (as shown in Figure 2.2), which divides the space between and around the transceiver into several concentric prolate ellipsoidal regions, i.e., Fresnel regions. In this case, the amplitude and phase of CSI data can be correlated with human movements. Specifically, once a human target moves across different Fresnel regions, the amplitude and phase of CSI will change. Based on Fresnel regions, the Fresnel penetration model [16] mathematically correlates the Fresnel phase differences between two different subcarriers with the target's location. Meanwhile, the CSI-ratio model [3] was proposed to quantify the relationship between human movements and the phase variations of CSI measurements. Specifically, the random carrier frequency offsets in the received CSI measurements can be removed by calculating the CSI ratio of the two receiving antennas. Then, based on the Mobius transform (i.e., the phase shift of CSI caused by human movement can be approximately equal to the phase shift of CSI ratio), the moving distance of human targets can be computed. As illustrated in Figure 2.3, three gestures can be distinguished according to the radian change of CSI, which is also the radian change of CSI-ratio caused by human movement.

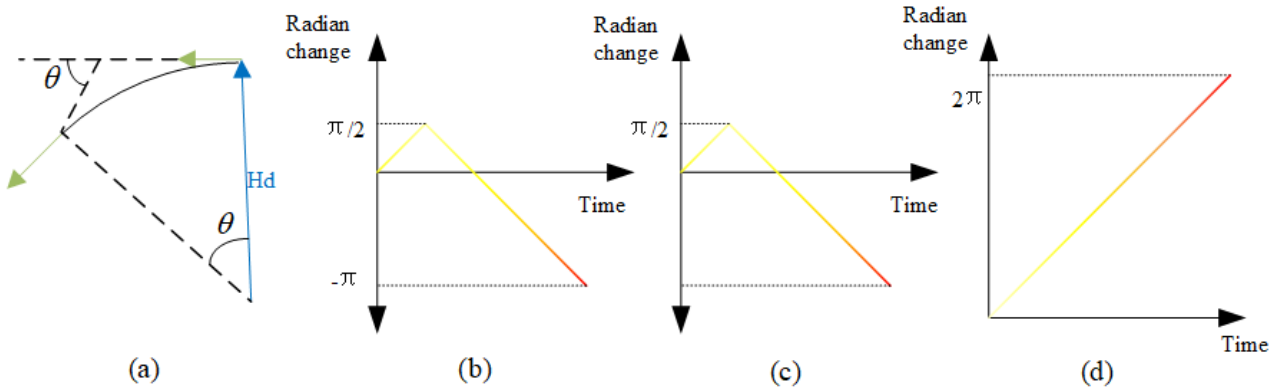


Figure 2.3: Radian changes of the arc caused by three different gestures. (a) Radian change can be estimated as the variation of arc tangent angle. (b)–(d) Radian changes caused by three different gestures. Adopted from [17].

2.1.2 Phase Offset Removal in Bi-static WiFi Systems

In commodity communication systems, e.g., Wi-Fi systems, unlike in the bistatic radar, the receivers are not tightly synchronized with the transmitter regarding carrier frequency and time, resulting in extra phase offsets in received CSI measurements. In this case, the frequency-domain CSI measurement between the ith transmitting antenna and the jth receiving antenna at the kth subcarrier is given by [4]

$$H_{i,j,k,t} = e^{j\phi_t} \sum_{l=1}^{L} b_l e^{-j2\pi(\tau_l+\tau_{o,t})nf_0} e^{j2\pi(f_{D,l}+f_{o,t}tT_s)} e^{ju_{l,p,q}}, \tag{2.2}$$

where f_0 is the subcarrier bandwidth, n is the index of the nth subcarrier, $e^{j\phi_t}$ is the random phase shift changing slowly over time t, including SFO and packet detection delay (PDD), $\tau_{o,t}$ is the TMO, $f_{o,t}$ is the CFO, $u_{l,p,q}$ is the angle-related term, T_s is the OFDM block period, t_l, $f_{D,l}$ and b_l are the signal propagation delay (i.e., ToF), the Doppler frequency, and the amplitude of the lth path, respectively, and L is the total number of paths. TMO can cause timing ambiguity and degrade the performance of range estimation. CFO leads to Doppler estimation ambiguity and further impacts the radial velocity estimation. SFO between every WiFi transmitter-receiver pair can result in additive noise for time-of-flight (ToF) estimation across packets.

Therefore, due to the time-variation characteristic of the unknown variables $\{\phi_t,\ \tau_{o,t},\ f_{o,t}\}$ caused by clock asynchronization, explicitly or implicitly estimating $\{\tau_l,\ f_{D,l}\}$ for accurate

sensing becomes challenging.

Several solutions have been proposed to compensate for the phase offsets at the receiving node. For a single receiver node with multiple antennas, one technique is to construct a reference signal from the line-of-sight (LOS) path. Then, the TMO in other reflected echoes can be eliminated by measuring the time-difference of arrival (TDOA) between the reference signal and the echoes. However, this technique is sensitive to the quality of the constructed reference signal. The other commonly used strategy exploits the fact that the clock offsets across multiple antennas at a receiving node are the same due to the shared oscillator clock. Two classes of methods have been developed based on this fact: Cross antenna cross-correlation (CACC) and cross-antenna signal (or CSI) ratio (CASR).

Specifically, the CACC approach [18], [19] removes the random phase offsets by computing the cross-product between signals from multiple receiving antennas. However, the conjugate multiplication operation introduces additional terms. To remove the extra terms and estimate $f_{D,l}$, it is widely assumed that there exists a dominating line-of-sight (LOS) with a much larger magnitude than non-line-of-sight (NLOS) paths. In this case, the results of the conjugate multiplication operation can be divided into three parts: the product of static paths of two antennas, the product of the dynamic paths, and the product of the static paths of one antenna and the dynamic paths of another antenna. Since the Doppler shift information is contained in the third part, the first two parts need to be removed with a bandpass filter (BPF). Furthermore, image components in the third part could degrade the performance of Doppler shift estimation and needs to be suppressed. In this case, a power adjustment strategy [18] was proposed to increase the power of the term containing the correct Doppler velocity information and make Doppler frequencies identified in the spectrogram. At the same time, other alternative schemes such as mirrored multiple signal classification (MUSIC) algorithm [19] were also proposed.

In contrast, CASR removes the random offsets of multiple receiving antennas via the ratio between CSI signals. Compared with the CACC method, CASR produces an expression where only one term in the denominator contains the parameters to be estimated. Experiments demonstrated that the CASR method could effectively cancel out the noise and improve the SINR of received sensing signals, enlarging the sensing ranges [20]. Based on CASR, a human respiration monitoring method was proposed [3], which quantifies the relationship between the phase variations of CSI ratio and the chest movement. Furthermore, a respiration monitoring

method for multiple persons was proposed [21]. In this method, multi-person respiration sensing was modeled as a blind source separation (BSS) problem when the reflected signals of multiple persons are linearly mixed. Then, the independent component analysis (ICA) algorithm was adopted to separate the mixed signals for further single-person respiration monitoring. However, although the potential of CASR has been well demonstrated in respiration pattern detection and straight-line movement, it has rarely been studied for more complicated scenarios involving irregular and fast movement.

2.2 Radar-based Human Activity Sensing

2.2.1 Radar Micro-Doppler Effect

The radar micro-Doppler (MD) effect was first proposed by Victor C. Chen in 2000 [22]. When a non-rigid target moves, in addition to the main Doppler frequency shift caused by the movement of the target's backbone, the small movements (e.g., vibration and rotation) of the target's other parts will also affect the frequency of the transmitted signals. The generated frequency side lobes next to the Doppler frequency are the MD frequencies. MD effect shows the moving characteristics of the target and is an important basis for target analysis and identification. The mathematical description of the MD effect is as follows.

Suppose there is a mono-static pulsed radar with a centre frequency f_0. When a target is moving in front of the radar with at speed v. Then, the frequency f_r of the received signals is denoted as

$$f_r = f_0(1 + 2v_r/c), \tag{2.3}$$

where c is the speed of the light, $v_r \triangleq v\sin(\theta)$ is the radial component of v, and θ is the angle between the moving direction of the target and the radial direction of the radar.

Then, the MD frequency shift is given by

$$f_D = f_r - f_0 = f_0(2v_r/c). \tag{2.4}$$

It can be seen from Equation2.4 that the (micro) Doppler frequency shift is proportional to

the radial velocity of the corresponding part of the moving target. When a person is moving, the torso and limbs of the non-rigid human body generally have different moving speeds, which will modulate transmitted signals on the frequency domain and produce main Doppler and MD frequency components.

It is assumed that the human target can be divided into K parts, and each part is regarded as a point target. Then, the echo signal s_h corresponding to the whole human body is the sum of the echo signals of the K point targets, which can be denoted as

$$s_h(t) = \sum_{i=1}^{K} a_{t,i} rect(\frac{\hat{t} - t_{d,i}}{\tau} e^{j(-2\pi f_0(t-t_{d,i}) + \pi\gamma(\hat{t}-t_{d,i})^2)}), \tag{2.5}$$

where a_t, i is the amplitude of the reflected signal, τ is the pulse width, γ is the slope of frequency modulation, and $\hat{t}$ is the fast time. Additionally, the total time for signalling $t = T(n-1) + \hat{t}$, where T is the pulse repetition interval (PRI), and n is the number of the transmitted pulse signal, i.e., the slow time. $t_{d,i}$ is the time delay (i.e., the round-trip time required for the radar signal to be sent to the target and reflected back to the receiving antenna), which is given by

$$t_{d,i} = \frac{2(R_i - v_i t)}{c} \tag{2.6}$$

where R_i is the range of the ith part of the human target, v_i is the radial velocity of the ith body component relative to the radar.

Meanwhile, the signal amplitude $a_{t,i}$ can be formulated as

$$a_{t,i} = \frac{G\lambda\sqrt{P_i\sigma_i}}{(4\pi)^{3/2} R_i^2 \sqrt{L_s}\sqrt{L_a}}, \tag{2.7}$$

where G is the antenna gain, λ is the wavelength, P_t is the signal transmission power, σ_i is the radar cross section (RCS) of the ith body component, and L_s and L_a are the system loss and the atmospheric loss, respectively.

From Equation 2.5, it can be seen that radar echo signal is a function of fast time and slow time, and can be represented as two-dimensional (2D) data form, as illustrated in Figure 2.4.

To extract the time-varying Doppler frequency components in the received radar signals, joint time-frequency transform (JTFT), e.g., short-time Fourier transform (STFT), is generally em-

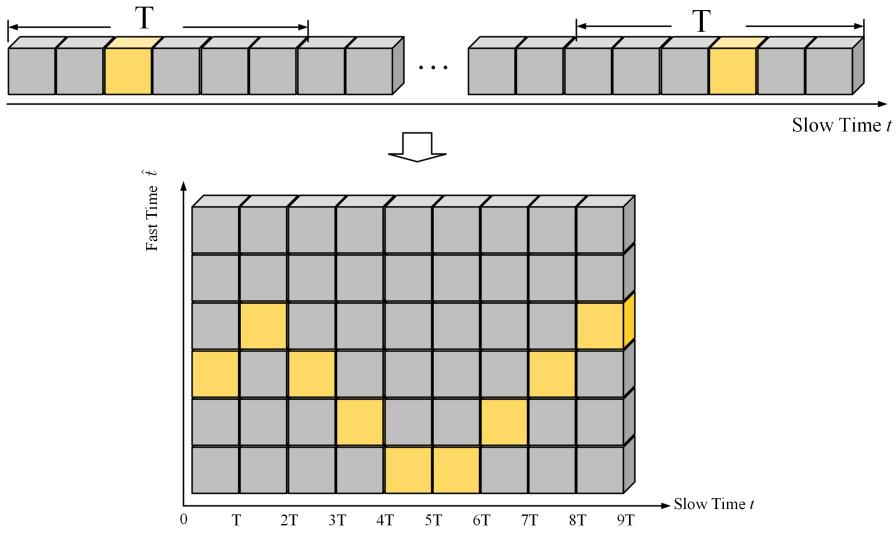


Figure 2.4: 2D radar raw data representation.

ployed to transform the 2D "slow time - fast time" raw data into a "slow time-Doppler frequency" radar map. The mathematical description of STFT is as follows.

$$STFT(i,K) = \sum_{n=0}^{N-1} x_i(n) e^{-j2\pi(nK/N)}, K = 0, 1, ..., N-1, \tag{2.8}$$

where $x_i(n)$ is a sliding window of length N. The ith window $x_i(n)$ is defined as

$$x_i(n) = \hat{S}_R(n + i(N/2))w(n), \tag{2.9}$$

where $w(n)$ is the weighting function.

Based on the above analysis, it can be inferred that the Doppler frequency resolution Δf after STFT can be approximated as the reciprocal of the window duration T_w, namely,

$$\Delta f = \frac{1}{T_w} = \frac{f_s}{N}, \tag{2.10}$$

where f_s is the sampling frequency. From Equation 2.10, we can find that the radar operating at higher frequencies produces wider MD bandwidths. Thus, the Doppler frequency corresponding to the smaller movement of the target is more significant and easier to detect in the time-frequency data. Figure 2.6 shows the radar time-Doppler frequency maps from the radars with different centre frequencies when the human target is running. It can be seen that the time-frequency maps all show the prominent periodic characteristics of the "running" movement.

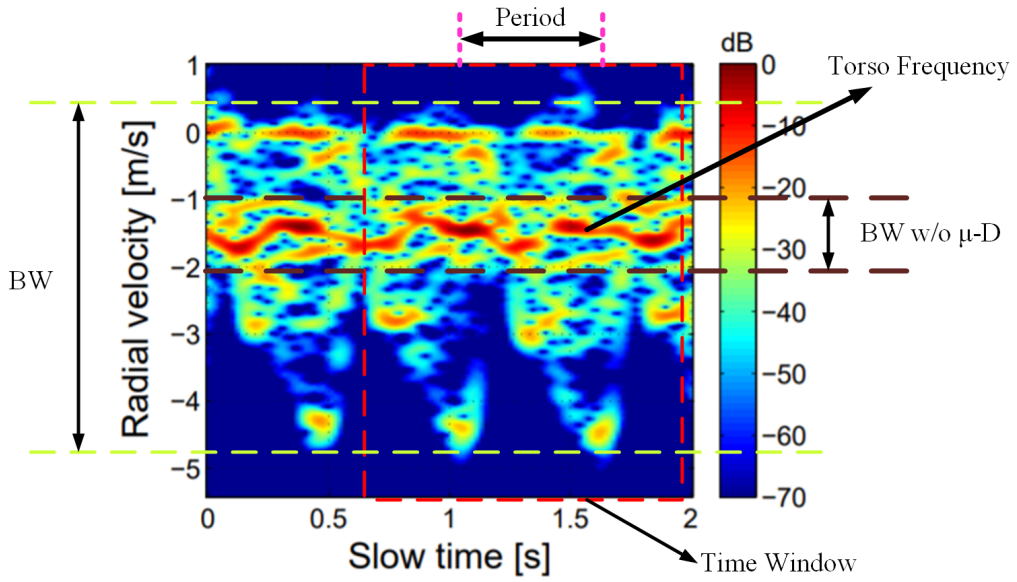
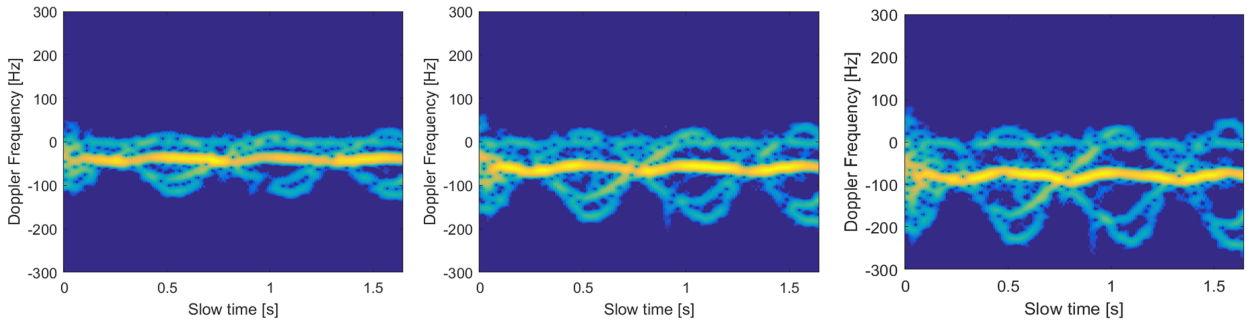


Figure 2.5: Human activity recognition with manually extracted features and an SVM model [23].

Specifically, the corresponding radar echoes are the strongest due to the large RCS of the body torso. Meanwhile, the MD frequency components corresponding to the movement of the limbs are located on both sides of the main Doppler frequency and change periodically.

Due to the activity-unique characteristic, the time-varying Doppler and MD frequencies have been widely utilized for radar-based human activity sensing. In addition to frequency information, time delay (i.e., range) is also adopted alone or in combination with Doppler frequency to perform diverse HAR tasks. Furthermore, with the multiple-input-multiple-output (MIMO) antenna system, angle information of the human target can also be attained with some angle estimation approaches, such as the MUSIC algorithm.

There are mainly two categories of radar-based HAR, i.e., traditional machine learning (ML)-based and DL-based HAR. Traditional ML techniques are based on rigorous theoretical derivation, so they are highly interpretable. Compared to DL models, the complexity of traditional ML tends to be lower, resulting in less computation and faster models. Support vector machine (SVM), Dynamic Time Warping (DTW) and random forest (RF) are commonly used ML models for radar-based HAR. For instance, as shown in Figure 2.5, Kim et al. [23] manually extracted activity-related features from radar time-frequency maps, such as activity period, Doppler frequency bandwidth, and Doppler frequency of the human torso. Then, a decision tree structure consisting of 6 SVMs was constructed to classify 12 human activities.



(a) $f_0 = 4$ GHz (b) $f_0 = 6$ GHz (c) $f_0 = 8$ GHz

Figure 2.6: The time-Doppler frequency maps from the radars with different center frequencies.

Although traditional ML algorithms have been widely used for radar-based HAR, there are still some defects in such algorithms, which hinder the further improvement of the robustness and generalization. First, traditional ML-based classification methods rely on heuristic and manual feature extraction, which is highly dependent on human experience and domain knowledge; secondly, manual features usually refer to some low-level statistical information, including mean, variance, frequency, amplitude, etc., which have weak transferability and generalization. As a result, when a trained model is applied to a new scene, the performance of ML models usually degrades. In contrast, DL models can automatically learn human target information from radar data through a hierarchical structure. Also, the automatic feature extraction process does not require specialized knowledge and human intervention. In addition, with the emergence of the Graphic Processing Unit (GPU), deep learning algorithms can fully use massive data and realize fast data processing and computation based on parallel computing technology.

The following subsection reviews the literature on radar-based HAR with DL methods.

2.2.2 Radar-based HAR with Deep Learning

Radar echo signals contain sufficient information about the targets in the environment, such as range and Doppler frequencies. How to design DL-based signal processing algorithms to extract the target-related information from the received echoes is an essential topic in radar-based human target sensing. In this section, we describe DL approaches for radar-based HAR according to the dimension of radar returns. Table 2.1 lists all the surveyed work in this section.

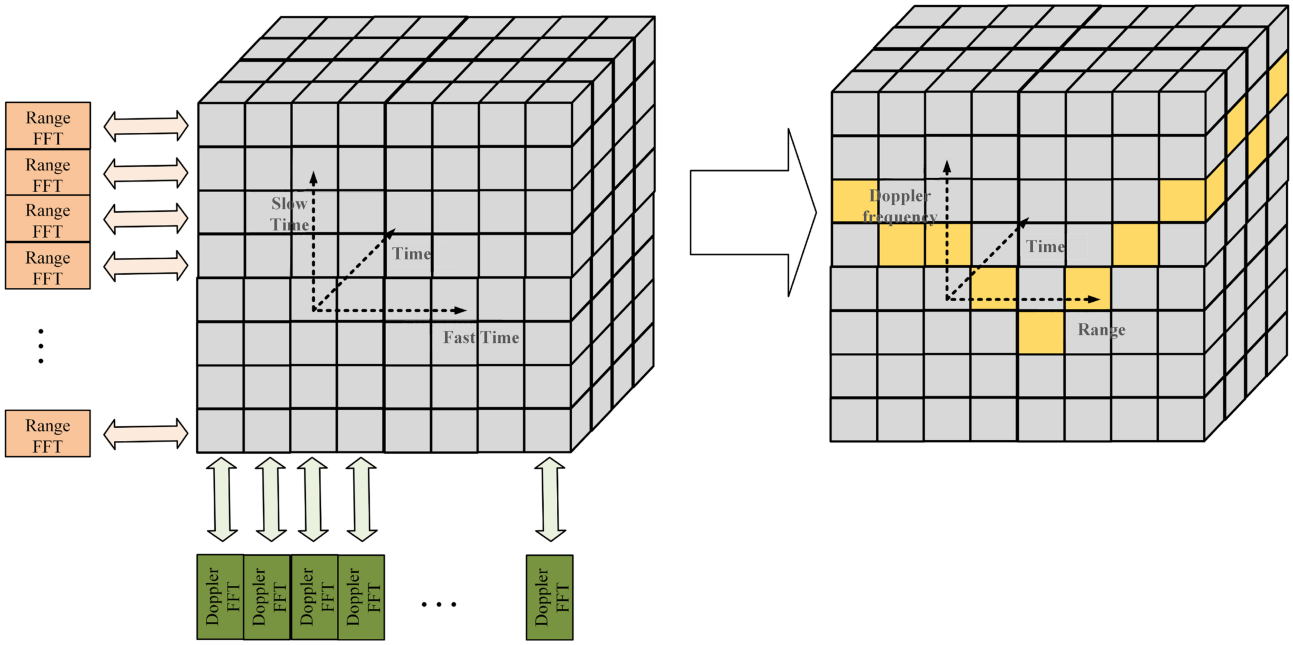


Figure 2.7: Illustration of range-Doppler processing.

Table 2.1: Review on DL-based human activity recognition methods with radar.

Echo Form		**Literature**	**Radar Type**	**Frequency**	**Deep Model**
3D	TRD maps	[24], [25]	FMCW radar	60 GHz	CNN + LSTM
2D	TD maps	[26]	CW radar	4 GHz	CNN
		[27]	CW radar	24 GHz	CNN
		[28]	CW radar	8 GHz	LSTM
		[29]	CW radar	6 GHz	SAE
		[30]	CW radar	4 GHz	CAE
		[31]	Doppler radar	24 GHz	CNN
		[32]	Doppler radar	25 GHz	LSTM
		[33]	Doppler radar	5.8 GHz	CNN
		[34]	UWB radar	4 GHz	CNN
		[35]	FMCW radar	24 GHz	CNN
	TR maps	[36]	UWB radar	3.9GHz	CNN
		[37]	FMCW radar	24 GHz	3D CNN + LSTM
	RD maps	[38]	FMCW radar	24 GHz	3D CNN
	TD, TR maps	[39]	FMCW radar	25 GHz	SAE
	TD, TR, RD maps	[40]	FMCW radar	24 GHz	SAE

Radar signals are transformed into 3D time-range-Doppler data cube by range-Doppler (RD) processing [41], which is illustrated in Figure 2.7. In this way, multiple components of a target are resolved not only in range but also in Doppler frequency. The 3D RD 'video' describes the slow-time evolution of the target's activity, as shown in Figure 2.8a. Radar signals can also be represented in 2D, namely time-Doppler map (Figure 2.8b), time-range map (Figure 2.8c) and rang-Doppler map (Figure 2.8d). In order to make full use of the information in echoes, DL methods should be designed more targeted for different forms of echoes.

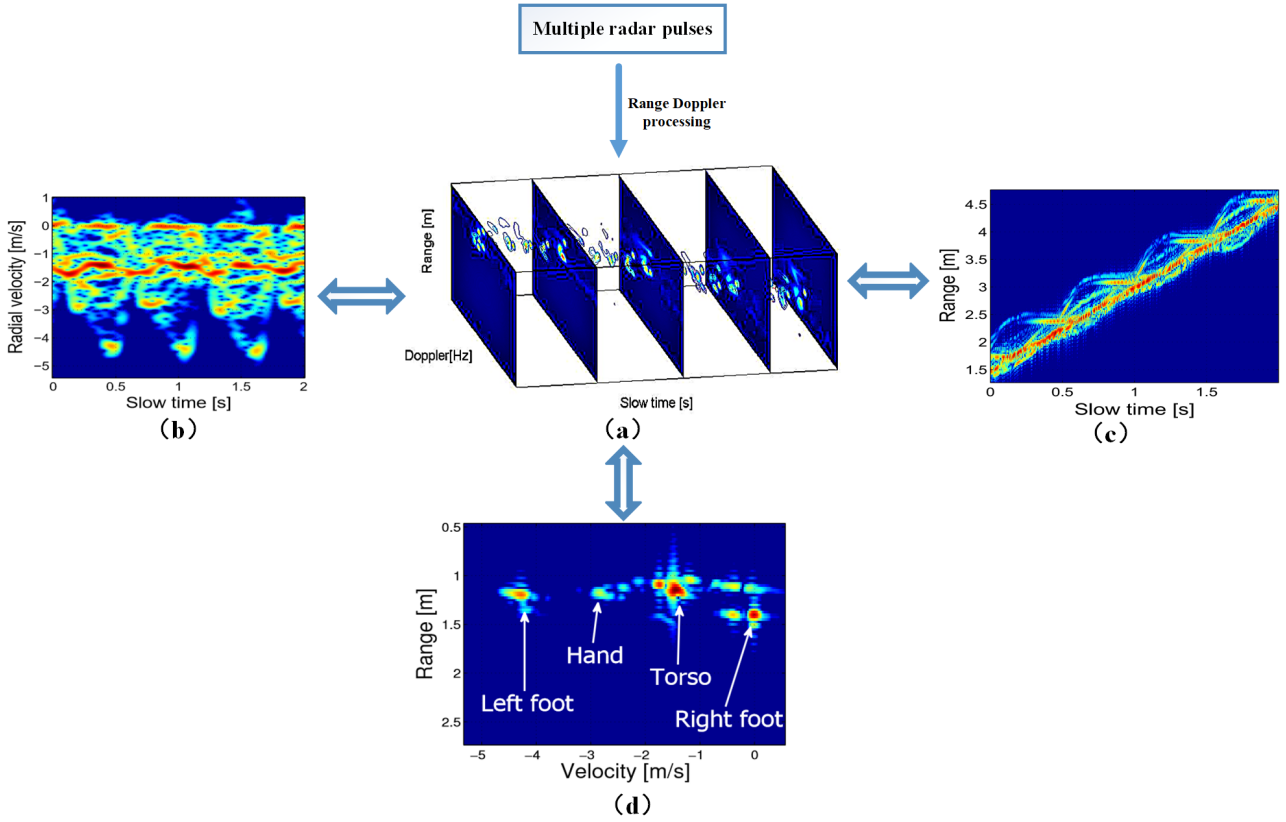


Figure 2.8: 1D, 2D and 3D radar echoes: (a) 3D time–range–Doppler data cube, (b) 2D time–Doppler map, (c) 2D time–range map, (d) 2D range–Doppler map.

- **Deep Learning Approaches in 3D Radar Echo**

Range-Doppler frames reveal moving properties, as well as Doppler properties of targets [42]. Consisting of N time-sampled 2D range–Doppler frames, the 3D RD video sequence conveys both spatial and temporal characteristics of human activities. Range and Doppler information consists in every RD frame while time information exists between frames. Compared with 1D and 2D echoes, the joint time–range–Doppler echoes contain almost all the activity information that radar receives. Models that can extract both temporal and spatial information are required.

Since it is challenging to design features manually from 3D echoes, DL methods are more feasible and preferable for 3D echo-based HAR, thanks to its capability of automatically extracting deep features. Furthermore, the advent of GPU makes it possible for DL models to process 3D data quickly and efficiently. Although few DL algorithms have been proposed for 3D radar echoes till now, DL approaches on 3D echoes are promising for HAR.

3D CNN is one of the most used models for processing 3D data recently [43]–[45]. It extends the spatial CNN into a spatio-temporal model, and spatial-temporal features are learned automatically. Z. Zhang et al. [37] proposed a recurrent 3D CNN model for continuous dynamic gesture recognition using an FMCW radar. 3D CNN was used for extracting short temporal-spatial features in continuous time–range maps, and then a long short term memory (LSTM) was adopted for global temporal feature learning. Experiment showed that when 3D CNN was substituted with a traditional 2D-CNN, the recognition was reduced by around 5%, which demonstrated that compared with 2D CNN, 3D CNN was able to learn better representations of hand gestures. Though the input of 3D CNN is time–range maps, this approach is also suitable for a 3D data cube because the cube contains almost all the activity information in continuous time–range maps.

A representative example using 3D radar echoes for HAR is *GoogleSoli*, as shown in Figure 2.9. *GoogleSoli* is the first gesture recognition system capable of recognizing a rich set of dynamic gestures based on short-range FMCW radar [24], [25]. It is based on an end-to-end trained combination of deep convolutional and recurrent neural networks, and the dataset is comprised of 3D radar echoes. Combining CNN and LSTM could enhance the ability to recognize different activities with varied time spans and spatial distributions. It was shown that the approach with 3D range–Doppler videos was better than the frame-level classification approaches, and the end-to-end 'CNN + LSTM' method could explore the gesture information more thoroughly than the single CNN or LSTM models. With the advent of *GoogleSoli*, other DL architectures have been proposed based on it [37], [45], [46]. Furthermore, Li *et al.* transformed radar echoes into 3D time-range-velocity point sets and proposed a hierarchical PointNet model to classify these point sets for HAR. However, when both range and Doppler frequency information is utilized, the complexity of the HAR method is often higher than that of using only Doppler information.

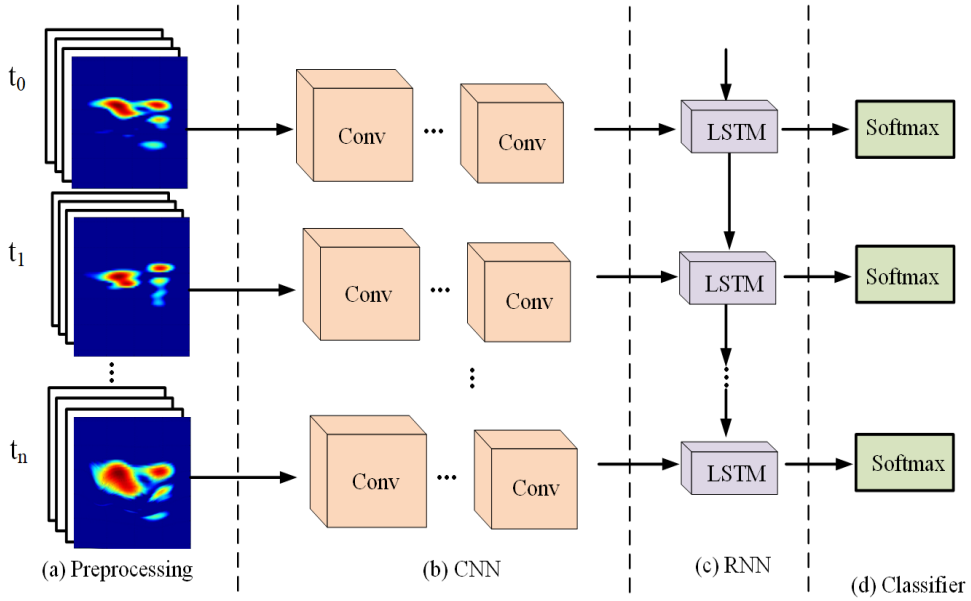


Figure 2.9: Deep learning architecture of Google Soli, a hybrid model that consists of CNN and LSTM. Adopted from [25].

• **Deep Learning Approaches in 2D Radar Echo**

Containing plentiful information of human activity, 3D human backscattering echoes are still complicated to process. 2D radar echoes, which are mainly referred to as time-Doppler maps, time-range maps and range-Doppler maps, also carry sufficient human activity information. Generally, 2D echoes are treated as images, so with the line of computer vision, CNN has become the most commonly utilized model for 2D echoes. Thus, 2D echo-based HAR is often transformed into an image classification task.

(1) **Time–Doppler map** (also referred to as MD spectrograms) includes sufficient time–varying Doppler information that is pivotal for radar-based HAR [47]. When a human target is moving, the main Doppler shift is caused by the torso, while MD is produced by rotating or vibrating parts, such as legs, feet, and hands. The range and velocities of every body part are often different, as shown in Figure 2.10. When the target acts differently, the time–Doppler maps corresponding to these activities are various. Time–Doppler maps are easily obtained by transforming raw echoes with STFT [48] and other joint time-frequency analysis methods. A simple CW radar with one transmitter and one receiver could be employed for identifying basic human activities with time–Doppler maps. In addition, time–Doppler maps are intuitive and explicable. As a consequence, compared with other 2D radar echoes, the time–Doppler maps

are most commonly used for radar-based HAR up to now [27]–[29], [32], [33], [49]–[53].

R.P. Trommel et al. [49] applied a 14-layer deep CNN (DCNN) on time–Doppler maps to classify human gaits. The experimental result showed that the DCNN architecture could extract useful MD features of human gaits even at lower frequencies or low SNR levels, which exceeded the performance of SVM and the artificial neural network. M.S. Seyfioglu et al. [30] employed a CAE architecture to discriminate 12 indoor human activities involving aided and unaided human motions, which often resulted in highly similar MD spectrograms. The CAE model is composed of 3 convolutional layers and three deconvolutional layers. It can learn nuances in the MD spectrograms and attains a good recognition performance of 94.2%. This HAR method shows the potential of radar-based health monitoring systems for assisted living. In [50], a DCNN-based hand gesture recognition system using time–Doppler maps was proposed. There were three convolutional layers and a fully connected layer in the model. In addition, how the DCNN effectively recognizes hand gestures in uncontrolled environments was investigated. Ref. [31] proposed a DCNN architecture composed of cascaded convolutional network layers to classify human activities with time–Doppler maps, as shown in Figure 2.11. The Bayesian optimization with Gaussian prior process was utilized to optimize the network. Experimental results showed that the performance of this method was better than three existing feature-based methods.

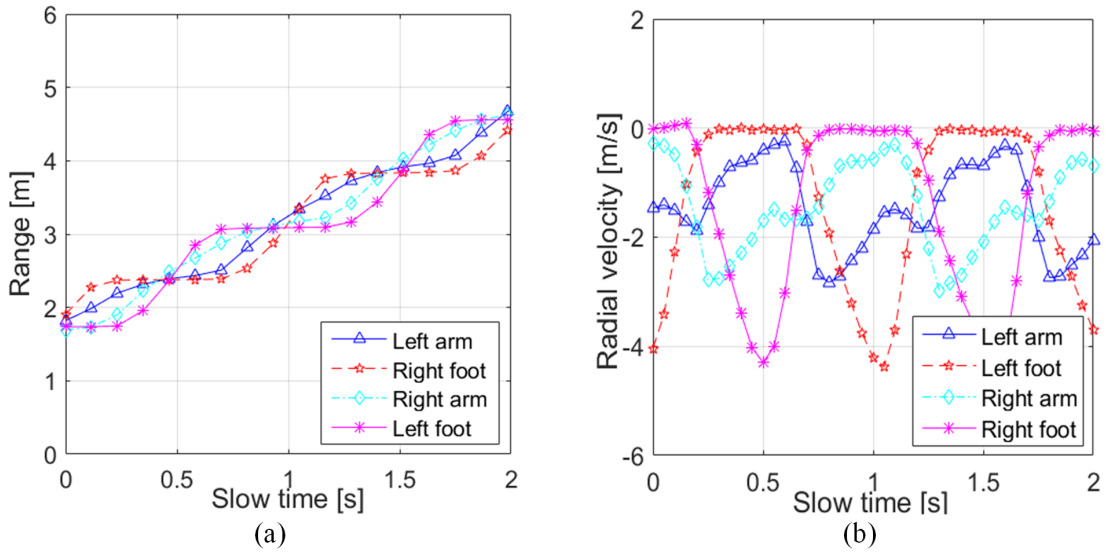


Figure 2.10: Moving trajectories of different body parts when a human target is walking: (a) Range of different parts. (b) Radial velocity of different parts. Adopted from [54].

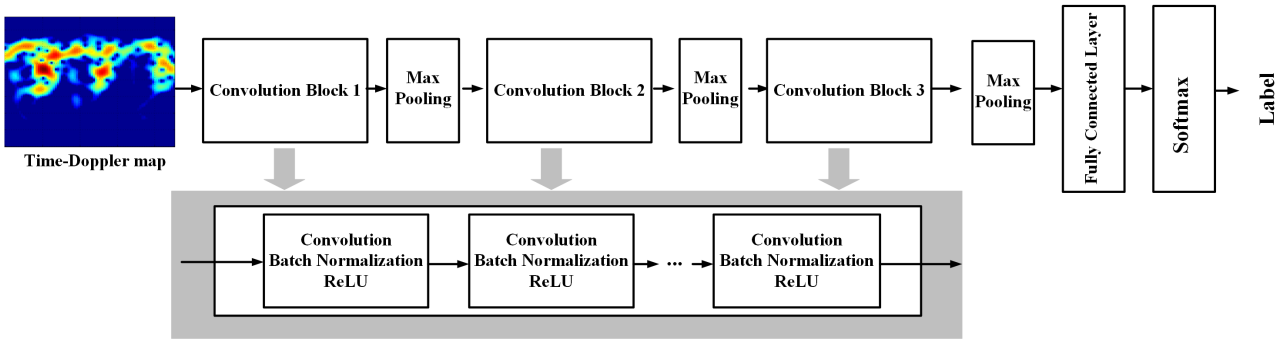


Figure 2.11: Cascaded DCNN optimized by Bayesian learning technique. Adopted from [31].

(2) **Time–range map** is composed of multiple pulses along time (see Figure 2.8c). It contains time-varying range information between the target and the radar. When a person is moving, different components of the human body have different relative distances from the radar, as illustrated in Figure 2.10a. As a result, although time–range maps neglect Doppler information, the time–varying range information of the human body can still be used for recognizing human activities [37]. In [55], time–range maps were utilized to detect falling in assisted living. By providing range information, the false alarms caused by fall-like activities such as sitting were reduced. In [36], Y. Shao et al. employed a three-layer DCNN to classify six human motions such as walking, running and boxing. It was shown that the time–range maps were more robust than the time–Doppler maps, especially when the radial velocity was low. Additionally, when increasing the incident angle, the recognition accuracy was maintained at a stable value because the range information barely changes with the signal to noise ratio.

(3) **Range–Doppler map** (see Figure 2.8d) illustrates range and Doppler information of a moving target at a specific time. It can separate different components of the moving human body parts and locate the target accurately. In addition, range-Doppler maps can track multiple targets simultaneously, promising for multiple human activity recognition. P. Molchanov et al. [56] utilized a short-range monopulse FMCW radar with one Tx and three Rx to sense dynamic hand gestures. A 4D vector representing the hand's spatial coordinates and radial velocity was attained with range-Doppler maps from three antennas. Similarly, in [38], a 4D vector obtained from three range-Doppler maps was combined with a mask from a depth image. Then a resulting velocity layer was fed into a 3D CNN to identify dynamic car-driver hand gestures. The 3D CNN can extract spatial-temporal features, which are indispensable for recognizing dynamic hand gestures of short durations. In [40], two sparse AEs were stacked

to learn sparse representation from range-Doppler maps gradually, and a Softmax layer was employed for classification. In [39], a stack AE was utilized to extract features from range-Doppler maps, and logistic regression was applied for identifying fall/non-fall. Ref. [39], [40] gave examples of using DL methods on range-Doppler maps for HAR.

(4) **Hybrid 2D maps** Up to now, most HAR systems based on 2D radar echoes only utilize one of the above three maps. However, sometimes it is observed that activities that can be easily distinguished with one map may not be correctly identified with another map. This motivates the use of multiple maps aiming at reducing false alarms. Ref. [57] utilized the time-Doppler map, time-range map and range-Doppler map for falling detection. By extracting range and Doppler information from the three maps, the false alarm rate of fall detection can be reduced. In [40], three stack AEs and three Softmax classifiers were employed to classify four human motions (falling, sitting, bending and walking), as described in Figure 2.12. This method applied time–Doppler maps, time–range maps and range-Doppler maps to fully exploit the motion information in radar echoes. Then, three classification results were combined to deliver the final result by voting strategy. Experiments showed that the performance was better than the one that only used one kind of map. In [39], fall detection was divided into two stages: using a stacked AE composed of two sparse AEs to distinguish fall/walk from sitting/bend with time–range maps, and then using another stacked AE with the same structure to distinguish fall from walk with time–Doppler maps. Detection accuracy of 97.1% was achieved.

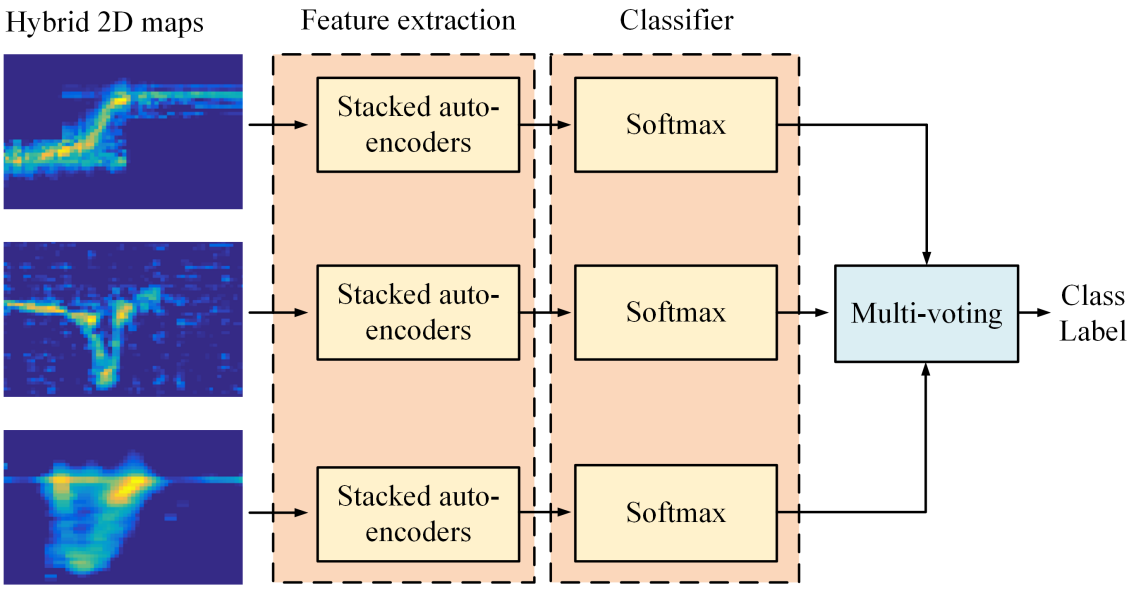


Figure 2.12: The scheme for hybrid 2D maps based recognition. Adopted from [40].

• **Deep Learning Approaches in 1D Radar Echo**

Projecting the time-range-Doppler data cube to the range dimension can result in 1D radar echoes, namely high-resolution range profile (HRRP), as shown in Figure 2.13. Though it is not as intuitive as 2D and 3D radar echoes, HRRP also carries enough information for identifying human activities. Ref. [58] applied HRRP to analyze human target gaits with an ultra-wideband radar. Ref. [59] combined HRRP and MD spectrograms to classify human gaits. Z. Zhou et al. adopted multi-modal signals, including HRRPs and Doppler signatures acquired from a terahertz radar system to recognize dynamic gestures and the recognition rate reached more than 91% [46]. Chen *et al.* [60] proposed a 1-D CNN network to extract features for HAR. Instead of using the conventional joint time-frequency transformation method, data transformation is achieved with the first two layers of CNN. Zhao *et al.* [61] fed the raw radar echoes after preprocessing into an attention-based encoder-decoder model to recognize continuous human activities.

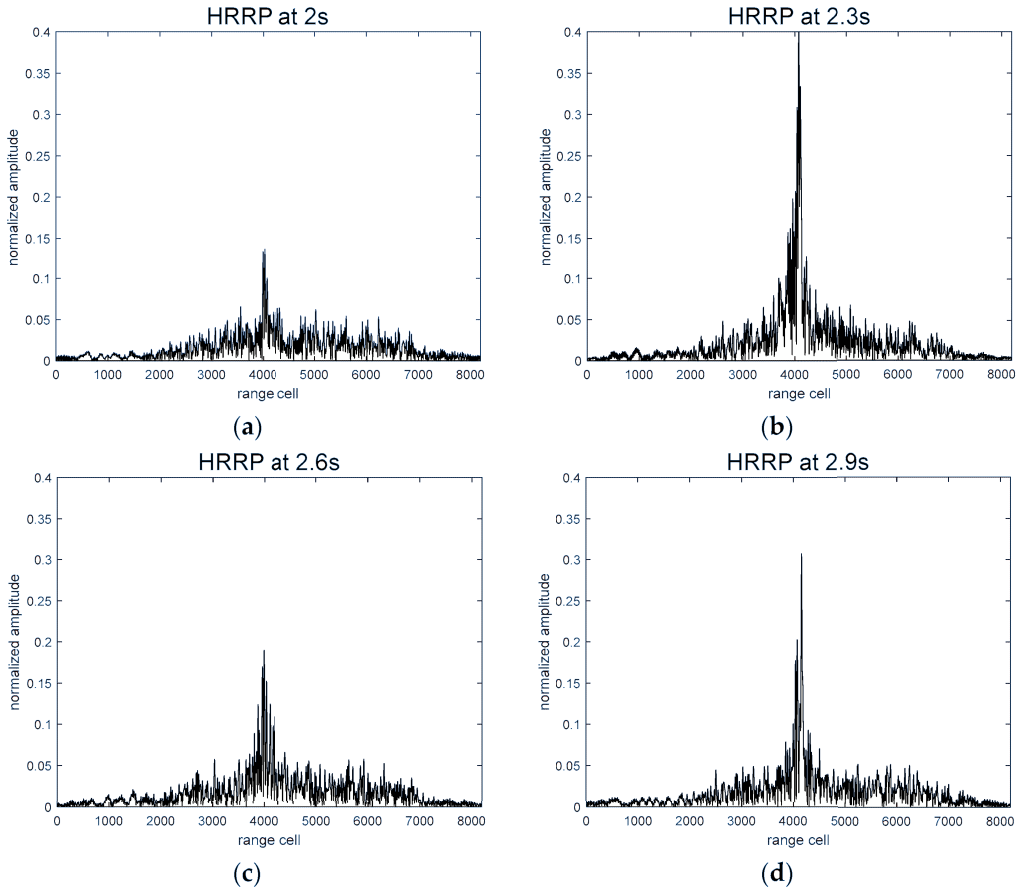


Figure 2.13: High resolution range profiles of a hand at a different time. Adopted from [46]. Each sub-figure illustrates the HRRP at a specific time.

1D radar echoes are time-series and similar to the data obtained from sensors like accelerometers and gyroscopes. Thus, many approaches used for time series could be adopted to 1D echo-based HAR. RNN is often utilized for 1D data due to the advantages of modeling sequential data. For instance, A. Graves et al. proposed a speech recognition architecture composed of LSTM and Connectionist Temporal Classification (CTC) algorithm that is suitable to label unsegmented sequence data [62]. This method provides insights into recognizing continuous activities without annotating manually in advance. A. Hamid et al. [63] applied 1D CNN to the hybrid NN-HMM model for speech recognition and proposed partial weight sharing for the first time. Although there are few DL-related studies for 1D radar echoes, DL approaches have the potential to extract sequential features and deliver good classification results for 1D radar echoes.

2.2.3 HAR with Limited Radar Training Samples

The radar spectrogram is the power distribution of target returns over time and frequency and is a typical 2D representation for analyzing radar MD spectrograms. These spectrograms are individual-unique and motion-unique and have been increasingly used for radar-based HAR [64]–[68].

However, since collecting and annotating radio data manually is time-consuming and expensive, most labeled radar datasets are pretty small-scale. In this circumstance, training a classification model from scratch with limited training data, especially a DL model, often leads to overfitting. Meanwhile, due to the differences in data distribution, directly using a trained model for the HAR task in a new scenario is generally ineffective. As a result, the performance of HAR approaches is often hindered by limited radar data, and data scarcity becomes a bottleneck for the emerging radar-based HAR field. Prompt solutions are required to make DL models generalize well from insufficient annotated radar data.

Current work about radar-based HAR with limited training data can be roughly divided into three categories. The first category is to build classifiers robust to limited training data, such as the models in [69]–[71], and the second category is labeled data augmentation with synthetic data, e.g., [72]–[74]. TL [75], which can take advantage of prior knowledge from an existing large-scale dataset (*source domain*) as a supplement for the tasks on a different but related small-scale dataset (*target domain*), is the third category.

Depending on whether the target datasets are labeled or not, two kinds of TL methods, *i.e.*, the

supervised and unsupervised TL ones, have been proposed for radar-based HAR. The supervised TL [6], [7], [76], [77] utilizes labeled radar data in the target dataset to transfer the source prior knowledge. Such an approach can perform well when sufficient labeled data is available for each class. For instance, Park et al. [78] presented a DCNN model pretrained on *ImageNet* and fine-tuned the network with measured radar MD spectrograms for human aquatic activity classification. Seyfioğlu et al. [6] proposed a residual learning model *DivNet* trained on the simulated radar spectrogram dataset and fine-tuned the model with a measured dataset to classify seven human activities. Additionally, a convolutional autoencoder (CAE) model [5] was first pretrained in an unsupervised manner, and then fine-tuned with a limited number of labeled spectrograms. The fine-tuning (FT) strategy adopted in these methods utilizes a few target data to fine-tune the pretrained DL models, and transfers the source knowledge to compensate for the insufficiency of target domain data. In this thesis, we refer to such an FT strategy as the *Conventional FT*.

However, the performance of *Conventional FT* approaches often degrades when the amount of labeled data decreases. Furthermore, the catastrophic forgetting effect [8] (the tendency of DL models to abruptly forget previously learned tasks after being trained for a new task) usually occurs in the *Conventional FT*. In other words, the performance usually decreases when the model fine-tuned on the target dataset is applied to classify the persons' motions in the source dataset. As a result, the *Conventional FT* method often lacks generalization and cannot scale well to the persons in different domains simultaneously. To deal with this, researchers have presented several solutions and applied the proposed algorithms to various fields. For instance, Jamal et al. [79] demonstrated that the *Conventional FT* suffered from obvious catastrophic forgetting for face detection. And Mallya et al. [80] proposed an iterative pruning method to deal with the catastrophic forgetting effect in the *Conventional FT* approaches.

On the other hand, the unsupervised TL [81], [82], based on domain adaptation with unlabeled training data, is employed to learn domain-invariant feature representation. By utilizing the motion capture database as the source dataset for knowledge transferring, Lang et al. [82] proposed an unsupervised domain adaptation (UDA) method to learn the domain-invariant features for classifying the measured radar data. Du et al. [81] utilized an unsupervised adversarial domain adaption method to reduce the domain discrepancy between the simulated radar spectrogram dataset and the measured spectrogram dataset. Chen et al. [83] proposed two adaptation networks that utilized domain adaptation to eliminate the impact of aspect

angle on HAR with MD spectrograms. However, due to the lack of label information, the performance of the unsupervised methods is generally not as good as the supervised ones.

2.3 Integrated Human Sensing and Communications

Advance in wireless communication and signal processing facilitates ISAC - a technology that combines sensing and communication functionalities to efficiently utilize congested wireless/hardware resources, and to pursue mutual benefits. Consequently, the future communications network will be *perceptive*. Particularly, with the merits of contactless, non-intrusive, and all-weather day-and-night availability, various standardized wireless signals (WiFi, LoRa, etc.) have been explored as a new medium to capture ambient human motions, relying on predefined channel estimation outputs such as the RSSI, and CSI [84]. Recently, such wireless sensing functionalities are primarily implemented by exploiting the reference signals of the standardized fourth-generation (4G) or fifth-generation (5G) waveforms. Hence, the quality of the above sensory data is fundamentally determined by the prefixed pilot structure, the standardized waveform, and the spatial relationships of deployed commodity wireless products. The rationale for the above problem is that existing communication commodities are not primarily designed for information extraction but for information communication.

In general, human activities impact the wireless signal propagation properties such as reflection, diffraction and scattering, which provides human activity sensing opportunities through analyzing and mapping the variations of the received signals with a specific activity [11]. However, its recognition accuracy is subject to the constraints of the communication protocols. On one hand, by complementing communication commodity with built-in sensing functionality [85], wireless devices can balance spatial/time/frequency wireless resources between sensing and communications, and explore untapped signal structures for sensing usage (e.g., data payloads), rather than the standardized pilot structure only [86]. On the other hand, compared to the contactless sensors embedded in the environment, ISAC shows the potential to construct an intelligent system fast and economically, which is sensitive and responsive to the surrounding variations. Therefore, it is essential to evaluate both wireless communication and human-related sensing from a systematic viewpoint, from raw data processing to recognition algorithms to provide a bird's eye view for new researchers in this area.

In this section, we provide a review of human-related sensing in the context of ISAC. We first

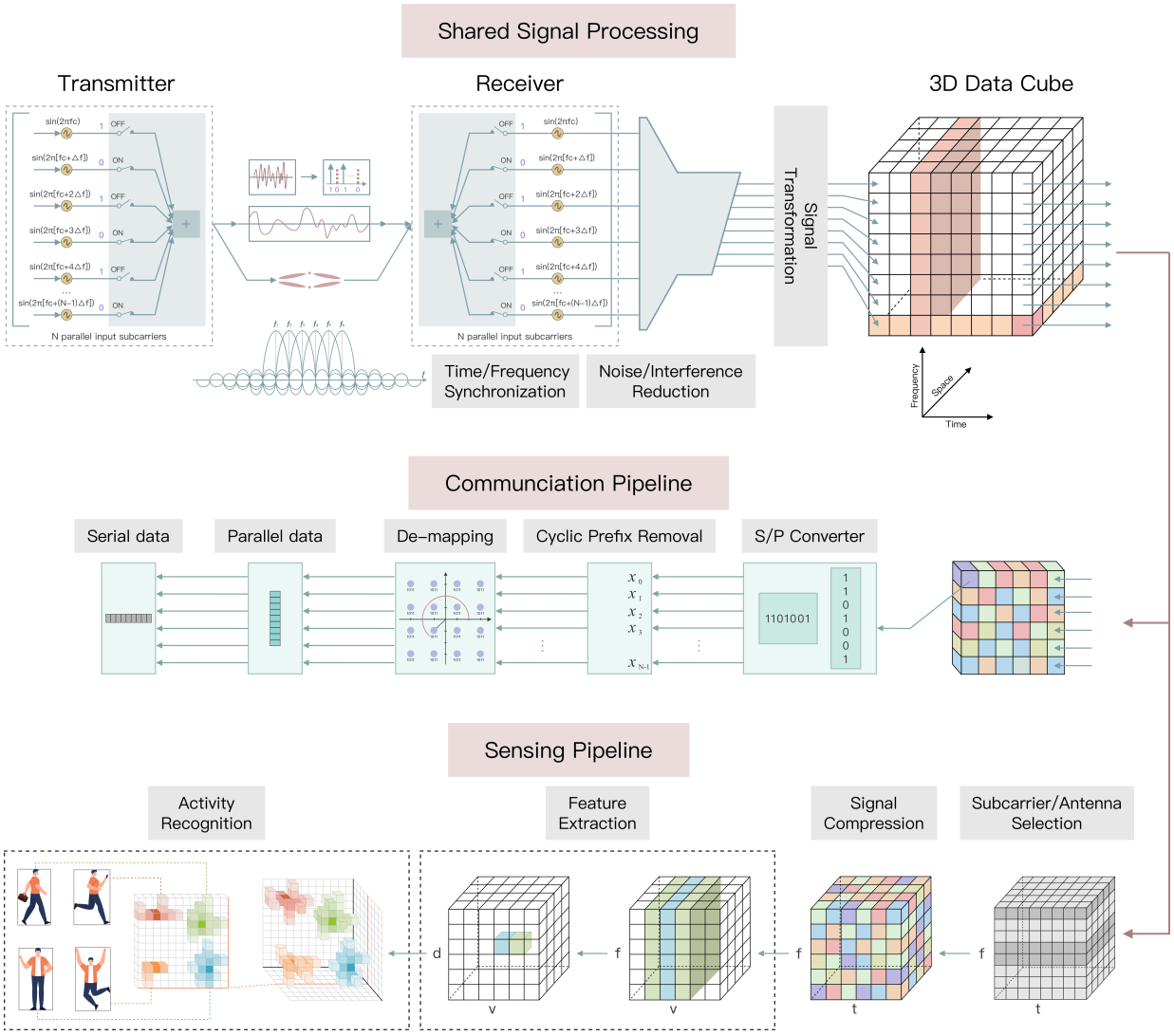


Figure 2.14: The general pipeline of IHASC.

elaborate on the overlap and divergence between wireless communication and sensing processing pipelines, by providing a systematic overview of the IHASC signal processing framework. After that, to explore the impact of wireless devices' geographical and spatial relationships on sensing performance, we identify three typical IHASC configurations (e.g., monostatic, bistatic, and distributed configurations), and discuss their respective key challenges in deployment and implementation. Furthermore, we analyze the impact of several physical system parameters on HAR performance and discuss the relevant optimization principle by jointly considering human sensing and communications. Then, some experimental results are provided to illustrate the potential and capabilities of different IHASC systems.

2.3.1 A Systematic View of IHASC Signal Processing

A communication- and sensing-capable radio emission can simultaneously extract the environmental information while conveying communication data from the transmitter to the intended receiver(s). Even though a number of signaling strategies are able to achieve a unified sensing and communication waveform, the most straightforward implementation is to reuse the communication infrastructures for wireless sensing, with a low-cost and fast-deployment footprint. To provide a systematic view of the evolution from the communication-only devices to the IHASC infrastructure, in this section, we introduce a general IHASC receiver signal processing framework by examining the similarities and differences of current communication and sensing signal processing procedures, with reference to the sensing application of HAR. For each procedure, various state-of-the-art technologies and corresponding challenges are detailed.

• The Shared Procedures

This subsection presents the shared receiving procedures between the communication and sensing signal processing pipelines, as shown in the upper of Fig. 2.14.

i. Time/Frequency Synchronization: Time/frequency synchronization is a fundamental requirement for IHASC systems to achieve both high communication data rates and accurate sensing. Generally, communication demands can be satisfied by compensating the offsets caused by clock asynchronism using embedded pilots, or by absorbing the offsets into channel estimation. However, for sensing, the residual offsets can still trigger estimation ambiguity and consequently produce ghost targets. Assume the carrier frequency of an OFDM system is 3.5GHz, and the oscillator's stability is 10 parts per million (ppm). Then, the CFO can be as large as 3.5 GHz $\times$ 10 ppm= 35 kHz. Even when 10 Hz residual CFO is left after a compensation algorithm dedicated to communications, the estimation error of human radial velocity still reaches 0.86 m/s. Therefore, sophisticated synchronization methods should be devised to ensure high-accuracy human sensing.

ii. Noise/Interference Reduction: Signal distortions such as interference constitute unintended but ubiquitous aspects of any radio system. It is well known that a low signal-to-interference-plus-noise ratio (SINR) severely degrades both communication quality and sensing performance. However, the communication and sensing functionalities show several divergences when dealing with interference. For instance, all transmitting paths contain effective signals for communi-

cation, while some paths (e.g., the paths that are not reflected by targets of interests) are non-desired for sensing and shall be treated as interference.

iii. Signal Transformation: After preliminary time-domain processing, data transformation of IHASC signal measurements is employed for space-time-frequency analysis. A typical transformation is shown in Fig. 2.14, which is a complex-valued 3D data cube and can be shared by communication and sensing. For instance, for an OFDM system, performing an FFT over time dimension transforms the time series signals to the frequency domain, and the Doppler information is attained. Additionally, an FFT over frequency dimension obtains delay presentation, which can also reflect the changes in the surrounding environment.

iv. Signal Separation: With the shared processing procedure, sensing and communication signals may be tightly integrated into a unified IHASC waveform, or be loosely combined in time, frequency, space, or code domains. Therefore, it is vital to distinguish sensing echoes from the entangled received signals for the subsequent HAR procedure, which remains an open issue now.

• Separate Procedures for Sensing

In this subsection, we mainly focus on the signal processing procedure of the sensing pipeline. Additionally, we summarize the sensing characteristics of several existing wireless networks from signal structure, network deployment, and data processing perspectives, as shown in Table 2.2.

i. Subcarriers/Antennas Selection: Due to frequency-selective fading and antenna deployments, the variation patterns of human echoes from different subcarriers/antennas may be diverse, and are susceptible to external factors such as the moving direction of human targets. The received sensing signals that show a weak response to the human target moving cannot improve the HAR performance. In this case, subcarriers/antennas selection is indispensable to retain the signals that show significant fluctuations with the human movement [87].

ii. Signal Compression The goal of signal compression is to remove the redundancy in the discrete 3D range-Doppler-angle sensing signals, such as static background clutter and outdoor environmental noise. There are mainly two signal compression strategies: statistical dimension reduction approaches and clustering approaches based on the range-Doppler-angle estimates. However, since each dimension in the 3D data cube has a clear physical meaning, dimension reduction techniques may destroy the structure and the physical explanatory nature of the data

Table 2.2: Sensing characteristics of diverse wireless networks

	802.11			LTE	NR	LoRa
	ac	ad	ax			
Sensing Range Resolution	∼1 m ∼2 m ∼5 m	∼0.05 m	∼1 m ∼2 m ∼5 m	∼10 m ∼50 m ∼100 m	∼0.5 m ∼5 m ∼30 m	∼500 m ∼1000 m
Signal for Sensings[1]	STF LTF	STF CEF	STF LTF	DMRS SRS, CRS CSI, PRS	DMRS PTRS SRS, CSI	Upchirps Sync Word Downchirps
Available Frequency	5 GHz	60 GHz	2.4 GHz 5 GHz	800 MHz 1.8 GHz 2.6 GHz	7.125 GHz 52.6 GHz	169 MHz 433 MHz 868 MHz 915 MHz
Signal Type[2]	OFDM		OFDMA	SC-FDMA OFDMA	DFT-S-OFDM CP-OFDM	Chirps
Coverage	∼5 m	∼5 m	∼8 m	∼15 m	/	∼50 m
Cooperative Sensing	Protocol supported				Protocol supported	Protocol not supported
Data Fusion	Decentralized fusion Non-cooperative fusion			Centralized/Decentralized fusion Cooperative/Non-cooperative fusion		
Computing Hardware	Access point Phone			BS Phone	BS Phone Edge device	
Measurement for HAR	CSI, RSSI Round trip time (RTT)			CSI RSS	RSSI, CSI Amplitude Phase	Amplitude Phase

[1] SFD: Start Frame Delimiter; STF: Short Training Field; LTF: Long Training Field; CEF: Channel Estimation Field; DMRS: Demodulation Reference Signal; SRS: Sounding Reference Signal; CRS: Cell Reference Signal; CSI: Channel State Information; PRS: Positioning Reference Signals; PTRS: Phase Tracking Reference Signal. [2]OFDMA: Orthogonal Frequency Division Multiple Access; SC-FDMA: Single Carrier-FDMA; DFT-S-OFDM: Discrete Fourier Transform-Spread OFDM; CP-OFDM: Cyclic Prefix-OFDM.

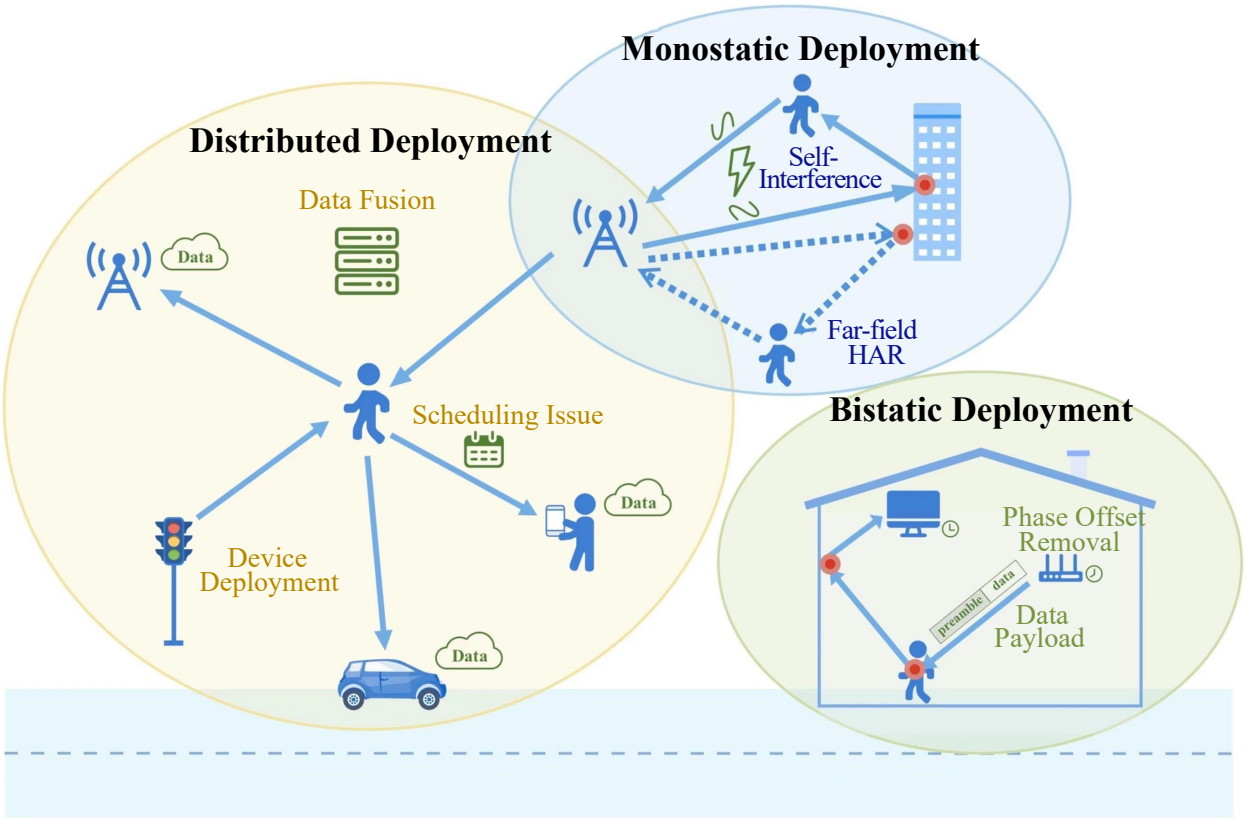


Figure 2.15: Three deployments of IHASC systems.

cube. In contrast, the clustering-based strategy can retain the 3D data structure [88], and the sensing data of the human target can be represented as a 3D range-Doppler-angle point cloud for the subsequent feature extraction.

iii. Feature Extraction: Radio features could be extracted by manual feature engineering, or by automatic DL algorithms. In manual feature engineering, the amplitude and phase of received sensing signals are two commonly used features because they can characterize the impact of human activity on signal propagation. In addition, the time-varying Doppler/micro-Doppler frequency shifts contained in signal phases, which correspond to radial velocities of different components of the human body, are also effective for single-person activity recognition. Additionally, in multi-person scenarios, spatial parameters such as range and angle are indispensable to separate the signals reflected by different targets. In the case of using cloud points data for feature extraction, the intensity of the reflected points and the shape of the point cloud can also be adopted to describe diverse human activities. Meanwhile, DL [89] is an effective tool to automatically extract HAR-related features. Early researchers have exploited the combination of CNN with radio-based HAR feature extraction. However, the temporal information of human

motions gradually vanishes during CNN training, which may seriously degrade the classification performance in the next stage. Hence, even with higher computational complexity, the memory-enabled recurrent NNs are commonly adopted to extract time-varying HAR-related features, resulting in significantly improved performances.

iv. Activity Recognition: The algorithms for HAR can be divided into two categories: model-based algorithms and learning-based algorithms [15]. Model-based approaches, e.g., the Fresnel Zone model and the Angle of Arrival model, mathematically characterize the underlying relationship between human motion and the resultant signal variations. Hence, human movement-related parameters can be quantitatively estimated with the signal dynamics. With clear physical interpretations, model-based algorithms have great potential in achieving fine-grained activity recognition tasks, and can promote the exploration of sensing limits (e.g., sensing coverage and performance bound) of HAR. On the other hand, learning-based approaches, including ML-based and DL-based approaches, aim to learn the mapping between sensing measurements and the label of the corresponding human activities by using pre-extracted features. Generally, both manually extracted and automatically extracted features can be employed in learning-based methods. However, in most cases, the DL-based feature extractor and classifier are combined to process input data and then, classify human activity in an end-to-end manner without any human intervention.

2.3.2 Unique Signal Processing for HAR with Various Deployments

There are mainly three deployments of IHASC systems, i.e., monostatic, bistatic, and distributed deployments, depending on the locations of transmitters and receivers. In this section, we discuss the unique properties of these deployments beyond the general signal processing operations, and present their key problems when used for HAR.

- **Monostatic Deployment**

A monostatic system transmits wireless signals to sense the environment and captures the target echoes via the sensing receiver co-located and synchronized with the transmitter. One example is a 5G New Radio (NR) base station (BS) that senses the environment using the echoes of its transmitted downlink communication signals. Self-interference (SI) and far-field HAR are two major concerns in such a deployment.

i. Self-Interference (SI): In a monostatic system, the leaked transmitted signals generally interfere with the desired echo signals, also known as SI. The strong SI can saturate the receivers and overwhelm the target echoes. Since most modern communication systems transmit continuous waveform, it is infeasible to use a dumb period for receiving echoes after an ultra-short transmission period, like in pulse radar, or adopt the transmission signal as the local oscillator input to remove SI, like in a frequency modulated continuous radar. Full duplex is a long-term solution to this issue, as described in [90], together with suboptimal near-term solutions, such as deploying a receiving antenna dedicated to sensing, widely separated from other antennas.

ii. Far-Field HAR: According to the spatial relationship between transmitters and targets, the sensing area can be divided into two distinct regions: near-field and far-field. In far-field, due to the low received signal power, low signal-to-noise ratio (SNR) and multipath propagation, the reflected signals from the human target may be overwhelmed by the background clutter [91]. Specifically, the micro-Doppler frequencies, which are efficient features for HAR, may be too weak to be captured. Alternatively, other features such as the time-varying range information of different human body segments can be employed for recognition [39].

Table 2.3: Properties and challenges of the three IHASC deployments.

Deployment	**Properties**	**Challenges**
Monostatic Deployment	Known sensing signals Synchronized transceiver	Self-interference Far-field HAR
Bistatic Deployment	Tx/Rx spatially isolated Compatible with existing networks	Phase offset removal Unknown data payload
Distributed Deployment	Wide coverage Multidirectional sensing Multi-node collaboration	Data fusion Device deployment Scheduling issue

• **Bistatic Deployment**

Bistatic deployment refers to an IHASC system where transmitters and receivers are spatially separated. One typical example is Wi-Fi sensing [12]. Compared with monostatic systems, the bistatic deployment is more compatible with existing communication networks such as WiFi and cellular networks. Furthermore, the SI issue is naturally avoided with the spatially bistatic scheme. Nevertheless, phase offset removal and unknown data payload are two pressing problems in bistatic deployments for performing accurate HAR.

i. Phase Offset Removal: Due to the oscillator instability, phase offset usually exists in the received signals of bistatic systems, leading to measurement ambiguity and accuracy degradation. For instance, SFO generally introduces high variations in the phase of sensing signals, and can even drown out the small phase changes caused by human movement. To compensate for the phase offsets and recover the information loss, cross-antenna correlation and cross-antenna ratio techniques are applied [90], based on the fact that the phase offsets between different receive antennas are the same. An alternative strategy discards the phase information and only uses the signal magnitude, which results in degraded sensing performance [11].

ii. Exploiting Data Payload for Sensing: Current communication-centric IHASC systems mainly utilize pilot signals for sensing applications. However, instead of only using a pilot, employing the entire frame as sensing signals can potentially achieve higher SNRs and a finer Doppler frequency resolution. In most existing bistatic IHASC systems, the data payload is unknown on the receiver side. To this end, a possible strategy is to first decode the unknown data payload according to the channel estimation results, and then employ the entire frame for sensing. However, the actual benefit of achieving improved SNRs with the sensing-after-decoding scheme is yet to be verified, and may only be prominent over a limited range of SNRs.

• **Distributed Deployment**

In a distributed system, all transmitting and receiving devices are distributed in different spatial locations, which can provide spatial diversity of the illuminated target and obtain human activity information to deal with target fluctuation [90]. In a distributed system, apart from the issues in the bistatic system, systematic design and arrangement, such as receiving data fusion, infrastructure deployment, and scheduling issues, are indispensable.

i. Data Fusion: An efficient data fusion strategy is essential to remove data redundancy from

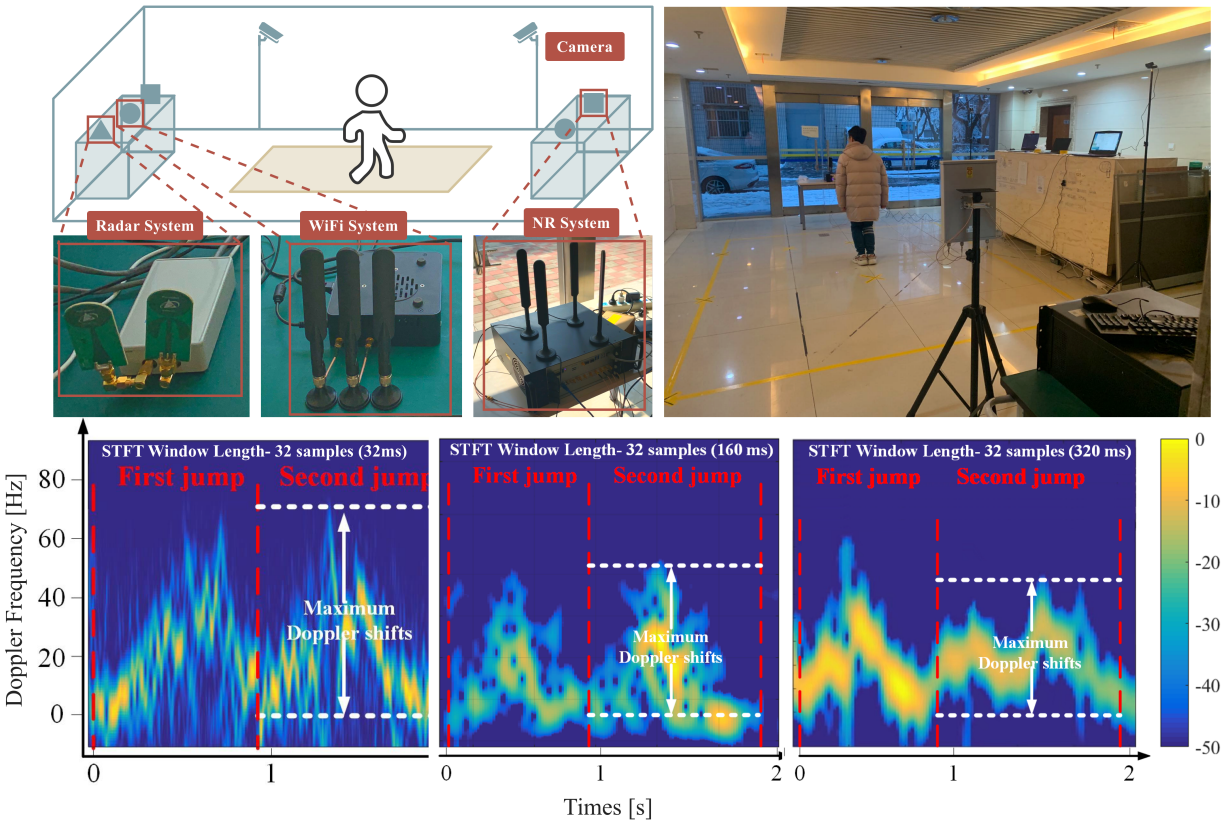


Figure 2.16: Experimental setup (top) and the resulting time-Doppler frequency spectrograms (bottom) of a human target jumping forward twice, for WiFi (left), radar (middle), and 5G NR (right) systems, respectively.

different sources and yield global feature representations. Data-level, feature-level, and decision-level fusion are three typical data fusion techniques [92]. In data-level fusion, raw data from different nodes are sent to the fusion center (FC) and then aggregated for extracting the HAR-related information. However, sending a huge amount of sensing data to the FC generally causes a large communication burden and high hardware costs. In feature-level and decision-level fusion, each node can preprocess its data and send the output features/decisions to the FC, which require less data exchange between FC and local nodes, greatly saving energy and computing resources at the FC. Furthermore, such decentralized strategies allow flexible algorithm designs at different branches and hence can extract unique information from various devices.

ii. Deployment of Host and Slave Devices: In a communication-centric IHASC system, both the host devices and the slave devices can act as sensing transceivers. However, host device deployments in a pure communication system and an IHASC system generally follow different

principles. For instance, in a cellular network, BSs are deployed with little signal overlap to avoid interference between cells, which is not suitable for IHASC systems, because interference also contains useful sensing information. Therefore, in IHASC systems, the deployment of host devices faces a trade-off between communication interference and sensing performance. On the other hand, though slave device deployment is independent of the communication performance, the placement of terminals can affect sensing factors such as coverage, orientation, and angles, and needs to be optimized for better sensing performance.

iii. Scheduling Issue with Target Echoes: In addition to data fusion and device deployment, sensing human activities also imposes challenges in resource scheduling of distributed IHASC systems. Since the human echoes could randomly appear in time, frequency, and spatial domains, novel IHASC scheduling algorithms are required to predict the appearance of random echo signals and schedule the echoes in an orderly manner [86]. Furthermore, intelligent resource allocation algorithms can be designed to generate scheduling strategies for all device nodes, based on context information like quality-of-service (QoS) requirements and battery consumption.

• **Summary and Design Factors**

In light of the discussions above, we summarize the properties and challenges of the three types of deployments in Table 2.3. It can be inferred that the distributed deployment can yield the best HAR performance due to spatial diversity and wide coverage, and has greater potential for through-the-wall HAR and compound activity recognition. However, more signal processing and computational resources are required. In real-world applications, one can select the system deployment by considering the site and resource constraints.

In addition to the deployment, the system parameters of IHASC systems, as inherited from communication designs, could also have notable impact on human-related sensing functionality. In Table 2.4, we present the impact of some major system physical parameters on HAR performance.

2.3.3 Over-the-air Experiments and Results

In this section, we show some experimental results of using different communication signals for HAR.

Table 2.4: Impact of Physical Parameters on HAR Performance.

Physical Parameters	**Impact on Sensing Pipeline**	**Impact on HAR performance**
Total Signal Bandwidth B	Larger B leads to finer range resolution	Better multi-targets separation ability along range domain with larger B
Carrier Frequency f_c	Greater f_c leads to finer velocity resolution but smaller unambiguous velocity[2]	Different moving components of the human target can be be recorded with finer granularity in velocity domain
Symbol Duration T_s	Larger T_s leads to longer unambiguous range but lower range resolution	Longer detectable range with T_s increasing, promising for far-field HAR
Subcarrier Interval T[1]	Smaller T leads to larger unambiguous velocity but lower velocity resolution	Wider coverage in velocity domain to record the activities with higher velocity components
MIMO Antenna Array	Larger antenna aperture and higher angular resolution with more antennas	Locating human target more precisely at the angular direction; Distinguishing multiple targets at closer angular directions
Transmission Power	Positively correlated with the coverage of the sensing system	Wider HAR coverage with higher transmission power; Stronger echo signals and more robust to interference

[1] For single-subcarrier signals, T is T_s, and is euqal to $1/B$.
For OFDM signals, T includes T_s and cyclic prefix, and is equal to N/B, where N is the number of subcarriers.

[2] Maximum unambiguous velocity $v_{max} = c/(2f_cT)$, and velocity resolution $\Delta v = v_{max}/M$, where c is the velocity of light, and M is the number of symbols.

In the experiments, the reflected sensing signals corresponding to a human jumping forward in an in-door scenario of 20 m^2 are collected with three wireless systems: 1) a ultra-wideband (UWB) radar with 1.0 GHz bandwidth, 1.0 ms pulse repetition interval (PRI), and 4.0 GHz central frequency, 2) a WiFi system with 40 MHz bandwidth, 5.0 ms PRI, and 5.8 GHz central frequency, and 3) a 5G NR BS system with 100 MHz bandwidth, 10.0 ms PRI, and 3.6 GHz central frequency. Noting that both WiFi and NR systems collect signal reflections from separately deployed receiving antennas such that they are working in the bi-static sensing mode. We preprocessed these sensing measurements using some of the methods described in Section 2.3.1 and 2.3.2, and transformed the measurements into time-Doppler frequency spectrograms by using the STFT with a window of 32 samples.

To compare the sensing performances, we plot their micro-Doppler signatures in Fig. 2.16. It can be observed that, although the radial velocity components are of similar magnitude, the Doppler shifts produced by the system with a higher center frequency are more pronounced. More distinct frequency shift components can enable the activity-related micro-Doppler features to be extracted from the reflected echoes manually, improving the performance of modeling-based HAR approaches. Meanwhile, low PRI leads to finer time resolution, enabling the echo signals to convey the time-varying characteristics of human activities in more detail. Therefore, the system with low PRI can be used to distinguish some similar human activities. Furthermore, the intensity of the spectrogram from the NR system is the strongest, indicating the NR system can be more robust to interference and has a wider coverage for HAR.

2.4 Summary

This chapter reviewed the current work on contact-free human activity sensing with WiFi and radar signals. Specifically, we first overviewed the state-of-the-art WiFi-based human sensing methods, which can be divided into modeling-based and learning-based approaches. Then, we introduced radar-based HAR approaches, emphasizing DL-empowered methods. Since DL is a data-driven approach whose performance could degrade when training data is insufficient, we further reviewed the current DL-based human sensing with limited radar training samples. Furthermore, we illustrated the general pipeline of IHACS signal processing and categorized IHASC systems into three typical deployments to elaborate on the characteristics and problems in these three deployments.

Chapter 3

Doppler Speeds Estimation of Moving Human Target with Cross-Antenna Signal Ratio

3.1 Introduction

In a variety of context-aware human sensing tasks, estimating the Doppler frequencies from the received signals is an essential step [2]. For instance, the estimated Doppler frequency can be used to determine respiration pattern and human moving velocity [11]. Passive WiFi radar (PWR) has been designed to estimate Doppler frequency indoors [93], [94]. In PWR, two clock-synchronized receivers collaboratively work, one receiving direct WiFi signals as references and the other receiving echo signals for sensing. Another mainstream of WiFi sensing, which has more joint communications and sensing flavor, employs the by-product of a WiFi communication receiver, namely the channel state information (CSI), which does not require a separate reference receiver [3], [18]. In such WiFi sensing systems, Tx and Rx are generally geometrically separated, like a bistatic radar system. However, unlike in the bistatic radar, there is typically no common clock between the spatially-separated Tx and (sensing) Rx, which leads to the clock asynchronism issue. To estimate CSI, a WiFi receiver would generally use the training signals to synchronize with the WiFi transmitter. This level of synchronization, however, is not sufficient for accurate sensing [4]. In particular, the CFO residual in a WiFi receiver can be tens hertz in general. Even it is 10 Hz, the Doppler estimation error can be as

large as 1.25($= 10 \times \frac{3\times10^8}{2.4\times10^9}$) m/s, for a 2.4 GHz WiFi system. Therefore, the communication synchronization errors must be further addressed for accurate sensing. As a result, when using complex signals instead of the power only for accurately estimating Doppler frequencies, removing the phase offsets in WiFi CSI induced by clock asynchrony is an essential prerequisite step.

In this chapter, we study Doppler frequency estimation using the CASR method for scenarios with general movement. We first develop a CSI-ratio expression disclosing more insights, using a more general CSI model as a function of delay, Doppler frequency and AoA, instead of the widely used and simpler one only based on signal propagation distance [3], [20], [21]. We then propose three algorithms for estimating Doppler frequencies: Mobius Transformation-based, signal difference-based, and periodicity-based. These algorithms exploit different features of the CSI ratio in terms of Doppler frequencies and can be applied to scenarios involving general and/or irregular movement. We describe these algorithms by referring to human tracking applications in this paper, but they can be easily adapted to other applications. Using a publically available WiFi CSI dataset *Widar* 2.0, we then validate the efficiency of the proposed Doppler frequency estimation algorithms.

The rest of the chapter is organized as follows. Section 3.2 presents and analyzes the CSI ratio based on a general CSI model. Three CSI-ratio-based Doppler frequency estimation algorithms are provided in Section 3.3. In Section 3.4, experimental results are presented to validate the efficiency of the proposed methods. Section V provides a more detailed review of related work on RF-based human sensing and phase offset removal in bi-static wireless sensing systems. Finally, the conclusion is provided in Section 3.5.

3.2 Sensing Signal Model

In this section, we first introduce the basic concept of CSI and present the relationship between the Doppler frequency of a moving human target and the change of CSI measurements. Then, we adopt the cross-antenna signal radio strategy to solve the clock synchronization problem, and describe the properties of the produced CSI-ratio measurements when being utilized for Doppler frequency estimation.

3.2.1 CSI Model

We consider an MIMO-OFDM system with M_T transmit antennas and M_R receive antennas. Let B denote the total signal bandwidth, N be the number of total subcarriers, $f_0 = B/N$ be subcarrier interval, and T_s be the OFDM symbol period.

Consider a planar wave-front signal propagation model. The array steering vector of a uniform linear antenna array (ULA) at the receiver is given by

$$\mathbf{a}(M_R, \alpha) = [1, e^{ju(\alpha)}, ..., e^{j(M-1)u(\alpha)}]^{\mathrm{T}}, \tag{3.1}$$

where $u(\alpha) = 2\pi d/\lambda \sin(\alpha)$, λ is the wavelength, d is the antenna spacing, and α is the angle-of-arrival (AoA) of a signal path. Similarly, we can define $\mathbf{a}(M_T, \beta)$ for the transmitter array, where β is the angle of departure (AoD).

The frequency-domain OFDM channel state matrix $\mathbf{H}(t)$ at the n-th subcarrier at time t can be represented as [95]

$$\begin{aligned}\mathbf{H}_n(t) = e^{j\phi_n(t)} \sum_{l=1}^{L} b_l e^{-j2\pi n(\tau_l+\tau_o(t))f_0} e^{j2\pi(f_{D,l}+f_o(t))t} \\ \cdot \mathbf{a}(M_R, \beta_l)\mathbf{a}^{\mathrm{T}}(M_T, \beta_l),\end{aligned} \tag{3.2}$$

where b_l is the amplitude of the l-th multipath; τ_l is the propagation delay; $f_{D,l}$ is the Doppler frequency caused by the moving human target; $\phi_n(t)$, $\tau_o(t)$ and $f_o(t)$ are the time-varying phase shift, TMO and CFO induced by clock asynchronism between transmitters and receivers, respectively.

Since our algorithms to be proposed can be similarly applied to each subcarrier, we drop the subscript n hereafter. We also consider the CSI with a single transmitting antenna. For multiple transmitting antennas, the channel matrix can be easily separated for each antenna as the training sequences across transmitting antennas are generally orthogonal. Assume that there is only one dynamic path and L_s ($L_s \geq 1$) static paths, which corresponds to approximating the single mobile human target to be sensed as a point source.

Under the above setup, we rewrite the channel CSI at the m-th receive antenna, $H^m(t)$, as the sum of static paths and dynamic path. Let

$$H_s^m(t) = \sum_{l_s=1}^{L_s} b_{l_s} e^{-j2\pi n\tau_{l_s} f_0} e^{j(m-1)u(\theta_{l_s})}, \tag{3.3}$$

and

$$H_d^m(t) = b_d e^{-j2\pi n \tau_d f_0} e^{j(m-1)u(\alpha)} \tag{3.4}$$

represent parts of the static and dynamic paths, respectively, where τ_{l_s} and θ_{l_s} are the TMO and AoA of the l-th static path, respectively; τ_d and α are the TMO and AoA of the dynamic path, respectively, and $u(\theta_{l_s}) \triangleq 2\pi d/\lambda \sin(\theta_{l_s})$. Note that the AoD related phase term has been absorbed into b_{l_s} and b_d. We can then represent $H^m(t)$ as

$$\begin{aligned} H^m(t) =& e^{j\phi_n(t)} e^{j2\pi(f_o(t)t - n\tau_0(t)f_0)} \cdot \\ & \left(H_s^m(t) + H_d^m(t) e^{j2\pi f_D t}\right). \end{aligned} \tag{3.5}$$

3.2.2 CSI-Ratio Model

The CSI ratio $R(t)$ between the m-th and $(m+1)$-th receiving antennas can be expressed as

$$\begin{aligned} R(t) &= \frac{H^m(t)}{H^{m+1}(t)} \\ &= \frac{H_s^m(t) + H_d^m(t) z(t)}{H_s^{m+1}(t) + H_d^{m+1}(t) z(t)}, \end{aligned} \tag{3.6}$$

where

$$z(t) \triangleq e^{2\pi f_D t}. \tag{3.7}$$

The varying speeds of sensing parameters, Doppler frequency, propagation delay and AoA are very different due to the movement of a human target. AoA typically changes most slowly, and then delay and Doppler frequency. When a human target is moving at a speed up to five meters per second, we can reasonably assume they are all fixed over a period of tens of milliseconds. Therefore, in such a short time period, $H_s^m(t)$, $H_s^{m+1}(t)$, $H_d^m(t)$ and $H_d^{m+1}(t)$ are constant, and can be denoted as H_s^m, H_s^{m+1}, H_d^m and H_d^{m+1}, respectively. In this case, $R(t)$ only varies with $z(t)$ and can be rewritten as

$$\begin{aligned} R(t) &= \frac{H_s^m + H_d^m e^{j2\pi f_D t}}{H_s^{m+1} + H_d^{m+1} e^{j2\pi f_D t}} \\ &= \frac{H_s^m + H_d^m e^{j2\pi f_D t}}{H_s^{m+1} + H_d^m e^{ju(\alpha)} e^{j2\pi f_D t}}. \end{aligned} \tag{3.8}$$

Collecting CSI measurements over this period, we can then estimate the Doppler frequency f_D based on Equation (3.8).

3.3 Proposed Doppler Frequency Estimation Methods

In this section, we propose three Doppler frequency estimation algorithms, by exploiting several different properties based on the CSI-ratio model developed above. These algorithms estimate a Doppler frequency for each segment of CSI-ratio samples, over a time window of tens-of-millisecond.

Firstly, we apply a calibration step to avoid the estimation of a pseudo Doppler frequency when the target is moving at almost zero velocity or is stationary. In this case, the phase of CSI data changes slightly due to the environmental noise. If this case is not identified in advance, our proposed algorithms will estimate a pseudo Doppler frequency. To deal with this issue, we calculate the square deviation of CSI amplitude in the time window, and compare the deviation value with a preset threshold. If the deviation is larger than the threshold, we then activate one of the proposed algorithms. Otherwise, the Doppler frequency in this time window is set to 0 Hz.

Next, we describe the three algorithms.

3.3.1 Doppler Frequency Estimation based on Mobius Transformation

In the complex plane, with t increasing, $z(t)$ is a unit circle rotating clockwise or anticlockwise, depending on the sign of f_D. When H_s^m, H_s^{m+1}, H_d^m and H_d^{m+1} are invariant and satisfy $H_s^m H_d^m$ - $H_s^{m+1} H_d^{m+1} \neq 0$, $R(t)$ can be treated as the Mobius transformation of $z(t)$ in the complex plane, as illustrated in Figure 3.1.

After the translation, complex inversion, and multiplication transform in Figure 3.1, we can see that the CSI ratio also changes along a circle in the complex plane, although the speed of variations become non-uniform, due to the inversion transform. When the magnitude of the static component is larger than that of the dynamic one, the CSI ratio $R(t)$ and $z(t)$ rotate in the same direction in the complex plane; otherwise, they rotate in opposite directions. For human sensing applications in real scenarios, the magnitude of the static component H_s^m is generally larger than that of the dynamic component H_d^m [3]. In this case, the rotation direction of $z(t)$ can be inferred based on the direction of $R(t)$.

Furthermore, according to the Mobius transformation, $z(t) \rightarrow R(t)$ is a conformal map, as

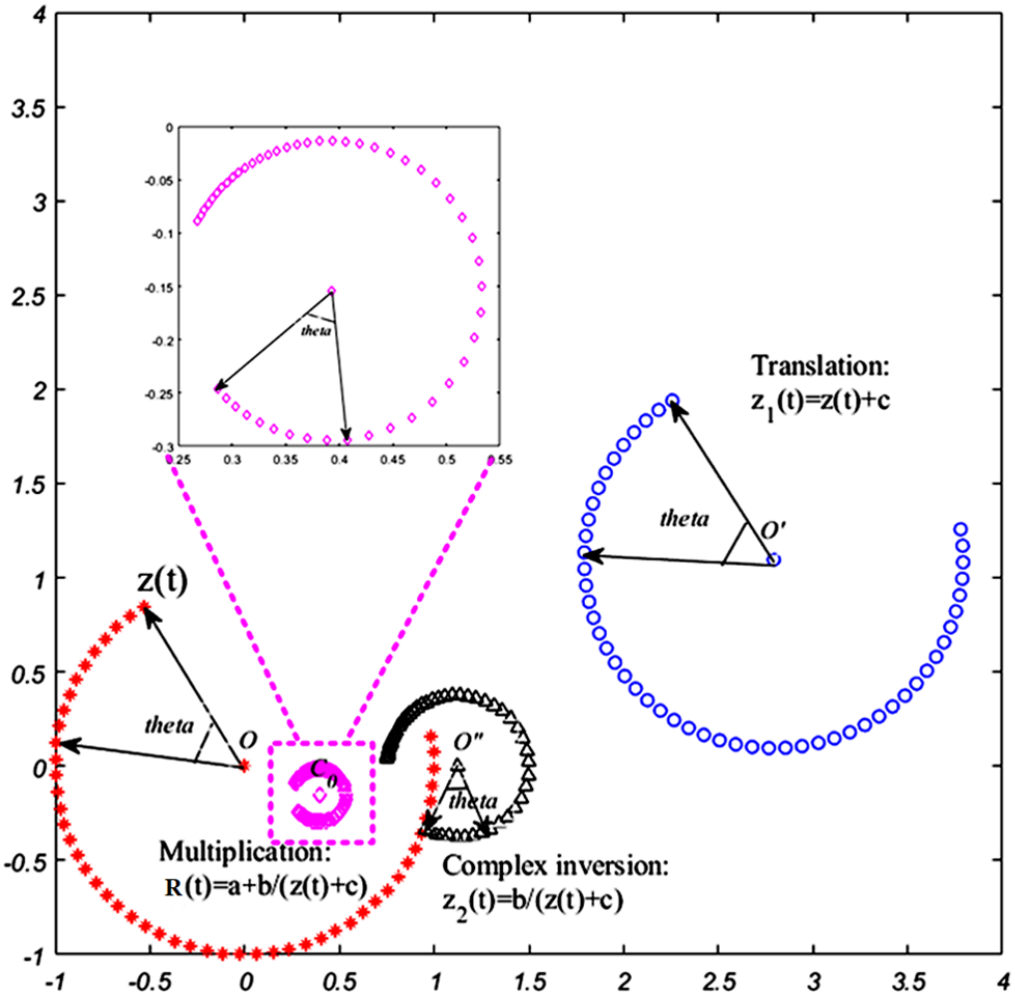


Figure 3.1: Illustration of Mobius Transform. With the translation, complex inversion, multiplication operations, z(t) is transformed to R(t), which is the Mobius transform of z(t).

shown in Figure 3.1 [96]. Let $\theta_R(t)$ be the angle of a point at $R(t)$ with respect to its center C_0. That is, $\theta_R(t) = \angle(R(t) - C_0)$. Then, at some relatively larger time interval Δt, the angle variation $\Delta(\theta_z)$ of $z(t)$ and $\Delta(\theta_R)$ of $R(t)$ is approximately equal, i.e.,

$$\Delta(\theta_R) \approx \Delta(\theta_z). \tag{3.9}$$

When Δt is a few milliseconds, the Doppler frequency $f_D = \frac{v_D}{c} f_c$ of the human target can be regarded as invariant, where v_D is the relative radical speed of the human target, f_c is the carrier frequency, and c is the speed of the light. Therefore,

$$\Delta(\theta_R) \approx \Delta(\theta_z) = 2\pi f_D \Delta t. \tag{3.10}$$

To obtain $\Delta(\theta_R)$, we need to estimate the center C_0 of the CSI-ratio samples $R(t_k)_{k=1}^n = \{R(t_1), R(t_2), ..., R(t_n)\}$ within Δt, where $\Delta t = t_n - t_1$. We adopt the least square method to estimate C_0.

Proposition 3.1: In the complex plane, let the coordinate of C_0 be (A, B) and the radius of the CSI-ratio circle be r, the coordinate of C_0 can be obtained by

$$\min_{A,B,r} \sum_{k=1}^{n} \delta_k^2 = \min_{A,B,r} \sum_{k=1}^{n} \left((x_k - A)^2 + (y_k - B)^2 - r^2\right)^2, \tag{3.11}$$

where (x_k, y_k) is the coordinate of $R(t_k)$ in the complex plane.

Proof: Please refer to the Appendix.

Note that not a whole circle samples are needed to get the estimate of C_0, and quite often a segment of arc is sufficient. Then, the coordinate of C_0 can be estimated as given in Equation (A.16). With the estimated C_0, the coordinate of $R(t_k)_{k=1}^n$ with respect to C_0 can be calculated. Then, the new CSI-ratio samples $R_s(t_k)_{k=1}^n$ is given by

$$R_s(t_k)_{k=1}^n = R(t_k)_{k=1}^n - C_0, \tag{3.12}$$

which equivalently shifts the center of $R(t)$ to the origin of the complex plane.

Therefore, the angle $\theta_R(t_k)_{k=1}^n = \{\theta_R(t_1), \theta_R(t_2), ..., \theta_R(t_n)\}$ and the magnitude (i.e., absolute value of the amplitude) $a_R(t_k)_{k=1}^n = \{a_R(t_1), a_R(t_2), ..., a_R(t_n)\}$ of the new CSI-ratio samples $R_s(t_k)_{k=1}^n$ with respect to C_0 can be obtained. Then, based on Equation (3.10), when f_D is invariant, $(t_k, \theta_R(t_k))_{k=1}^n = \{(t_1, \theta_R(t_1)), (t_2, \theta_R(t_2)), ..., (t_n, \theta_R(t_n))\}$ form a linear mapping, i.e.,

$$\theta_R(t_k) = \beta_1 t_k + \beta_2, \tag{3.13}$$

where β_1 is equal to $2\pi f_D T_s$.

To get the estimate for β_1, we adopt a weighted linear fitting method

$$\min_{\beta_1,\beta_2} Q(\beta_1, \beta_2) = \min_{\beta_1,\beta_2} \sum_{k=1}^{n} w_k \left(\theta_R(t_k) - (\beta_1 t_k + \beta_2)\right)^2, \tag{3.14}$$

where we use the magnitude $a_R(t_k)$ as the weight w_k. The weights can reduce the impact of phase errors of small ratio. Then, the estimated f_D can be calculated as $\beta_1/(2\pi T_s)$.

The main steps of estimating f_D with the Mobius-Transformation-based algorithm are illustrated in Algorithm 1. Note that this method can directly estimate the sign of the Doppler frequency, and is the only one with this capability in the proposed three algorithms in this paper.

Algorithm 1 Doppler Frequency f_D Estimation based on Mobius Transformation

Input: A sequence of CSI-ratio samples $R(t_k)_{k=1}^n = \{R(t_1), R(t_2), ..., R(t_n)\}$ within Δt, where $\Delta t = t_n$ - t_1

Output: The estimated f_D within $[t_1, t_n]$

1: The coordinate of the center C_0 of $R(t_k)_{k=1}^n$ is denoted as $[A, B]$
2: Find the values of A and B that make the sum of deviation δ_k^2 in Equation 3.11 reaches the minimum value
3: The new CSI-ratio samples $R_s(t_k)_{k=1}^n$ are calculated as $R(t_k)_{k=1}^n - C_0$
4: The angle of $R_s(t_k)_{k=1}^n$ are denoted as $\theta_R(t_1)$, $\theta_R(t_2)$, ..., and $\theta_R(t_n)$, and the magnitude are $a_R(t_1)$, $a_R(t_2)$, ..., and $a_R(t_n)$
5: Estimate the value $\beta_1 = 2\pi T_s f_D$ with the angle and the magnitude of $R_s(t_k)_{k=1}^n$, by using the weighted linear fitting methods in Equation 3.14
6: $f_D = \beta_1 / 2\pi T_s$
7: **return** Estimated f_D within the time period $[t_1, t_N]$

3.3.2 Doppler Frequency Estimation based on Periodicity of CSI Ratio

Since $z(t) \rightarrow R(t)$ is a conformal map, $z(t)$ and $R(t)$ have the same periodicity. Therefore, when the radian of $z(t)$ changes by 2π, $R(t)$ goes exactly one period. In this case, we can estimate f_D based on the periodicity of $R(t)$. Specifically, for a sequence of CSI-ratio samples $R(t) = \{R(t_1), R(t_2), ..., R(t_n)\}$, denote their angles relative to the point O=(0, 0) on the complex plane as $\gamma(t) = \{\gamma(t_1), \gamma(t_2), ..., \gamma(t_n)\}$. Then, if there are two angles separated by a reasonable amount of time are similar, we can infer that $R(t)$ goes through a cycle within this period Δt. Then, f_D can be estimated as

$$f_D = \frac{2\pi}{2\pi\Delta_t} = \frac{1}{\Delta_t}. \tag{3.15}$$

The detailed f_D estimation process based on the periodicity of CSI ratio is presented in Algorithm 2.

3.3.3 Doppler Frequency Estimation based on Signal Difference/-Correlation

Since H_s^m, H_s^{m+1}, H_d^m and H_d^{m+1} are time-invariant, two segment of signals of $R(t)$ separated by $n^\star T_s$ ($f_d n^\star T_s \approx 1$) will exhibit high correlation. In this subsection, we exploit such signal difference (or correlation) to estimate f_D.

Rewrite Equation (3.8) as

$$R(t) = a + \frac{b}{c + z(t)}, \tag{3.16}$$

where $b \triangleq (H_d^m H_s^{m+1} - H_s^m H_d^{m+1})/H_s^{m+1}$, $a \triangleq H_s^m/H_s^{m+1}$, and $c \triangleq H_d^{m+1}/H_s^{m+1}$. Here a, b, and c are also time-invariant. With $t_k = kT_s$, $R(t_k)$ can be simplified as $R(k) = a + b/(c + z(k))$.

Then, the square of the difference $\Delta_R^2(n)$ between two CSI-ratio samples $R(k)$ and $R(k+n)$ can be represented as

$$\begin{aligned} \Delta_R^2(k,n) &= |R(k+n) - R(k)|^2 \\ &= |b|^2 \cdot |\frac{e^{j(k+n)v} - e^{jkv}}{(c + e^{jkv})(c + e^{j(k+n)v})}|^2, \end{aligned} \tag{3.17}$$

where $v \triangleq 2\pi f_D T_s$. From Equation (3.17), we can see that when $e^{jn2\pi f_D T_s} = 1$, i.e., $n f_D T_s = 1$, $\Delta_R(k,n)$ is 0. This is the ideal case when a sequence of signals fully repeat themselves and noise is absent. In practice, it may not be exactly zero. To exploit the averaging effect, we instead compute the difference (or correlation) between two segment of signals spaced at different distances, and then look for the minimum (or maximum correlation) of the output. Using the difference, this can be presented as

$$n^\star = \arg_n \min \frac{1}{K} \sum_{k=k_0}^{k_0+K-1} \Delta_R^2(k,n), n = n_0, \cdots, N; \tag{3.18}$$

$$f_D = 1/(n^\star T_s), \tag{3.19}$$

where N does not have to be equal to K. Note that two pieces of $R(k)$ slightly separated may also exhibit high correlation as the phase differences caused by the Doppler frequency could be small. Therefore, we shall either start from a relatively large n_0 and then look for the first local minimum, or start from $n_0 = 1$ but look for the first "bottom" of the difference power (where

Algorithm 2 Doppler Frequency f_D Estimation based on Periodicity of CSI Ratio

Input: The angle of a sequence of CSI-ratio samples $\{\gamma(t_1), \gamma(t_2), ..., \gamma(t_N)\}$ within the time period $[t_1, t_N]$;

The minimum sample size S_{min} in one cycle;

The maximum sample size S_{max} in one cycle;

step I;

Output: The estimated f_D within $[t_1, t_N]$

Initialize $k = 1$

Initialize an empty list F_list

while dok $\leq N$ - S_{max}

    The angle sample $\gamma(t_k)$ is denoted as γ_{base}

    $j = k + 1$

    while do$j - k \leq S_{max}$

        if (**then**$\gamma(t_j) - \gamma(t_k))(\gamma(t_j + 1) - \gamma(t_k)) \leq 0$ and $S_{min} \leq (j - i)$

            $S = j - k$

            break

            $S = S_{max}$

        end if

        $j = j + 1$

    end while

    $f_D = \frac{1}{ST_s}$

    Add f_D to F_list

    $k = k + I$

end while

Average the f_D values in F_list, and f_D in $[t_1, t_N]$ is estimated to be the averaged value

return Estimated f_D within the time period $[t_1, t_N]$

curves go down first and then up). The latter is easier to implement without the trouble of determining the value of n_0 and is generally more reliable.

The detailed implementation of this method is illustrated in Algorithm 3.

Algorithm 3 Doppler Frequency f_D Estimation based on Signal Difference

Input: A sequence of CSI-ratio samples $\{R(n)\}_{n=1}^{N}$ within the time period $[t_1, t_N]$; The length k_0 of the reference sample dataset $\{R(1), R(2), ..., R(k_0)\}$; Pseudo-periodicity n_0

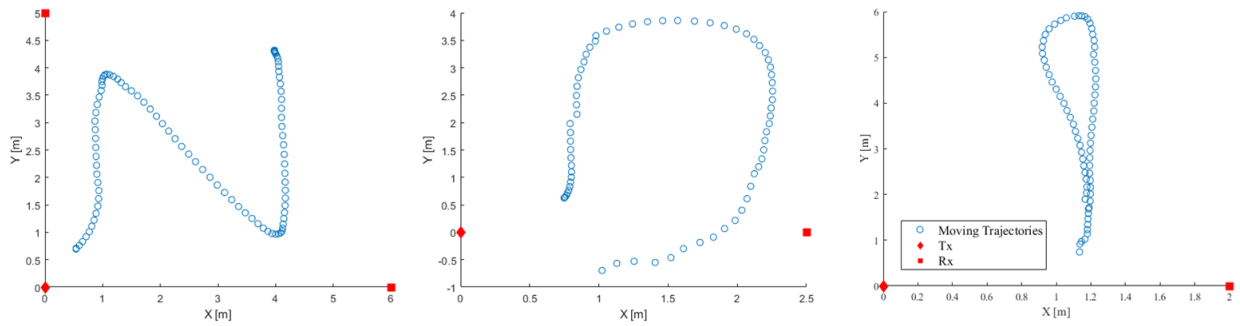
Output: The estimated f_D within $[t_1, t_N]$

Initialize an empty list D_list
while do$n_0 \leq n \leq N$
    Initialize $\Delta_s = 0$
    while do$k_0 \leq k \leq k_0 + K - 1 \leq N$
        $\Delta_R^2(k, n) = |R(k + n) - R(k)|^2$
        $\Delta_s = \Delta_s + \Delta_R^2(k, n)$
    end while
    Add Δ_s to D_list
end while
Find the value $n^\star$ from D_list that satisfies $n^\star = \arg_n \min \frac{1}{K} \sum_{k=k_0}^{k_0+K-1} \Delta_R^2(k, n)$
$f_D = \frac{1}{n^\star T_s}$
return Estimated f_D within the time period $[t_1, t_N]$

3.4 Experimental Results and Analysis

To verify the effectiveness of the proposed methods, we conduct experiments by employing the public WiFi CSI dataset *Widar* 2.0 [97]. The system uses one single transmitting antenna and three receiving antennas. The Intel 5300 WiFi card is used for CSI collection. The human motion channel CSI were collected in three environments: classroom, office and corridor. Several typical trajectories of human movements are shown in Figure 3.2.

We estimate the Doppler frequency of a human target moving in the three environments, using the three proposed algorithms, and compare the results with those obtained by the CACC method in *IndoTrack* [18]. The estimation results are illustrated in Figure 3.3 - Figure 3.5 for classroom, office and corridor, respectively. Specifically, for the CACC method, we perform



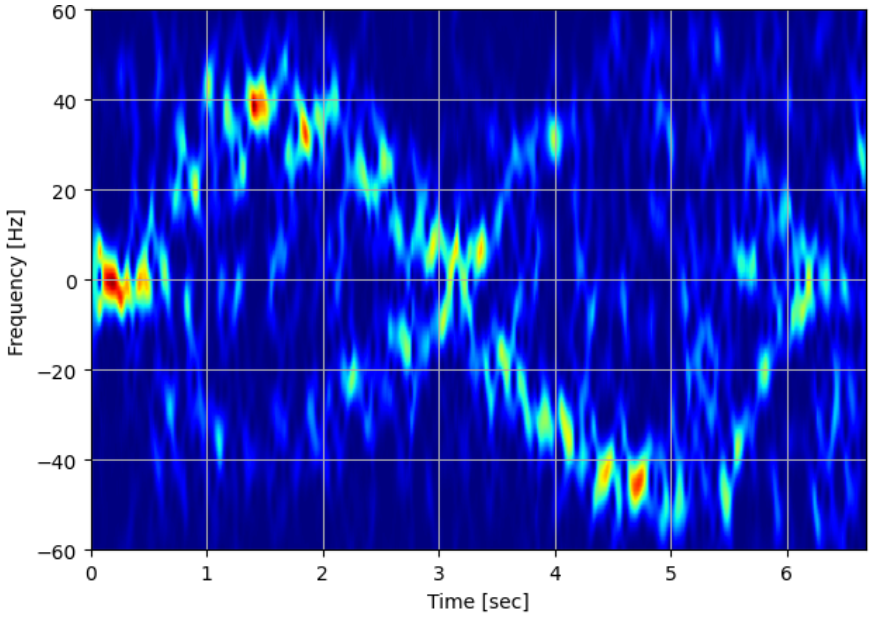
(a) Classroom (b) Office (c) Corridor

Figure 3.2: Three trajectories of human movements in three scenarios.

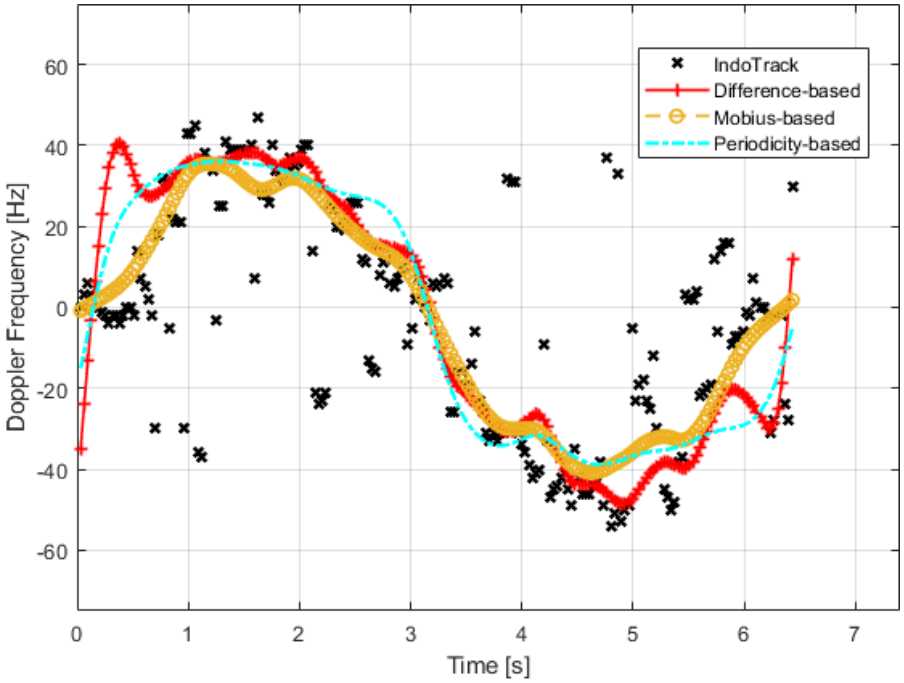
STFT on the CSI data after the antenna selection and the conjugate multiplication operations, and transform the data into a 2-dimensional time-Doppler frequency map, as shown in Figure 3.3(a) - Figure 3.5(a). Furthermore, at every time point, we select the frequency component that has the maximum power value, and treat these frequency components as the estimated Doppler frequencies, as shown in the black marks in Figure 3.3(b) - Figure 3.5(b). The sign of the Doppler frequency estimate in the Mobius-based algorithm is used as the sign in the other two algorithms.

From the estimated Doppler frequency results, we can see that the proposed three algorithms generally perform well, demonstrating much smaller fluctuations, compared to the CACC method in [18]. The image components are not effectively removed in [18], and cause large deviations in the estimates. Additionally, among the three methods, the signal difference-based algorithm achieves the best performance in the three scenarios, although it is incapable of estimating the sign of target Doppler frequency. The calibrated results of the difference-based method is also shown in Figure 3.6. Comparing with the results in Figure 3.3(b) and 3.4(b), we can see that with calibration, the performance is notably improved when the target moves at a low speed.

Overall, the performance of the Mobius-based algorithm is inferior to that of the difference-based one. Despite of this, the Mobius-based method is capable of estimating both the value and sign of a Doppler frequency. This is a very important feature. Furthermore, compared with the other two proposed algorithms, the periodicity-based method appears to achieve less accurate estimates. This is probably because unlike the other two algorithms, it does not explore the diversity effect of exploring multiple samples in the estimation and hence is susceptive to the noise. The advantage of this approach is that it is easy and simple to implement, and can be

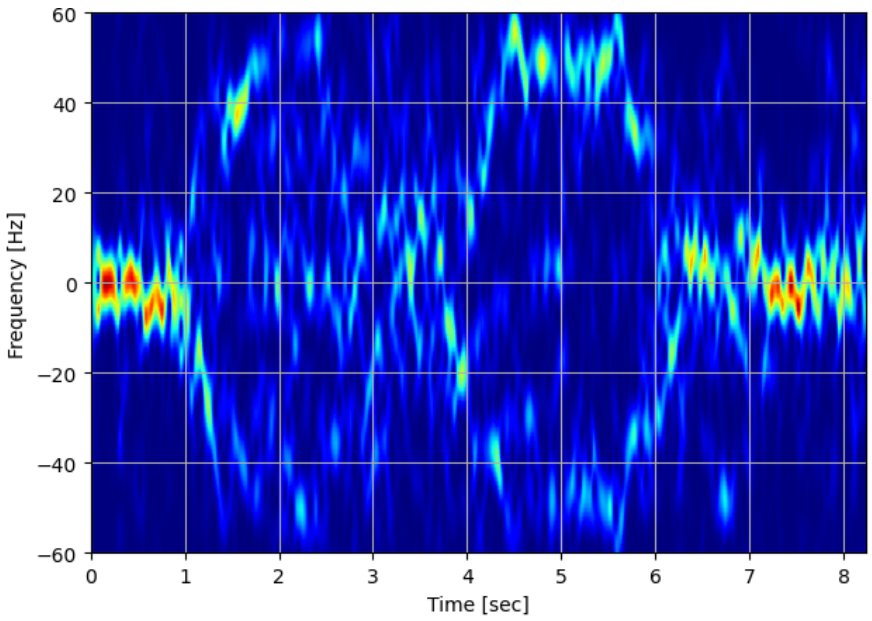


(a) The estimation results with the *IndoTrack* method [18].

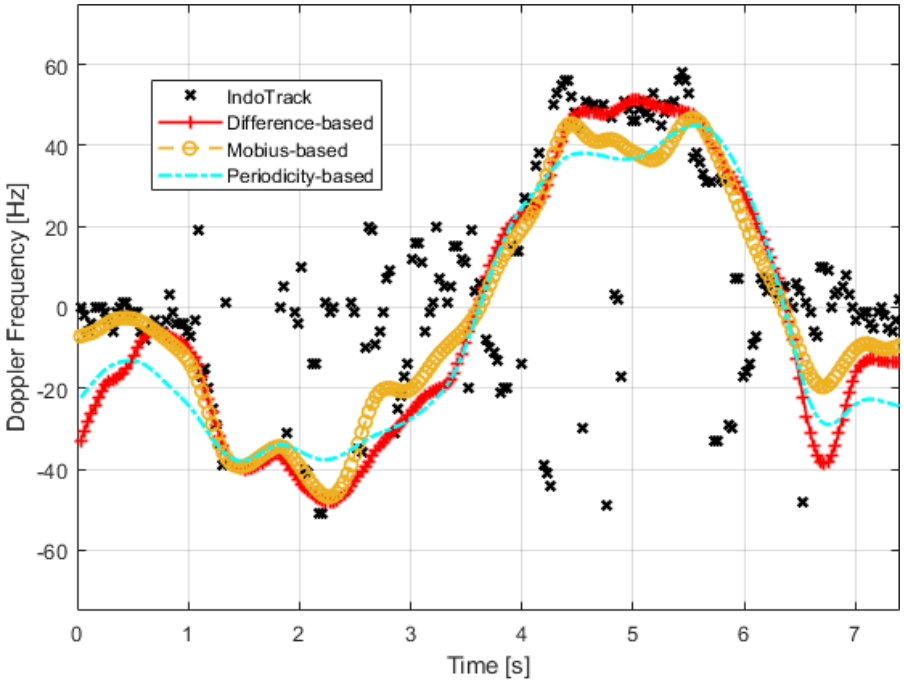


(b) The estimation results with the three proposed methods.

Figure 3.3: Doppler frequency estimation results in the classroom scenarios

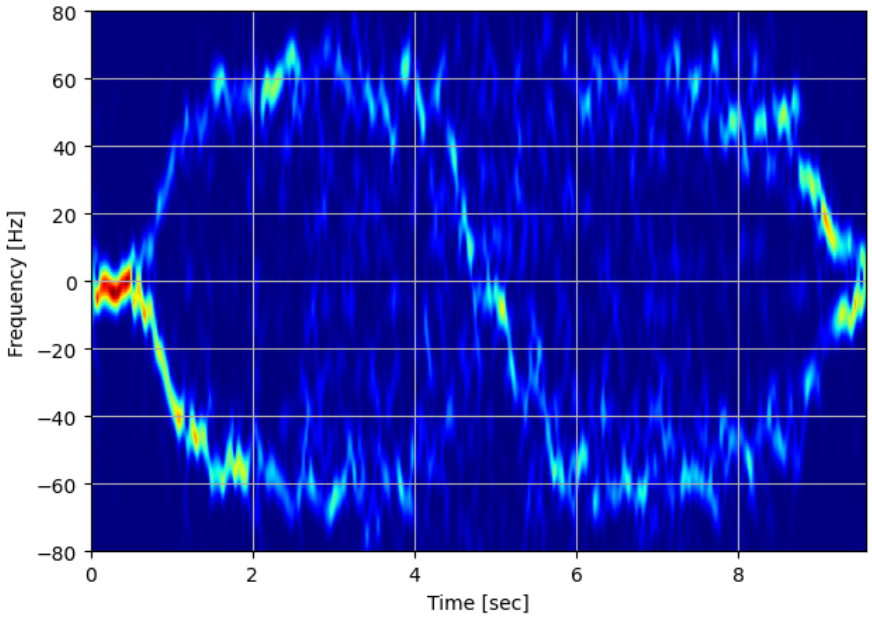


(a) The estimation results with the *IndoTrack* method [18].

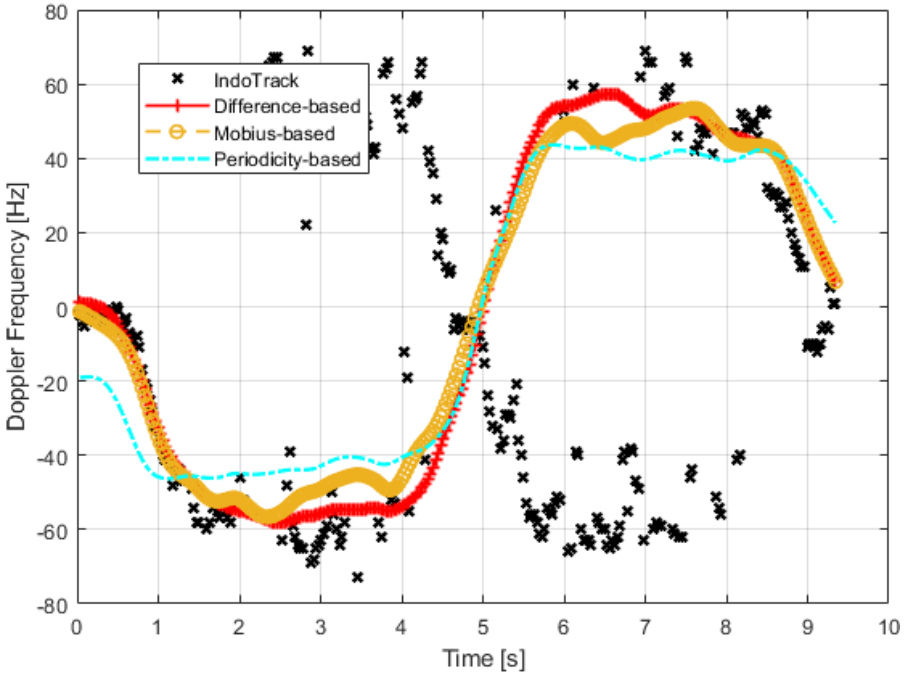


(b) The estimation results with the three proposed methods.

Figure 3.4: Doppler frequency estimation results in the office scenarios



(a) The estimation results with the *IndoTrack* method [18].



(b) The estimation results with the three proposed methods.

Figure 3.5: Doppler frequency estimation results in the corridor scenarios

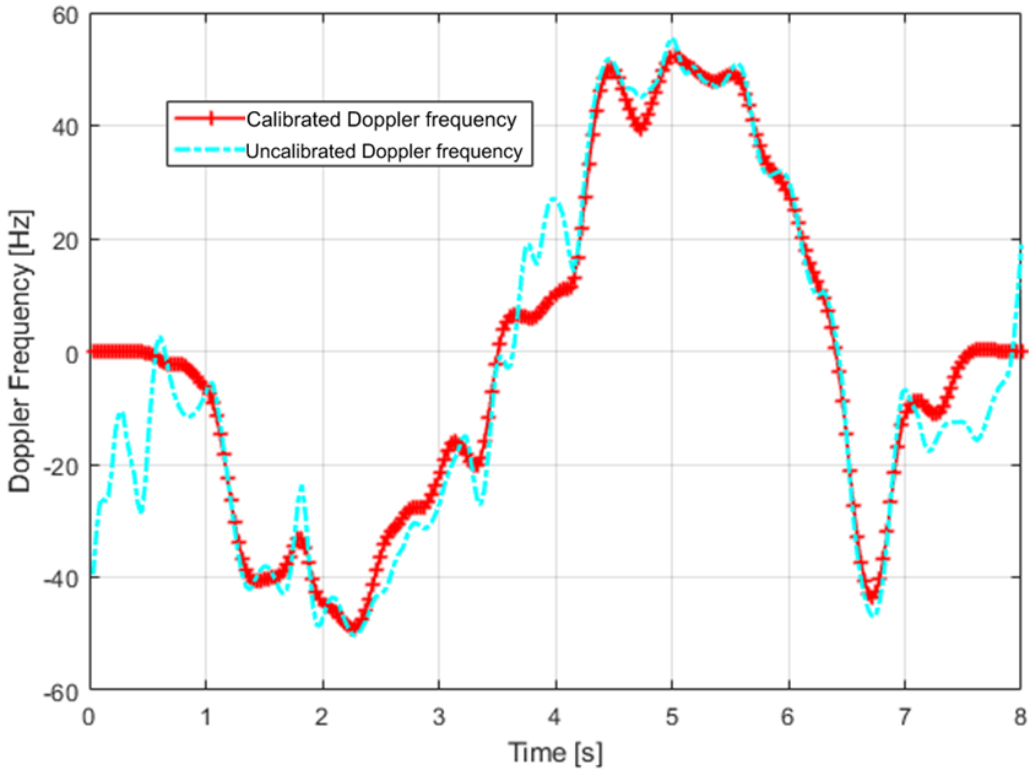


Figure 3.6: The estimated Doppler frequency in the office scenario with the calibrated difference-based approach.

used for a rough or initial estimation of target Doppler frequency.

3.5 Summary

In this chapter, we have proposed three Doppler frequency estimation algorithms based on the CSI ratio across antennas for applications involving sensing of moving targets, such as human activity recognition and mobile tracking. Among them, the signal difference-based algorithm has the best and the most robust performance for Doppler frequency estimation, while the periodicity-based algorithm is the easiest one to implement. However, these two algorithms cannot estimate the sign of Doppler frequency. In contrast, the Mobius-based method can evaluate both the sign and value of Doppler frequencies. As a result, the best solution may be to combine the strengths of the three algorithms. The current work may also be the cornerstone for estimating other moving parameters based on the CSI-ratio model, e.g., time delay and AoA of human targets.

Chapter 4

Cross-target HAR with Limited Radar Micro-Doppler Signatures

4.1 Introduction

Due to the human individual discrepancies, such as the differences in appearances and behaviors, the measurements of the same activity from different persons are generally diverse. When using a trained deep learning (DL) model to recognize the activities from various persons, the HAR performance of this model can be different. Furthermore, when a DL model trained with the activity data of a known person is applied to identify a new person's activity, the performance of this model generally degrades. TL, which utilizes prior knowledge to make a trained model generalize well on new tasks, is one of the potential solutions for cross-target HAR [5]–[7]. For instance, Park et al. [78] presented a deep convolutional neural network (CNN) pretrained on *ImageNet*, and fine-tuned the network with measured radar MD spectrograms for human aquatic activity classification. Seyfioğlu et al. [6] proposed a residual learning model *DivNet* trained on the simulated radar MD spectrogram dataset, and fine-tuned the model with a measured dataset to classify seven human activities. The fine-tuning (FT) strategy used in these methods utilizes the target data to fine-tune the pretrained DL models, and transfers the source knowledge to compensate for the insufficiency of target domain data. We refer this strategy to the *Conventional FT*.

However, the performance of *Conventional FT* approaches often degrades when the amount of labelled data drops. Furthermore, the catastrophic forgetting effect [8] (the tendency of

DL models to abruptly forget previously learned tasks after being trained for a new task) usually occurs in the *Conventional FT*. In other words, when the model fine-tuned on the target dataset is applied to classify the persons' motions in the source dataset, the performance usually decreases. As a result, the *Conventional FT* method often lacks generalization, and cannot scale well to the persons in different domains simultaneously.

In this chapter, aiming at enhancing the generalization ability of DL-based HAR on human individual differences and improving the HAR performance in different persons' activities, we propose a novel instance-based TL approach *ITL* for cross-target activity recognition.

The overall flow of the proposed *ITL* is shown in Figure 4.1. Firstly, we design a deep CNN *MNet* for radar-based HAR as the backbone of *ITL*, and pretrain it with all available source data (see Figure 4.1(a)). At the same time, a correlated source data selection (CSDS) algorithm is designed to pick up partial instances from the source domain as supplements for the target data (see Figure 4.1(b)). Then, an adaptive collaborative fine-tuning (ACFT) algorithm (see Figure 4.1(c)) is presented to fine-tune the pretrained *MNet* with the whole target dataset and the selected source data. With ACFT, *ITL* can perform the target task while retaining partial source knowledge. This property allows the fine-tuned *MNet* to be used for classifying the target activity data and accurately identifying the activities in the source dataset. In other words, *ITL* is more generalized to cope with the data distribution discrepancy between the two domains, which is often caused by human activity differences.

The major contributions of this chapter can be summarized as follows.

1) We propose an instance-based TL approach '*ITL*' for radar-based cross-target activity recognition with limited training data. The proposed approach can generalize well to human activity differences, and achieve good performance when used to recognize the activities of diverse persons.

2) *ITL* is a unique algorithm that consists of three interconnected parts, including DL model pretraining, correlated source data selection and adaptive collaborative fine-tuning. Any of the three components cannot be excluded; otherwise, the performance of the entire algorithm for HAR decreases.

3) The experimental results demonstrate that *ITL* has good performance for recognizing the activities of six persons with limited radar data, outperforming several state-of-the-art HAR

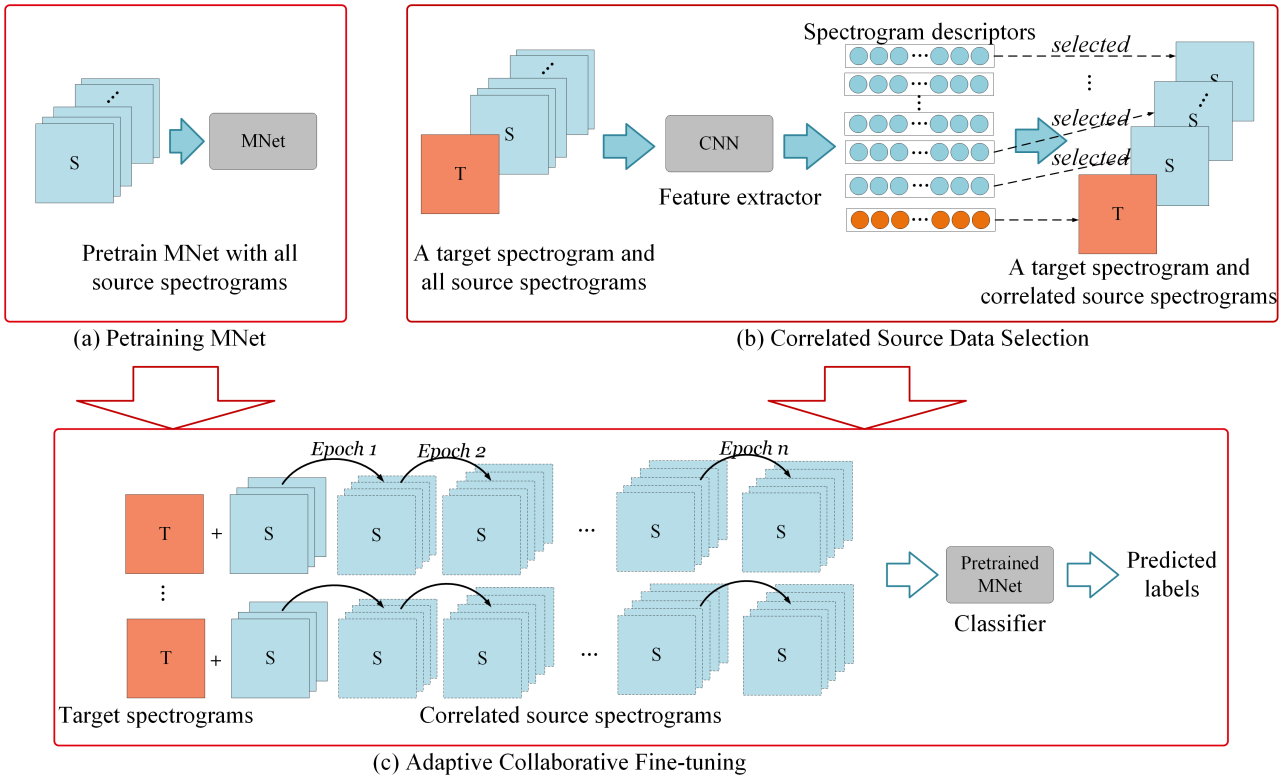


Figure 4.1: The pipeline of the proposed *ITL* method for cross-target HAR.

methods. Furthermore, though trained for the classification task in the target dataset, *ITL* can still recognize different persons' activities in both the source and target domains.

The rest of this chapter is organized as follows. Section 4.2 describes the measured data collection and preprocessing process and presents some data analysis. Section 4.3 introduces the structure of *ITL*. Section 4.4 presents the analysis and discussion of the experimental results. Furthermore, some ablation studies on *ITL* are performed in Section 4.5. Finally, Section 4.6 concludes this paper.

4.2 Data Collection, Preprocessing and Analysis

4.2.1 Data Collection

We utilize a UWB radar *PulsON* 440 for the experiments. *PulsON* 440 is composed of two antennas for transmitting and receiving C-band radio signals. The waveform generator generates chirp signals with a bandwidth of 1.8 GHz and a centre frequency of 4.0 GHz. The UWB radar can distinguish the main scattering points of the human target due to its high range resolution. The SNR of the received signals can be improved by accumulating the echo signals of multiple

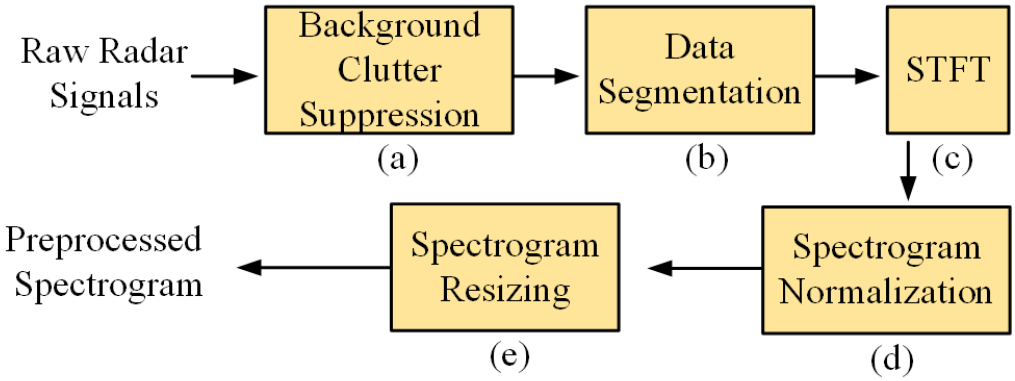


Figure 4.2: The pipeline of radar raw signal preprocessing. (a) MTI for background clutter suppression. (b) Data segmentation. (c) STFT. (d) Data normalization. (e) Resizing spectrograms.

strong scattering points. Thereby, the target recognition ability of the UWB radar is enhanced.

The experiments are conducted in an indoor environment. The radar is set at the height of 1 m, and six activities (M1: running forward, M2: running in a circle, M3: jumping ahead, M4: sitting on a chair, M5: walking along, M6: boxing in place) are performed by six persons in the line-of-sight of the radar with an aspect angle of 0 degrees. All the subjects are limited to moving within the range from 1.5 m to 7.5 m. Each of the six activities is continuously performed by an individual for approximately 1.5 minutes. And in each scenario, the process is repeated one to three times. The basic physical information of the six subjects is listed in Table 4.1.

Table 4.1: Basic physical information of the six subjects

	Sub ♯1	Sub ♯2	Sub ♯3	Sub ♯4	Sub ♯5	Sub ♯6
Age	23	25	23	23	23	24
Height (cm)	173	178	172	166	188	169
Weight (kg)	73	71	75	66	92	52

4.2.2 Data Preprocessing

In this study, we employ MD spectrograms as input to the network, treating HAR as a spectrogram classification problem. Figure 4.2 illustrates the radar data preprocessing process.

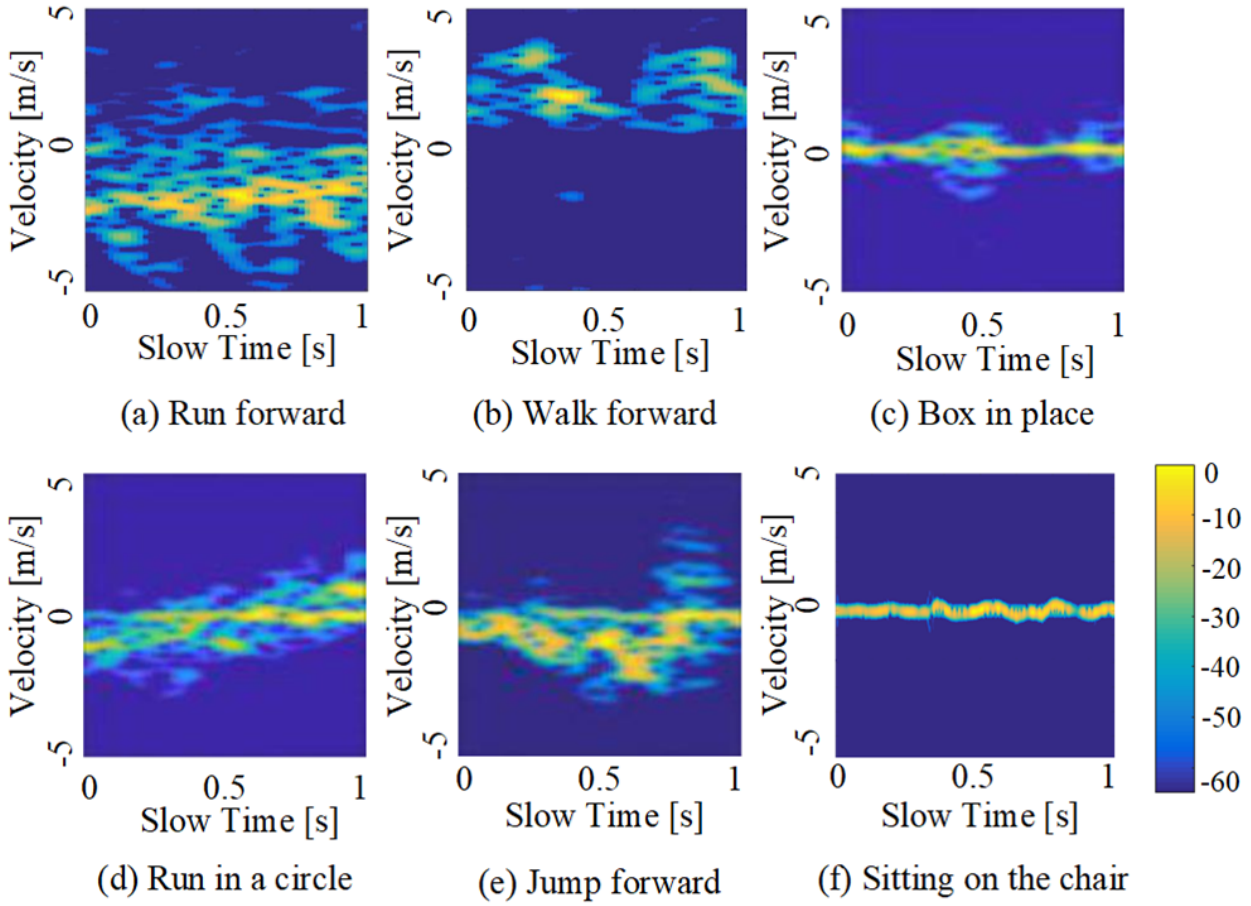


Figure 4.3: Several typical MD spectrograms of human activities.

Firstly, the moving target indicator (MTI) is adopted on the raw radar echo signals to remove the static background clutter. Next, the processed radar data is divided into several segments of 1 s so that there is an approximately complete cycle of each of the six activities. The overlap between adjacent segments is 0.36 s.

Based on this, a 1024-point STFT is used to process these data segments. Since the human bodies are distributed targets, the scattered data from the bodies are spread over a few range cells. Thus, the STFT is performed on the radar data that are summed over several resolution cells. The obtained 2-dimensional radar data after the STFT is still complex-valued, and the modules of the 2D complex data are utilized to form the 2D spectrograms.

Then, we normalize all the spectrograms to make the values in the spectrograms fall into [0,1]. Data normalization can prevent the value of a particular dimension from being too large. In this way, the convergence of DL models can be facilitated [98]. Finally, the spectrograms are resized into 150 $\times$ 150 pixels for further processing. The radar MD dataset has 300 spectrograms per

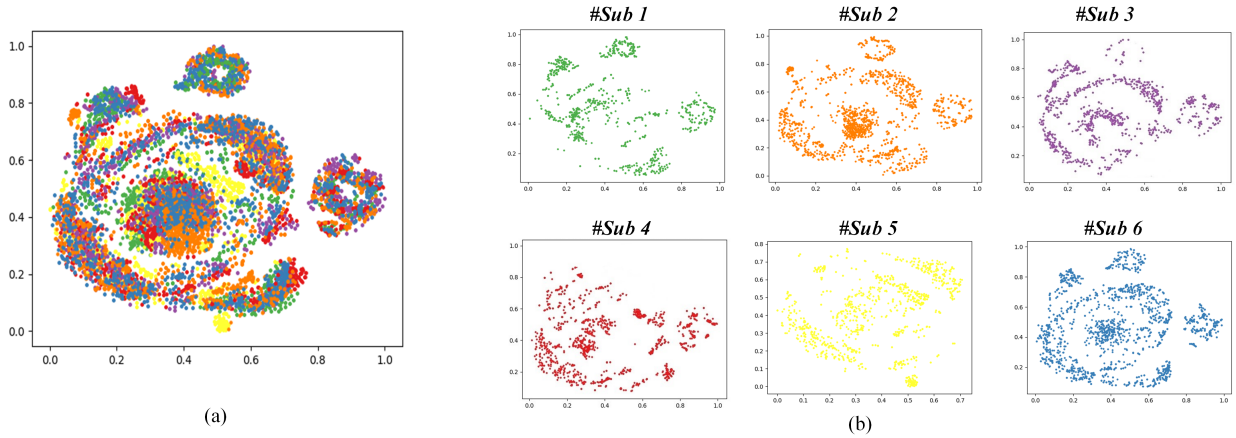


Figure 4.4: Visualization results of the whole spectrogram dataset with t-SNE. (a) The distribution of all activity data of the six persons. (b) The distributions of the six persons' activity data, separated for each individual.

person per activity. Several typical preprocessed spectrograms are shown in Figure 4.3.

4.2.3 Data Analysis

Due to the human individual activity differences, different persons' activity data often have some discrepancies and are varied in distribution. In this circumstance, when a DL model trained with several persons' activity data is directly applied to recognize the activities of new persons, the model's performance often decreases.

To show the differences in the distributions of the six persons' activity data, we reduce these activity data to a series of two-dimensional vectors and visualize the dimensionality-reduced data with t-Distributed Stochastic Neighbor Embedding (t-SNE) [99]. The visualization results are shown in Figure 4.4. It can be seen that, though related, the distributions of the six persons' activity data are different, indicating individual activity differences between them. Furthermore, a quantitative similarity comparison between the six persons' activity data is also performed. In detail, we assume the six persons' activity data follow independent multivariate Gaussian distributions. Then, the KL divergence from the activity data distribution of one person to the others can be calculated. The KL divergence $KL(p||q)$ is shown in Figure 4.5, where p and q are the probability distributions of the activity data of any two of the six people. It can be seen that when the KL divergence between the two distributions is slight, the similarity between the data is relatively high.

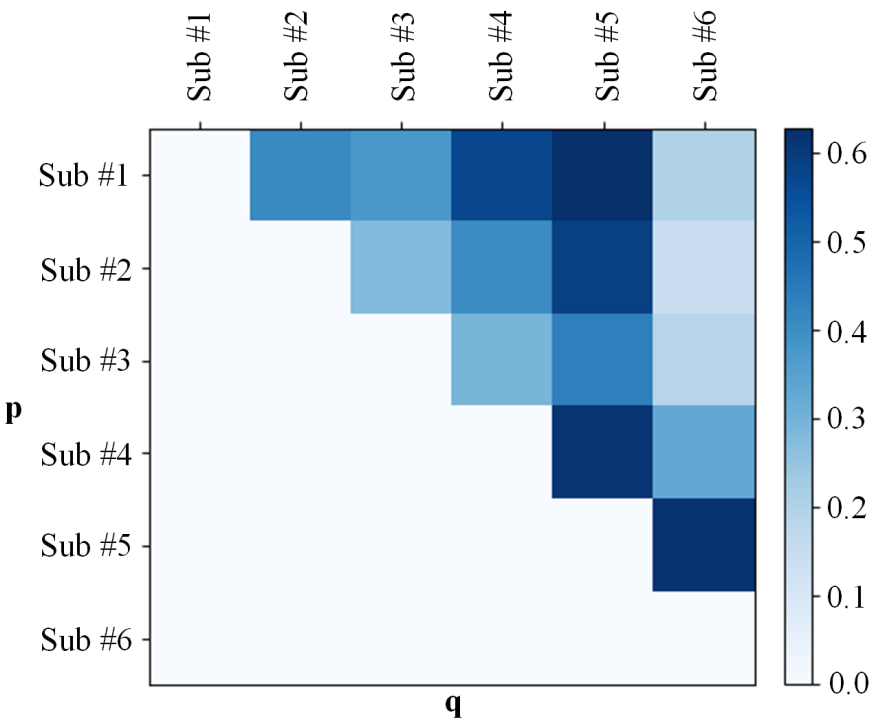


Figure 4.5: KL divergence $KL(p||q)$ between the activity data of one person to the others. p and q are the probability distributions of the activity data of any two of the six people.

4.3 Description of ITL

In this section, we introduce the algorithmic components of our proposed HAR approach, *ITL*, in detail.

4.3.1 Problem Formalization

Mathematically, the problem is described as follows. Let the source domain training dataset $D_s = \{x_i^{(s)}, y_i^{(s)}\}_{i=1}^{N_s}$, where there are N_s data in the source domain. $x^{(s)} \in \mathbf{R}^{m \times n}$ denotes an $m \times n$ matrix corresponding to the radar MD signature of the human activities in the source domain. $y^{(s)}$ denotes the corresponding label of C_s categories. A source classification network $f_s(\cdot)$ is trained with D_s from scratch. Let the target dataset $D_t = \{x_i^{(t)}, y_i^{(t)}\}_{i=1}^{N_t}$, where $x^{(t)} \in \mathbf{R}^{m \times n}$ denotes an $m \times n$ matrix corresponding to the radar MD signature in the target domain. The data in the target domain belong to the same C_t categories as those in the source domain. However, there is a distribution discrepancy between D_s and D_t, which makes $f_s(\cdot)$ not suitable to classify the target data. Furthermore, there are only a limited number of instances in D_t, which are insufficient to train a sufficiently generalized classification model.

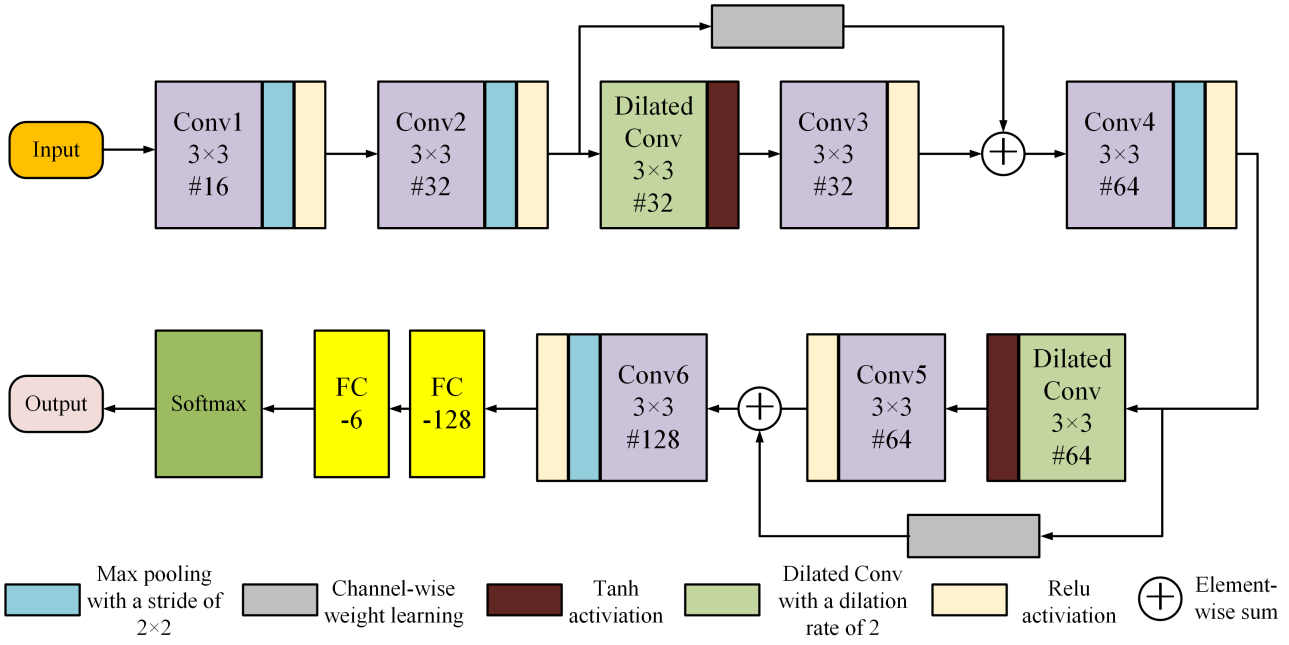


Figure 4.6: The architecture of the proposed backbone (*MNet*) for HAR. The proposed *MNet* is composed of six convolutional layers, two dilated convolutional layers, two channel-wise attention layers and two fully connected layers.

Our goal is to train a target classification network $f_t(\cdot)$ to recognize the activities accurately in D_t when there is limited target training data. To this end, an instance-based deep TL approach *ITL* is presented. The proposed *ITL* transfers the relevant knowledge from the sufficient activity data (*source domain*) as a supplement to classify the activity data in a new dataset (*target domain*). The details of our proposed algorithm are summarized in Algorithm 4.

4.3.2 Structure of the pretrained Deep Model

In this paper, we design a deep neural network *MNet* for radar-based HAR and use it as the backbone of *ITL*. In radar spectrograms, each pixel of the spectrogram has both an intensity and a sample of time and frequency values, distinguishing it from optical images. Due to the unique properties of MD signatures, the proposed DL approach is designed to be more tailored to the radar data. The architecture of *MNet* is illustrated in Figure 4.6.

As shown in this figure, convolutional layers, together with max-pooling, are the basic components of the network. Furthermore, to extract more discriminative features from the MD signatures, we apply the dilated convolution mechanism and the channel-wise attention mechanism within *MNet*. Then, two fully-connected layers are connected with the last convolutional

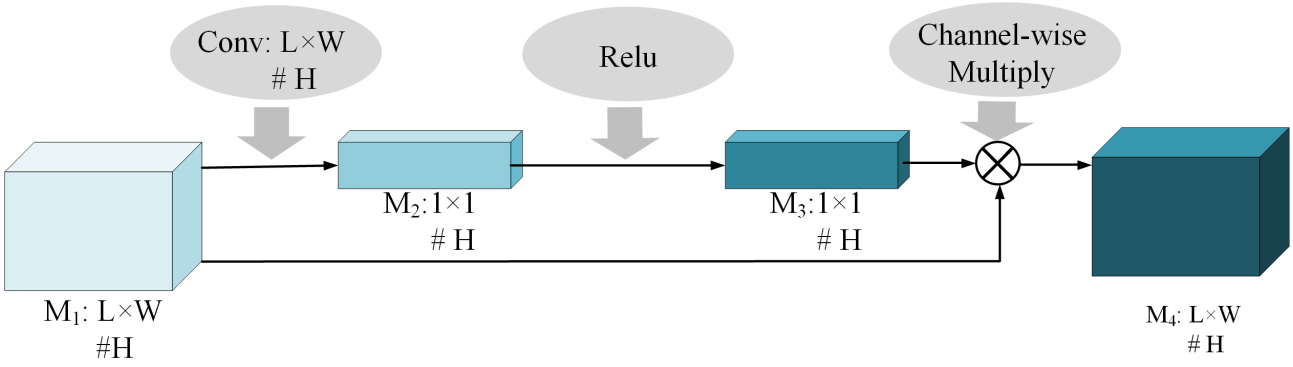


Figure 4.7: Illustration of the channel-wise attention mechanism. M_1 represents the input feature maps of $L \times W$ from H channels. So M_2, M_3, M_4. *Conv* represents the convolution with H kernels of $L \times W$.

layer sequentially. The softmax function is employed at the end of *MNet* to predict the labels of the input spectrograms.

- **Channel-wise Attention**

The channel-wise attention mechanism enhances the network performance by accounting for the different importance that each feature channel has in the classification process. The more helpful feature channels are weighted accordingly to emphasize their contribution, and the other way round for less important feature channels [100]. By explicitly modeling the channel interdependencies and recalibrating the features, the proposed network is more focused and oriented to the more informative data.

The channel-wise attention mechanism is illustrated in Figure 4.7. Firstly, the feature maps M_1 from H channels are fed into a convolutional layer. To obtain the importance of every channel, the convolutional layer is designed with H kernels with the size of $L \times W$, which has the same size as the input feature maps. Hence, the output feature maps M_2 are H real numbers and have a global receptive field. Next, the 1×1 feature maps are excited with an activation function, and the output values (feature maps M_4) are treated as the weights of importance corresponding to these channels. Finally, the channel recalibration is completed by multiplying the weights with the original feature maps M_1 channel-by-channel. In this way, the initial M_1 is transformed into the weighted feature maps. And the channels with larger weights are paid more attention.

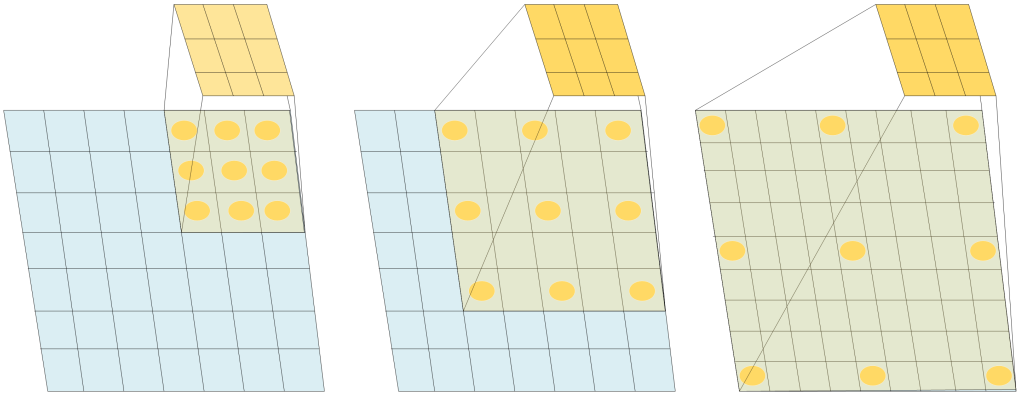


Figure 4.8: Dilated Convolution with different dilation rates. The blue area is the input feature map, and the yellow area is the convolution kernel. The pale yellow area is the receptive field. The yellow dots are the pixels that are convolved with the convolution kernel.

- **Dilated Convolution**

In CNNs, pooling is utilized to decrease the redundancy of the feature maps and enlarge the receptive fields. The receptive field is the size of the activation area on the feature map during a convolution operation. However, pooling has many drawbacks, such as missing spatial and small-object information. For example, when there are three pooling layers with a kernel of 2 $\times$ 2, the knowledge of the objects smaller than 8 $\times$ 8 is lost.

To tackle this problem, a dilated convolution [101] is adopted in this paper. Instead of down-sampling, dilated convolution is achieved by zero paddings on the convolution kernels, as shown in Figure 4.8. This mechanism allows the dilated convolution to increase the receptive fields without losing the structured data information. The size of the receptive field is proportional to a parameter called dilation rate due to the number of zero padding increasing as the dilation rate increases. When the dilation rate is set to 1, dilated convolution is equivalent to the conventional convolution. The dilated convolution operation can retain more valuable information of the input without increasing the parameters of the network and helps obtain more globally representative details on the original data.

4.3.3 Correlated Source Data Selection

In this section, we propose a novel correlated source data selection (CSDS) algorithm to select the most appropriate data for the collaborative fine-tuning. Instead of only using the target

data to fine-tune the pretrained network, we make a partial selection of source data with high similarity to the target data, and utilize them to fine-tune *MNet*, along with all target data.

Algorithm 4 *ITL*: An Instance-based TL Method for HAR with Limited Radar Data

Input: motion network *MNet*, a source dataset $D_s = \{x_i^{(s)}, y_i^{(s)}\}_{i=1}^{N_s}$, a small-scale labeled target dataset $D_t = \{x_i^{(t)}, y_i^{(t)}\}_{i=1}^{N_t}$, and number of epochs I_t

Output: the fine-tuned *MNet* for classifying unlabeled target data

Pretraining *MNet* with D_s
For the kth sample, refining $\{h_k\}_1^M$, and obtaining $\mathbf{H}^k = \{\mathbf{h}_0^{'},\mathbf{h}_1^{'},...,\mathbf{h}_{255}^{'}\}$
Calculate $EMD(P,Q) = \min_{F=\{f_{ij}\}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}}$
$i \Leftarrow 0$
For each spectrogram $x^{(t)}$ in D_t, selecting source spectrograms based on Equation (4.1)
while not converged **or** $i < I_t$ **do**
    Calculating w_j in Equation (4.3)
    if $i \leq 5$ **then**
        $E = -\sum_{c=1}^{C_t} p_c log(p_c)$
        Adjusting N^{i+1} in Equation (4.6)
        Update w_j
        $i = i + 1$
    end if
end while
return The fine-tuned parameters $w^{(m)}$, $b^{(m)}$ of *MNet* (m=1,2,...,M and M is the number of layers to be optimized in *MNet*)

• MD Signature Descriptor

As efficient feature extractors, deep convolutional neural networks can learn the high-level semantic representation of the input data. And the representation can be used for describing input data. In this paper, we utilize *AlexNet* [102], a typical convolutional neural network for image classification, to attain the descriptors of the input MD signatures. It is noted that instead of *AlexNet*, many other CNNs such as VGG-Net, ResNet, Inception-Net can also implement this function. We select *AlexNet* because it can extract semantic information effectively, and its structure is relatively simple. In detail, we treat the convolution kernels of the last convolutional layer in *AlexNet* as filters $\{F_0,F_1,F_2,...,F_{255}\}$. All data in D_s and D_t are input

to a pretrained *AlexNet*. Then, the feature maps output from the last convolutional layer are represented in histograms corresponding to the input MD signature. Let $M_i(x, y)$ denotes the output feature map of the ith filter F_i, and $\mathbf{h}_i$ its histogram, where i={0,1,2,...,255}. In the beginning, the pixel value range of all histograms is set from 0 to 255, and the width of every histogram bin is set to 0.5.

To obtain more discriminative descriptors, we refine the histograms and avoid a large percentage of pixels falling into the same bin. Specifically, we first obtain the maximum pixel value p_{max}^{u} and the minimum pixel value p_i^{min} of $M_i(x, y)$ by scanning the whole ith feature maps in the source dataset D_s. Then the pixel value range of $\mathbf{h}_i$ is set from p_i^{min} to p_{max}^{u}. Furthermore, we iterate through the original $\mathbf{h}_i$s of D_s and D_t, and adaptively set the width of the histogram bins so that there is a roughly equal percentage of pixels in each bin. The percentage is set to 2% so that there are no more than 50 bins in every type of histograms $\{\mathbf{h}_i\}_{i=0}^{255}$. This setting makes a compromise between computing complexity and representation efficiency, which allows the further designed descriptor to have a proper dimension and be discriminative simultaneously.

In this way, the inhomogeneous intervals are acquired, and the new ith histogram $\mathbf{h}_i^{'}$ of a spectrogram is obtained. Figure 4.9 illustrates the refined histograms of a radar spectrogram corresponding to a filter F_i. Finally, for the spectrogram x_k, the corresponding histograms $\{\mathbf{h}_i^{'}\}_{i=0}^{255}$ are concatenated to form an MD signature descriptor, namely, $\mathbf{H}^k = \{\mathbf{h}_0^{'},\mathbf{h}_1^{'},...,\mathbf{h}_{255}^{'}\}$.

• Similarity Metrics of MD Signatures

The Earth Mover's Distance (EMD) [103] is the minimal cost that must be paid to transform one distribution into another distribution. It is proposed based on the solution to a typical transportation problem, but can be used to measure the distance between two generalized distributions irrespective of the underlying application. As a similarity metric of two histograms, EMD is more efficient than other possible histogram matching techniques due to its feasibility of operating on variable-length representations of the distributions.

A histogram can be formulated as a set $S = \{s_j = (w_j, m_j)\}_{j=1}^{N}$, where the histogram values are denoted as the weights w_j, and the indices of bins are denoted as positions m_j. N denotes the number of bins in the histogram. Given two histograms $P = \{(p_i, u_i)\}_{i=1}^{m}$ and $Q = \{(q_i, v_i)\}_{i=1}^{n}$, with size m, n respectively, the EMD of P and Q is defined as the minimum work required to resolve the supply-demand transports, namely

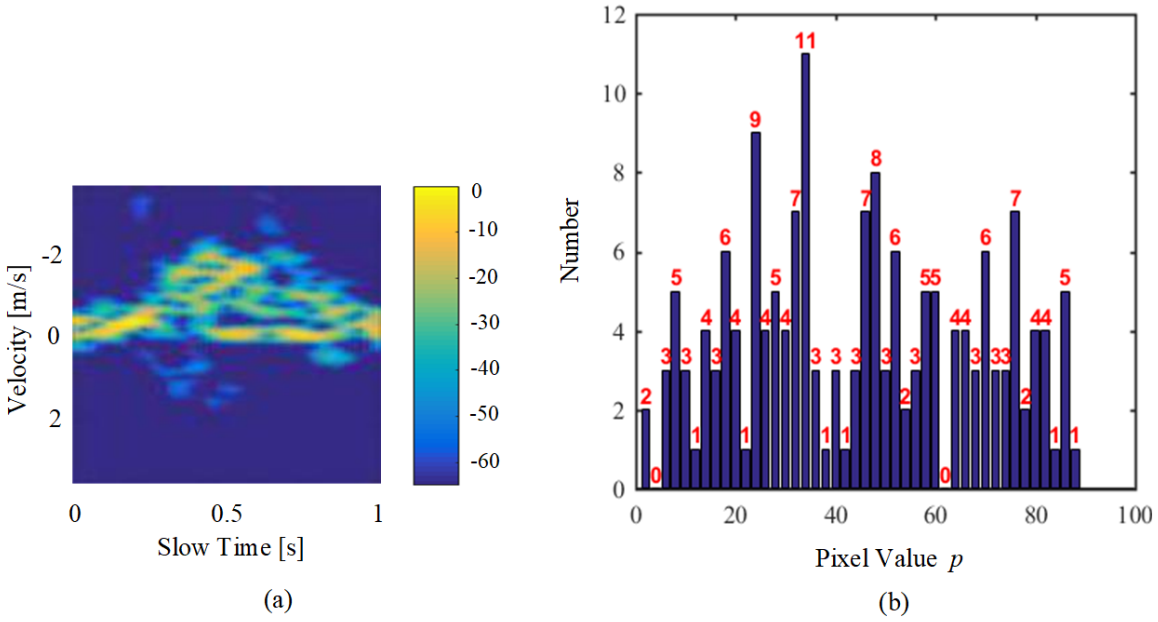


Figure 4.9: A typical spectrogram and its histogram. (a) A radar spectrogram. (b) The histogram corresponding a specific convolution kernel in the last convolutional layer of *AlexNet*.

$$EMD(P,Q) = \min_{F=\{f_{ij}\}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}} \tag{4.1}$$

with the constrains:

$$\begin{aligned} &\sum_j f_{ij} \le p_i, \qquad \sum_i f_{ij} \le q_j, \\ &\sum_{i,j} f_{ij} = min\{\sum_i p_i, \sum_j q_j\}, \qquad f_{ij} \ge 0, \end{aligned} \tag{4.2}$$

where p_i represents the histogram values of P, and q_i represents the histogram values of Q. u_i and v_i represent the indices of bins of P and Q, respectively. Furthermore, $F = \{f_{ij}\}$ denotes a flow set. Each flow f_{ij} represents the amount transported from the ith supply to the jth demand. d_{ij} denotes the distance between the position u_i and v_j.

In this paper, EMD is employed to measure the similarity of the histogram descriptors **H**s corresponding to the MD signatures in D_t and D_s. Given the histogram descriptors **H**s of a specific target spectrogram $x^{(t)}$ and all source spectrograms $\{x_i^{(s)}\}_{i=1}^{N_s}$, the EMDs $\{EMD_i\}_{i=1}^{N_s}$ of $x^{(t)}$ and $\{x_i^{(s)}\}_{i=1}^{N_s}$ are calculated with $\mathbf{H}^{(t)}$ and $\{\mathbf{H}_i^{(s)}\}_{i=1}^{N_s}$ according to Equation 4.1 and 4.2. A small EMD value between $x^{(t)}$ and $x^{(i)}$ means that these two spectrograms are highly correlated.

Furthermore, the source radar spectrograms with smaller EMD values are preferred as the correlated instance of $x^{(t)}$. In detail, the source spectrograms are ranked in ascending order based on their EMD values. For each target spectrogram, K source spectrograms (corresponding to the top 2.0% of the whole set of source spectrograms) are chosen as the most correlated set at the outset, based on the EMD metric. This ensures that the initially selected spectrograms are the most similar to the target spectrogram. Then, these instances are utilized to fine-tune the pretrained *MNet* along with all the target data.

4.3.4 Adaptive Collaborative Fine-tuning

• Source Instances Re-weighting

During fine-tuning the pretrained *MNet*, a series of source spectrograms are selected as the correlated instances by more than one target spectrogram. Compared with treating them equally, attaching more importance to the selected source instances more than once can make the fine-tuning process more efficient. Thus, the importance of the selected source instances differs. Specifically, we emphasized the loss function and re-weight the fine-tuning loss of the selected source instances. Suppose that in an epoch, the ith source instance $x_i^{(s)}$ is selected as the correlated instance by w target spectrograms. Then we design the loss function L of fine-tuning as follows,

$$L = \sum_i L_{cls}(y_i^{(t)}, \hat{y}_i^{(t)}) + \sum_j sin(\frac{\pi}{2} * \frac{w_j}{w_{max}}) * L_{cls}(y_j^{(s)}, \hat{y}_j^{(s)}) \tag{4.3}$$

where w_j denotes the number of target instances that select $x_j^{(s)}$ as its correlated source instance. w_{max} denotes the maximum among ws corresponding to all the selected source instances in an epoch. $y_i^{(t)}$ and $\hat{y}_i^{(t)}$ are the true label and the predicted label of the ith target instance $x_i^{(t)}$, respectively. Similarly, $y_j^{(s)}$ and $\hat{y}_j^{(s)}$ are the true label and the predicted label of the jth source instance $x_j^{(s)}$. The classification loss L_{cls} adopts the cross-entropy loss, whose definition is as follows:

$$L_{cls} = -[plog(\hat{p}) + (1-p)log(1-\hat{p})] \tag{4.4}$$

where p and $\hat{p}$ are the ground-truth one-hot label and the predicted probability, respectively.

• **Adaptive Source Data Search**

Subsequently, we present the adaptive searching scheme to employ more nearest source spectrograms in the following fine-tuning epochs, which is able to facilitate the target spectrograms classification. We calculate the information entropy E_i^m to measure the classification uncertainty of the target training sample $x_i^{(t)}$ after the mth epoch.

$$E_i^m = -\sum_{c=1}^{C_t} p_{i,c}^m log(p_{i,c}^m), \tag{4.5}$$

where C_t is the number of activity categories in D_t, $p_{i,c}^m$ is the probability that $x_i^{(t)}$ is classified as the cth class by the softmax layer of *MNet* in the mth epoch. The larger E_i^m, the higher classification uncertainty of $x_i^{(t)}$. We set the threshold θ for the classification uncertainty E. When E_i^m is larger than θ, we increase the number of correlated source samples for $x_i^{(t)}$ in the next epoch.

Furthermore, we stop the adaptive searching scheme after five fine-tuning epochs because too many epochs can lead to more source instances that are not highly correlated with the target data employed in the fine-tuning process. The overall adaptive source data search then is given as:

$$N_i^{m+1} = \begin{cases} N_i^m + \alpha, & m \leq 5 \ \ and \ \ \hat{y}_i^{(t)} \neq y_i^{(t)} \\ N_i^m + \beta, & m \leq 5 \ \ and \ \ \hat{y}_i^{(t)} = y_i^{(t)} \\ & and \ \ E_i^m \geq \theta \\ N_i^m, & others \end{cases} \tag{4.6}$$

where N_i^m and N_i^{m+1} are the number of the selected nearest source samples for $x_i^{(t)}$ in the m-th and m+1-th epochs, respectively. $N_i^1 = K$. $\hat{y}_i^{(t)}$ and $y_i^{(t)}$ are the predicted label and the true label of $x_i^{(t)}$, respectively. α and β are set to $K/2$ and $K/4$. θ is set to 0.25 empirically so that every target spectrogram tends to be classified into a particular category with high probability.

4.4 Experimental Implementation and Results

4.4.1 Evaluation Methodology

In the experiments, a leave-two-individual-out cross-validation method is adopted to split the dataset into the source and target datasets. Specifically, we randomly select the activity data

of (n-2) persons as the source dataset D_s, where n equals 6 in the experiments. And the data of the other two persons are utilized as the target dataset D_t. Hence, the process is repeated $\binom{n}{2} = 15$ times to obtain the average performance. Since it is infeasible to perform thousands of trials to get a statistical characterization of the experimental results, we assume the leave-two-individual-out cross-validation can approximate the statistical results. Furthermore, with the leave-two-individual-out cross-validation, the generalization to human activity differences of *ITL* can be demonstrated well.

To evaluate the efficiency of *ITL*, the activity data per person per class in the source domain is divided for training and validation according to the ratio of 8:2. The target dataset is also divided according to the ratio of 2:1 as the same way. Furthermore, we randomly select N instances per person per class from the target training set for fine-tuning, and evaluate the classification performance of *ITL* on the target validation dataset.

4.4.2 Implementation Details

We employ Tensorflow [104], a widely used deep-learning framework developed by Google Brain, to train our model. The proposed *MNet* is pretrained from scratch with D_s. The batch size is set to 32, and the learning rate λ_1 is set to 10^{-3}. The model is pretrained for 400 epochs, and L2 normalization is employed during the training process. For each human individual, 70% of the activity spectrograms in D_t are selected for fine-tuning, and the others for validation. During fine-tuning, the basic learning rate λ_2 is set to 10^{-5}, and an exponentially learning rate decay γ is set to 0.9. The model is fine-tuned for 50 epochs. The batch size is also set to 32. All experiments are performed on a CPU and Ti 1080 GPUs with CUDA for acceleration.

4.4.3 Comparison Methods

To further investigate the performance of *ITL*, we compare the model with several state-of-the-art TL methods, including two radar-based TL approaches and three typical instance-based approaches designed for optical image classification. These comparison approaches are also implemented with the dataset that is described in 4.2.2.

DivNet [6] is specially designed for radar-based HAR with radar MD spectrograms. The network is pretrained with diverse Kinect-based simulated activity data and fine-tuned with a limited number of measured radar data.

DuNet [34] is presented for radar-based HAR. The residual network (ResNet) is adopted as the backbone of the method. And the prior knowledge from simulated MOCAP radar data is transferred by fine-tuning the pretrained backbone with the limited target samples.

NgiamNet [105] is an instance-based TL approach. In this method, the source data are firstly re-weighted based on their similarity to the target data. A DL backbone is firstly pretrained with the re-weighted source data.

GeNet [106] is an instance-based TL approach. The source data are re-weighted based on a similarity metric between the source and target data. Then, a pretrained backbone model is fine-tuned with the re-weighted source data and all target data.

AsgarianNet [107] is another instance-based TL approach that re-weights the source data and uses them for fine-tuning. Distinctively, this approach proposes the hybrid weight for source data, which measures the similarity of a source sample to the target domain and the importance of the sample in the target task.

4.4.4 Experimental Results

• Performance with Limited Numbers of Target Samples

To evaluate the performance of the proposed *ITL* for HAR with limited training data, different amounts of target samples per class per person are provided for training *ITL*. The experimental results are shown in Figure 4.10. Furthermore, we select two baseline methods for comparison to demonstrate the efficiency of *ITL*. In detail, we train the proposed backbone model *MNet* from scratch with the limited target samples, and the test F1 scores are shown with gray marks in Figure 4.10. Then, a *Conventional FT* method that utilizes target samples to fine-tune a pretrained model is adopted for comparison. Specifically, the *MNet* is pretrained on the source training dataset and fine-tuned with the available target samples. The results are shown with orange marks in Figure 4.10.

From Figure 4.10, we can see that the proposed *ITL* yields the best performance among the three methods. Especially when there are 100 target samples per class available, *ITL* outperforms the *Conventional FT* by the largest margin of 4.4% F1 score. In detail, since there are insufficient samples for training *MNet* from scratch, the *Target Model* is susceptible to overfitting, and the performance is the worst. Furthermore, the performance of the *Conventional FT* and *ITL*

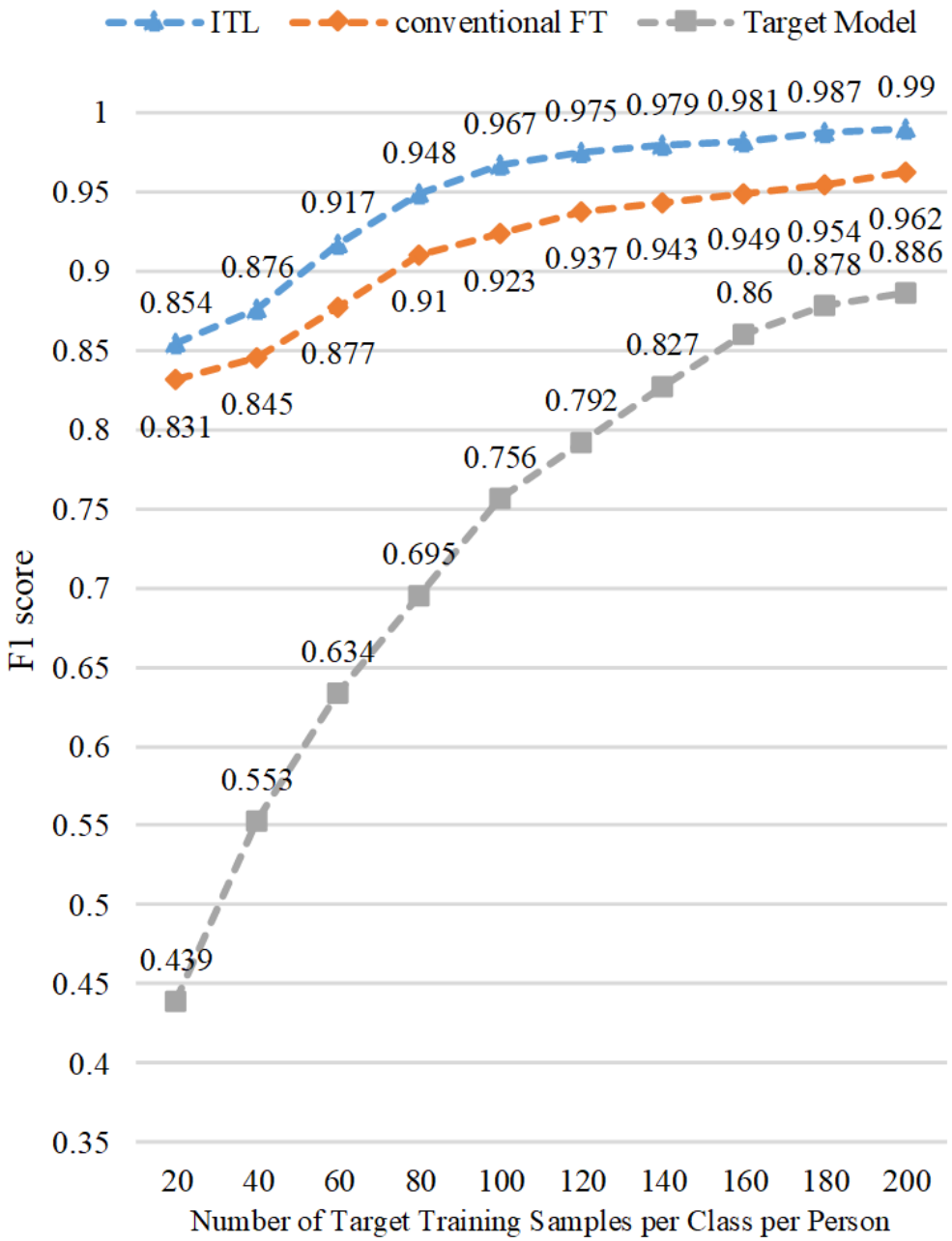


Figure 4.10: The F1 score performance of *ITL*, the *Conventional FT*, and the *Target Model* for classifying the target validation data when diverse amounts of target samples are used for training.

improves as the number of the target samples increases. *ITL* outperforms the *Conventional FT* all the time, demonstrating its better performance for the classification task with limited training data. Since *ITL* has an obvious advantage over the *Conventional FT* in F1 score when there are 100 target samples per person per class, we select 100 target training samples as the typical setting in the following experiments.

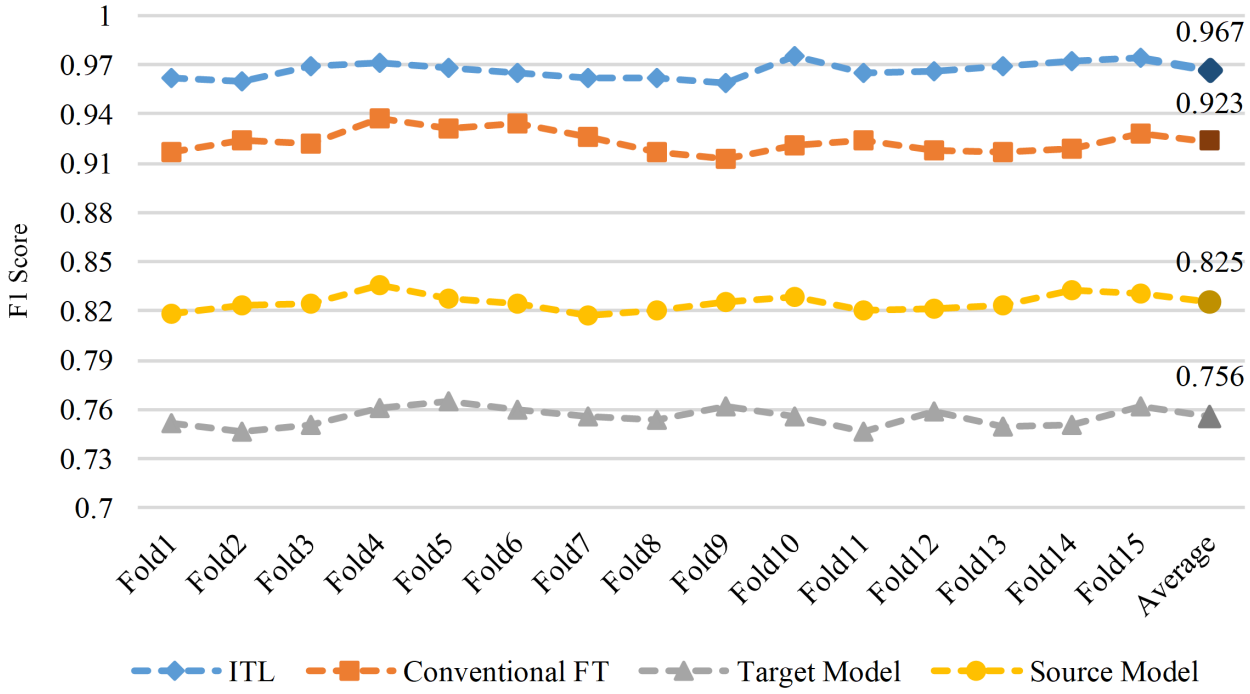


Figure 4.11: The results of the leave two-individual-out cross-validation when there are 100 samples per class per person.

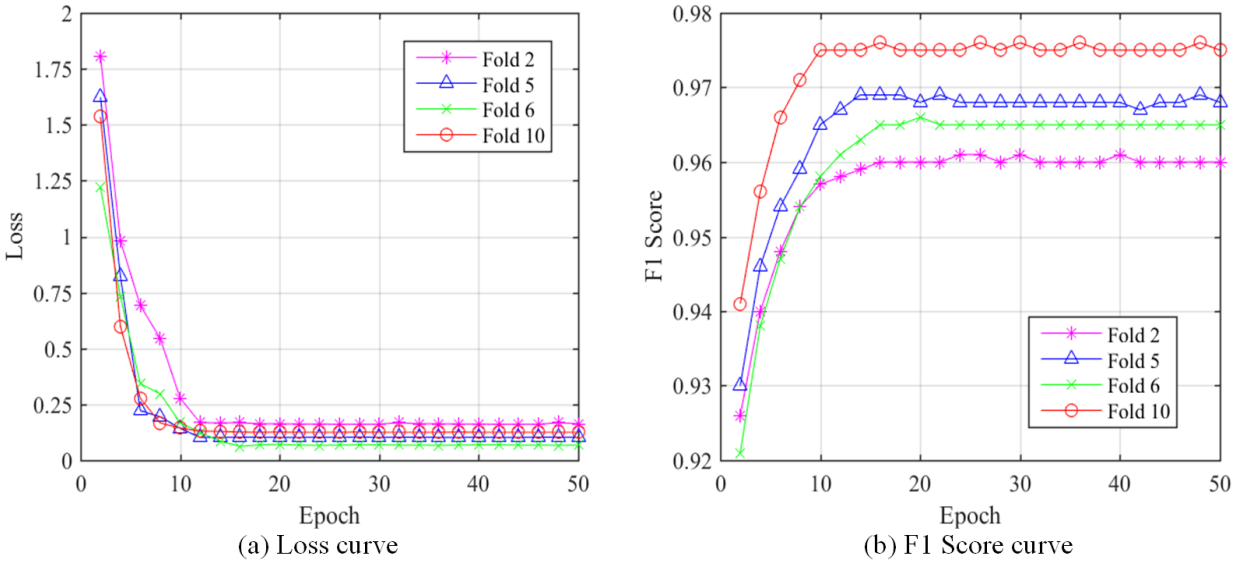
Cross-validation Performance with 100 Target Samples per Person per Class

The results of the leave-two-individual-out cross-validation when there are 100 target samples per class per person are further shown in Figure 4.11.

It can be seen that the performance of *ITL* during the two-individual-out cross-validation is steady. Regardless of the difference between the source and target domain data, the method can achieve an F1 score of about 96.7%. The standard deviation of the 15 folds in the F1 score is merely 4.88e-3. In detail, the average F1 scores of 96.7%, 92.3% and 75.6% are achieved by *ITL*, the *Conventional FT* and the *Target Model*, respectively.

Besides, to demonstrate the infeasibility of directly using the pretrained *MNet* to classify the target samples, we introduce another baseline model *Source Model*. The *Source Model* is obtained by training the backbone *MNet* with the whole source dataset, and no fine-tuning is involved. It can be found that *Source Model* achieves an average F1 score of 82.5% for classifying the target samples, indicating some differences between the source data and the target data.

Furthermore, we select 4 folds (*Fold 2*, *Fold 5*, *Fold 6*, and *Fold 10*) from the 15-fold two-



(a) Loss curve (b) F1 Score curve

Figure 4.12: The loss curves and F1 Score curves of *Fold 2*, *Fold 5*, *Fold 6*, and *Fold 10*.

individual-out validation experiments without any adjective, and their convergence properties are shown in Fig 4.12 in detail. From the loss and F1 score curves, we can see that *ITL* often begins to converge after 10 epochs and yields a stable performance after 30 epochs.

4.4.5 Analysis on Generalization of *ITL*

Generally speaking, the *Conventional FT* method often forgets how to perform the source task as training the new target task progresses. As a result, the DL model fine-tuned with unknown persons' activity data often cannot achieve good performance to recognize the persons' activities in the previous (source) domain. In contrast, the proposed *ITL* is more generalized to the human activity differences, and the fine-tuned model can also scale well to the persons' activities in the source domain.

• Generalization of *ITL* with Diverse Numbers of Target Samples

Firstly, as shown in the leave two-individual-out cross-validation results of Figure 4.11, *ITL* has good performance in recognizing these persons' activities in the target domain.

Specifically, *ITL* is firstly pretrained with the source data and fine-tuned with varying numbers of target samples for the task on the target domain. Then, the source validation samples are employed for classification to test the generalization ability of *ITL*. The experimental results are shown in Figure 4.13. For comparison, the *Conventional FT* model that is fine-tuned with

different amounts of target data is also employed to classify the source samples. The *Source Model* is trained with the source training samples, and no fine-tuning is involved.

As shown in Figure 4.13, since the *Source Model* is trained with the source data, the performance of classifying the source validation samples is the best. In contrast, the performance of the *Conventional FT* is poor. Additionally, the performance of the *Conventional FT* decreases with the increase in the number of target samples. It is because with the amount of the target samples increasing, the distribution of the available target samples tends to be closer and closer to the actual distribution of the target domain data, which is different from that of the source data. In this circumstance, when the fine-tuned model performs well on the target domain, its performance for the source domain usually drops.

In contrast, no matter how many target samples are available for fine-tuning, the performance of *ITL* is better than that of the *Conventional FT*. Especially when there are more than 100 samples per person per class, the performance of classifying the source samples is about 90.0% F1 score, exceeding that of the *Conventional FT* by over 11.0%. Furthermore, when there are more than 160 target samples per class for fine-tuning, the F1 score of *ITL* is only about 7.0% lower than that of the *Source Model*. Good performance is achieved because in *ITL* some source samples highly correlated with the target samples are selected for collaborative FT. As a result, *ITL* can be adapted to the new target task while retaining partial source knowledge. This property makes *ITL* generalized to the activity differences between different domains and scales well to the persons' activities in both the source and the target domains.

• Impact of the Value of K on the Performance of *ITL*

According to Equation 4.6, a certain amount of source samples is selected for every target training sample to perform the collaborative FT. Based on this setting, we change the number of the selected source data by adjusting the value of hyperparameter K, and explore the impact of K on the generalization ability of *ITL*. The experimental results are shown in Figure 4.14. The blue marks represent the results of using *ITL* to classify the target data. The orange marks represent the results of using *ITL* to classify the source data.

From this figure, we can find that with the value of K increasing, the performance of *ITL* for classifying the source validation samples improves. It is because with more source training samples similar to the target data selected and involved in the fine-tuning process, *ITL* can

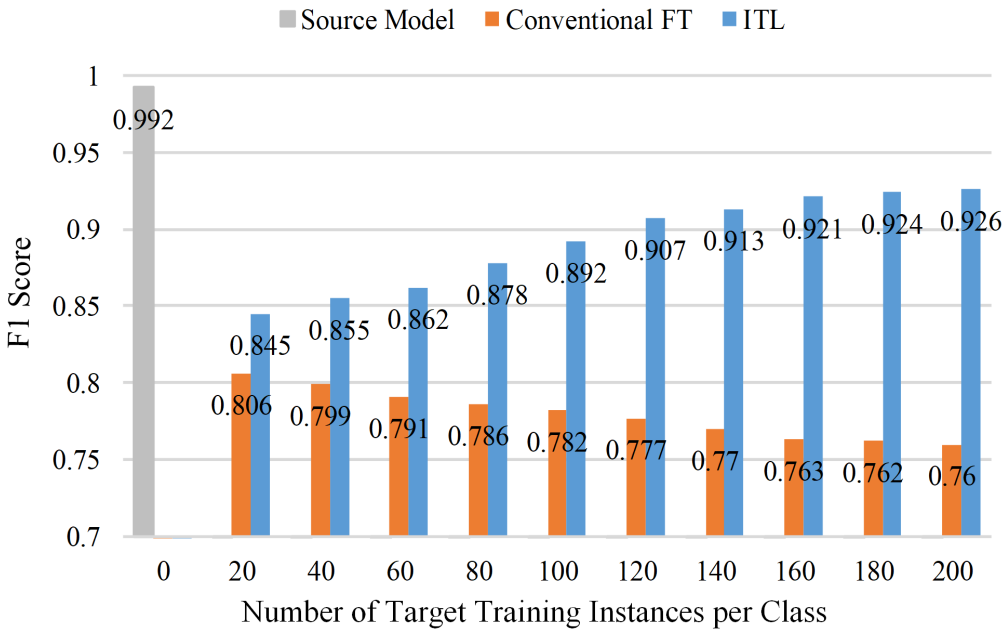


Figure 4.13: The performance of *Source Model*, *Conventional FT* and the proposed *ITL* for classifying the validation source samples.

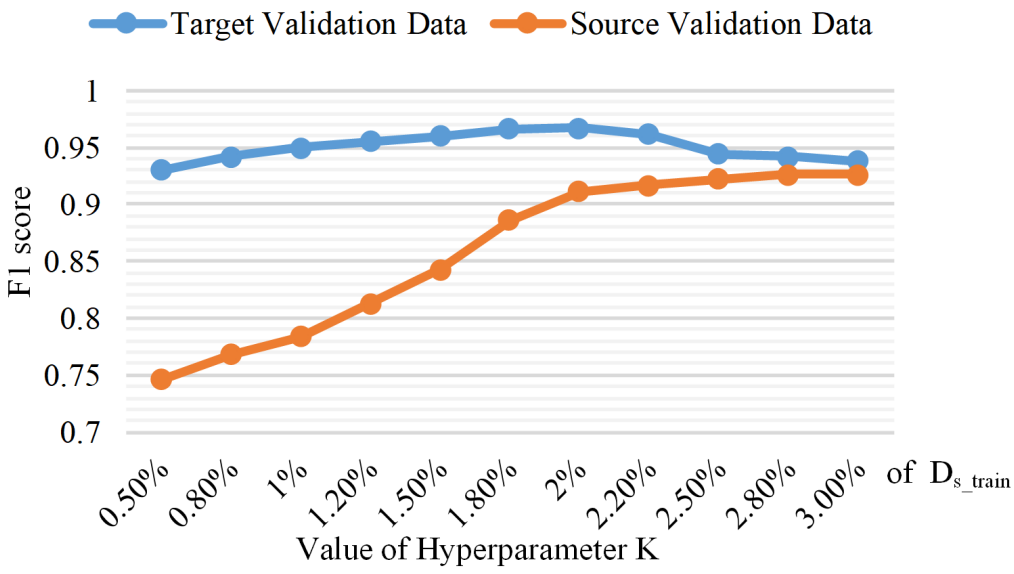


Figure 4.14: The performance variation of *ITL* to diverse values of K when there are 100 target samples available for fine-tuning.

preserve more knowledge of the source domain while performing well on the target domain. At the same time, the performance for classifying the target samples improves with K increasing when K is less than 2.0%. However, a decreasing trend is shown when K is more than 2.0%. It is because with the value of K increasing, more source samples that are not highly correlated

Table 4.2: Comparison with the state-of-the-art methods in F1 Score

	♯Target Training Instances									
	20	**40**	**60**	**80**	**100**	**120**	**140**	**160**	**180**	**200**
[34]	0.866	0.876	0.881	0.894	0.903	0.908	0.914	0.919	0.926	0.931
[6]	0.863	0.867	0.882	0.898	0.909	0.911	0.915	0.924	0.930	0.941
[105]	0.858	0.864	0.883	0.911	0.922	0.934	0.941	0.940	0.946	0.952
[106]	**0.865**	0.881	0.904	0.926	0.937	0.940	0.945	0.952	0.954	0.959
[107]	0.854	0.873	0.905	0.919	0.926	0.938	0.941	0.947	0.950	0.953
Ours	0.861	**0.882**	**0.921**	**0.951**	**0.967**	**0.972**	**0.974**	**0.980**	**0.984**	**0.988**

[1] '♯Target Training Instances' denotes the number of target instances per activity that are used for fine-tuning.

are selected, which is of little help to the task of the target domain.

Furthermore, when K is between 2.0% and 2.5%, the F1 score of *ITL* for classifying the target data is over 94.0%, and the performance of classifying the source samples is over 90.0%. Thus, we can conclude that when K is set between 2.0% and 2.5%, *ITL* is generalized to the differences between the source and the target domains and can scale well to recognize the activities of diverse persons. In further experiments, we set K to 2.0%.

4.4.6 Comparison with the state-of-the-art

To verify the efficacy of *ITL*, we compare it with several state-of-the-art TL approaches. Specifically, we vary the number of the target training samples and perform the leave-two-individual cross-validation on these methods. Then, the average F1 scores are obtained for comparison. The results are depicted in Table 4.2.

Comparison in F1 score: from Table 4.2, we can find that the proposed *ITL* obtains the best performance when there are more than 20 target samples per class for fine-tuning, indicating

Table 4.3: Comparison of the number of model parameters

[34]	[6]	[105]	[106]	[107]	***Ours***
33.16M	54.52M	24.73M	119.69M	22.55M	22.88M

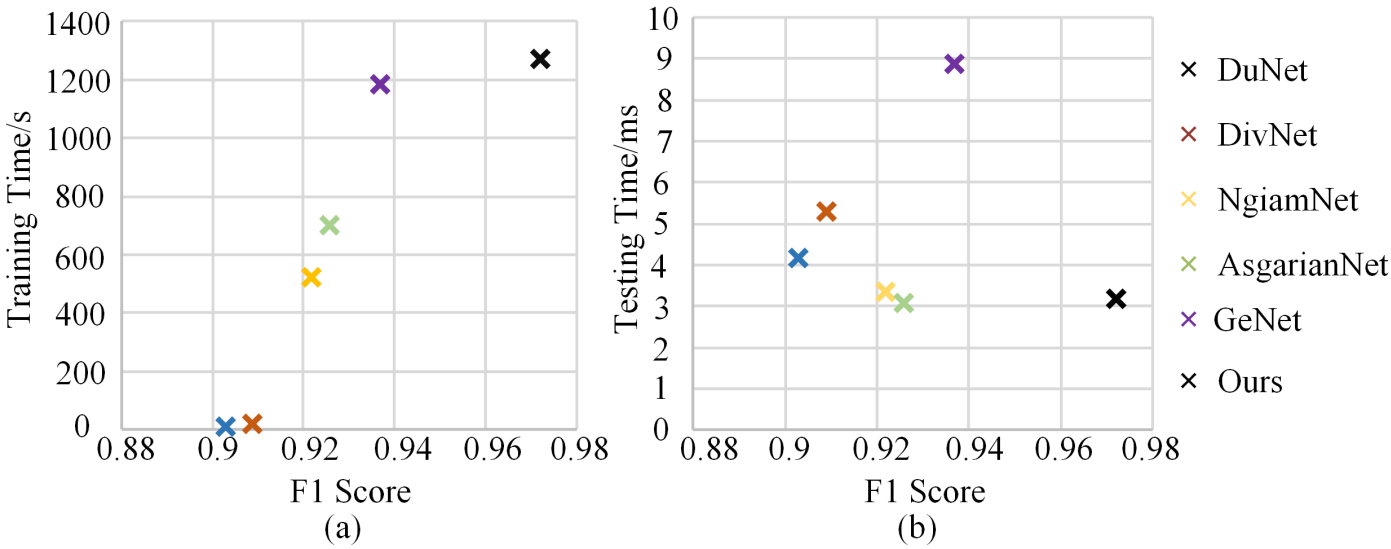


Figure 4.15: Comparison in terms of computational time and F1 score for different methods. In detail, (a) depicts the training time and the F1 scores of the six approaches, and (b) depicts the testing time per sample and the F1 scores of the six approaches.

the feasibility of *ITL* for radar-based HAR with limited data. Though the performance of *ITL* is not the best when there are 20 target samples, the F1 score of *ITL* is merely 0.4% lower than that of *GeNet*, which yields the best performance.

Comparison in the number of parameters: the number of parameters in these DL models is listed in Table 4.3, reflecting the spatial complexity of these methods. It can be seen that *GeNet* has the most parameters due to the complicated backbone. As for our method, with the effective but relatively simple structure of *MNet*, there are only 22.88M parameters in *ITL*.

Comparison in training/testing time: Figure 4.15 illustrates the training time and the testing time of the six methods when there are 100 target training samples per person per class. In detail, in Figure 4.15(a), the model training time and the F1 scores of the six approaches are shown. It can be seen that the training time of *DuNet* and *DivNet* is much shorter than the other methods. The reason is that the two methods only use the target samples to fine-tune their backbone models. As a result, the similarity between the source and target domains is

Table 4.4: Performance comparison with other deep models as backbone

	Conventional FT	***ITL***	**Difference**
MNet-v1	0.914 ±0.03	0.949 ± 0.03	+0.035
MNet-v2	0.913 ± 0.01	0.950 ± 0.01	+0.037
MNet	**0.923** ± 0.01	**0.967 ± 0.02**	+0.038
VGG16	0.918 ± 0.02	0.962 ± 0.01	**+0.044**
ResNet10	0.921 ± 0.03	0.959 ± 0.02	+0.036
Inception-v3	0.905 ± 0.02	0.944 ± 0.01	+0.039

not required to be calculated. In this way, their training time is significantly shortened.

As for the other four instance-based methods *NgiamNet*, *AsgarianNet*, *GeNet* and our approach, the training time includes two parts: the time of calculating the similarity of the source data and the target data and the time of fine-tuning the DL backbone model. Due to the operation of the similarity calculation, the training time of the four methods is much longer than that of *DuNet* and *DivNet*. Among the four methods, the training time of *ITL* is the longest since this method requires selecting the correlated source samples for every target sample. However, though it takes more training time, *ITL* has gained a performance boost and yields the highest F1 score.

Furthermore, Figure 4.15(b) shows the comparison results regarding testing time per sample and F1 score for the six methods. In general, the training process of a model is often performed offline. Compared with the training time, the testing time per sample significantly impacts whether the model can be applied in practice. The similarity calculation operation is not required during testing, and the running time is considerably shortened. As shown in this subfigure, though the training time is long, the proposed *ITL* takes a short time to classify a sample.

4.5 Ablation Study on *ITL*

To verify the effectiveness of different components in *ITL*, some ablation studies on *ITL* are performed. During the ablation study, we performed the experiments in a typical setting where

100 target samples per person per class are available. The 15-fold cross-validation is employed to obtain the average F1 score.

4.5.1 Ablation Study on MNet

To demonstrate the good performance of *MNet* for recognizing human activities with radar MD spectrograms, we slightly change the structure of *MNet*. Three variants of *MNet* are designed, which are referred to as *MNet-v1* and *MNet-v2*, respectively. Specifically, to obtain *MNet-v1*, the two channel-wise attention modules in *MNet* are removed. In *MNet-v2*, the dilation rate in the two dilated convolutional layers is set to 1, and the dilated convolution operations are converted into the general convolutions.

Then, we compare the performance of the two TL methods(the *Conventional FT* and *ITL*) when using *MNet*, *MNet-v1* and *MNet-v2* as the backbone, respectively. The comparison results in F1 score are listed in Table 4.4. As shown in this table, when the channel-wise attention modules are removed, the performance of both the *Conventional FT* and *ITL* decreases, indicating that the two channel-wise attention modules are vital to the performance of *MNet*. At the same time, the performance of *MNet-v2* is not as good as that of *MNet* regardless of whether the source data is fine-tuned or not, demonstrating the efficiency of dilated convolution operations.

Furthermore, to demonstrate the superiority of *MNet* for radar-based HAR, we replace *MNet* with several typical CNN models, including *VGG16*, *ResNet10* and *Inception-v3*. We compare these three DL models among the existing state-of-the-art models because they have a similar number of convolution layers to *MNet*. Then, their performance as the backbone of *Conventional FT* method and *ITL* is compared. The comparison results are listed in Table 4.4. We can see that regardless of whether the source data is used for fine-tuning, our model achieves the best results and is more suitable for the HAR tasks with radar MD spectrograms than the other DL models. Furthermore, when using *VGG16* as the backbone, *ITL* outperforms the *Conventional FT* the most, with a difference of 0.44% F1 score. Additionally, *VGG16* yields similar performance to *MNet* when used as the backbone of *ITL*, with an F1 score of 96.2%. It is indicated that compared with *ResNet10* and *Inception-v3*, *VGG16* is more suitable to transfer the activity characteristic in radar spectrograms.

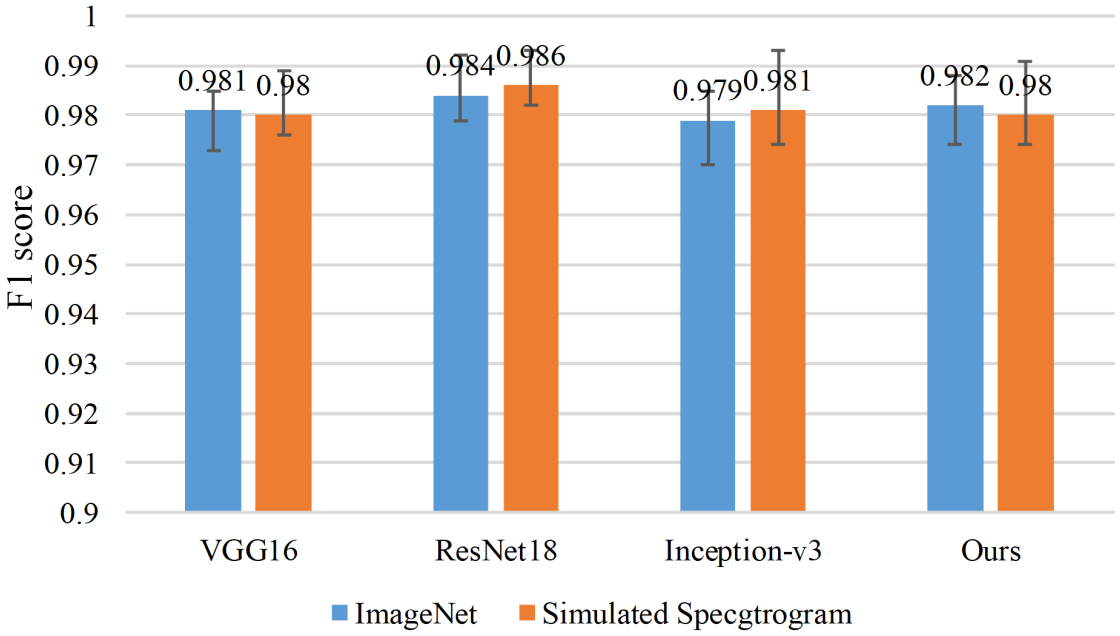


Figure 4.16: The performance of *ITL* in average F1 score when using different deep models that are pretrained on *ImageNet* and a simulated radar dataset, respectively.

4.5.2 Ablation Study on Correlated Source Data Selection

• Analysis on the MD Signature Descriptor

Firstly, we replace *AlexNet* with three typical CNNs *VGG16*, *ResNet18* and *Inception-v3* as the feature extractor and utilize the last convolutional layers of these models as filters to obtain the MD signature descriptors. Furthermore, the optical image dataset *ImageNet*, instead of a radar image dataset, is utilized for training the feature extractor. Though radar spectrograms have different characteristics from optical images, several radar-based HAR literature demonstrated the feasibility of extracting features from radar spectrograms with a model trained on a large-scale optical image dataset [34], [78], [108], [109]. Furthermore, to make our work more complete and comprehensive, we utilize a simulated MOCAP radar dataset [110] instead of *ImageNet* to train the feature extractor. The results are illustrated in Figure 4.16.

As shown in Figure 4.16, the performance of using the four deep models to obtain the MD signature descriptors is similar, demonstrating the feasibility of applying these typical deep models as the feature extractor of *ITL*. However, despite the excellent performance, the complexity of these models in obtaining MD signature descriptors is diverse due to different numbers of kernels in the last convolutional layers. In this circumstance, by balancing complexity and

Table 4.5: The classification results on the target validation dataset with different fine-tuning algorithms.

Method	**Transfer Performance**
S1: Fine-tuning with D_t only	92.3%
S2: Fine-tuning with D_t and the whole D_s	91.8%
S3: Fine-tuning with D_t and the randomly selected source data	91.5%
S4: Fine-tuning with D_t and the source data selected with EMD-based algorithm *(Ours)*	**96.7%**

F1 score, we can conclude that *AlexNet* is a better choice. Additionally, the performance of using simulated radar spectrograms and *ImageNet* is broadly similar. Though the simulated radar data is more similar to our measured data, there is no noticeable performance advantage. However, as the error lines in the figure show, the maximum F1 score when using simulated data to train the feature extractor is often more significant than the top F1 score when using *ImageNet*. Furthermore, the performance of *ResNet18* trained with simulated radar data is the best, with an average F1 score of 97.1%. Based on these results, we have reasons to believe that using simulated radar data to train the feature extractor has more potential to achieve good performance for HAR [6], [111].

• **Analysis on Source Instance Selection**

To demonstrate the efficiency of the EMD-based source instance selection algorithm, we compare this solution with the other three source instance selection solutions:

S1: Fine-tuning the pretrained *MNet* with only target dataset D_t (*Conventional FT*).

S2: Fine-tuning the pretrained *MNet* with the whole source dataset D_s and the target dataset D_t.

S3: Fine-tuning the pretrained *MNet* with randomly selected source data and the whole D_t.

The comparison results are listed in Table 4.5. The table shows that the proposed EMD-based source instance selection algorithm yields the best performance. Specifically, the F1 score of the

Conventional FT (**S**1) is 92.3%. The F1 score of fine-tuning with both D_s and D_t (**S**2) is 91.8%. The F1 score of fine-tuning with the randomly selected source data and D_t (**S**3) is 91.5%. It can be seen that the performance of **S**2 and **S**3 is not improved and is even worse than **S**1. It is because using the whole source dataset or randomly selected source data for fine-tuning can bring some negative knowledge transfer to the network. In contrast, the EMD-based source instance selection algorithm achieves the best performance, outperforming the solution **S**1 by 4.4% F1 score.

4.5.3 Ablation Study on Adaptive Collaborative Fine-tuning

Analysis on the comparison experiments

To investigate the effect of the two elements (adaptive source data search and source instance re-weighting) in ACFT on the performance of *ITL*, we perform the three following comparison experiments:

C1: Assigning equal importance to the selected source instances, and setting the same loss weight to all of the instances in Equation 4.3.

C2: In each of the first five epochs, if $\hat{y}_i^{(t)} = y_i^{(t)}$, β correlated source instances are selected for each target instance, without the limitation of information entropy E.

C3: Replace the loss weights $sin(\frac{\pi}{2} * \frac{w_i}{w_{max}})$ of source samples in Equation 4.3 with $\frac{\pi}{2} * \frac{w_i}{w_{max}}$.

Table 4.6: Comparison study for the proposed ACFT algorithm.

Method	**Transfer Performance**
C1: Assigning equal instance importance	93.8%
C2: Neglecting the limitation of E	95.7%
C3: Reconstruction loss function	96.0%
C4: The proposed ACFT algorithm (*Ours*)	**96.7%**

Table 4.6 shows the comparison results. It can be seen that our method yields the best performance. When assigning equal importance to the selected source instances (**C**1), an F1 score of 93.8% is yielded, which is 2.9% lower than that of the proposed ACFT algorithm. When selecting the same number β of correlated source instances for each target instance if $\hat{y}_i^{(t)} = y_i^{(t)}$

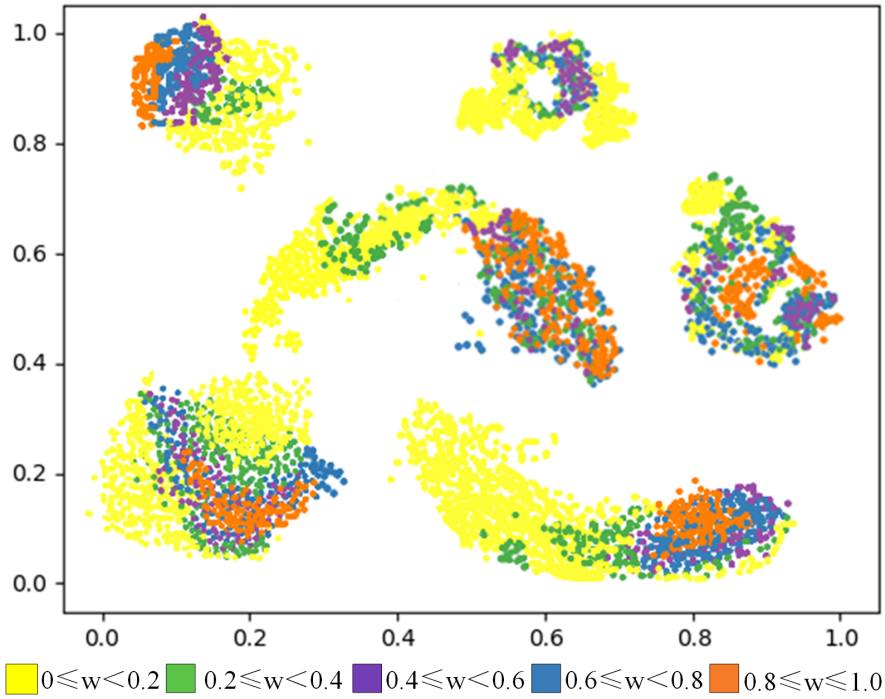


Figure 4.17: Visualization of the loss weights w assigned to the source instances.

(**C**2), the performance drops to 95.7%. Furthermore, when replacing $sin(\frac{\pi}{2} * \frac{w_i}{w_{max}})$ with $\frac{\pi}{2} * \frac{w_i}{w_{max}}$ in the loss function L (**C**3), the performance of *ITL* decreases to 96.0%. Though using *sine* function is the result of heuristic attempts, the comparison results demonstrate the efficiency of using *sine* instead of linear loss weights.

- **Visualization of Diverse Importance of the Source Samples**

To reveal the diverse importance of the selected source instances, we visualize the loss weights assigned to the source data with t-SNE. In particular, the loss weights of the selected source samples in the fifth fine-tuning epoch are recorded when there are 100 target samples per person per class available for training. For those source training samples that are not used for fine-tuning, the loss weights are 0. Then, all source training samples are input to the *Source Model*, and the feature vectors output by the last convolutional layer are visualized with t-SNE. The visualization results are shown in Figure 4.17. A more significant loss weight means that the source instance is attached more importance to the HAR task. In contrast, a smaller weight implies the sample is less correlated to the target data and is less critical to the collaborative fine-tuning process. From Fig, 4.17, we can see that *ITL*, only partial source domain data is helpful for the classification task of the target domain and selected for the ACFT process.

4.6 Summary

This chapter proposed an instance-based TL approach *ITL* for radar-based cross-target activity recognition. The approach comprises three interconnected and necessary parts (*MNet* pretraining, CSDS and ACFT) rather than a collection of three distinct pieces. This study collected six types of measured activity data from six human subjects using a pulsed UWB radar for experiments.

Experimental results showed that the proposed *ITL* was able to accurately recognize the activities of six different persons with limited radar data, with an F1 score of 96.7% when there were only 100 samples per person per class. Furthermore, when the model was trained to recognize a new person's activities, it could still perform well on the previous HAR task, effectively alleviating the catastrophic forgetting problem. Additionally, some ablation studies were conducted to demonstrate the uniqueness of the components in *ITL*. Any exclusion of these components resulted in performance degradation. Finally, despite the effectiveness of *ITL*, how to reduce the computational cost of the model needs to be further researched.

Chapter 5

Supervised Domain Adaptation for Few-shot Radar-based HAR

5.1 Introduction

Most of the existing DL solutions for radar-based HAR are trained with a large volume of labeled data in a supervised manner. However, obtaining a large-scale radar dataset is often difficult because annotating radar signals is complex and time-consuming. To this extent, data scarcity becomes a bottleneck for the emerging radar-based HAR application. Prompt solutions need to apply DL techniques to insufficient annotated radar data.

Few-shot learning (FSL) utilizes prior knowledge to make a trained model generalize on new tasks of limited supervised experience [112]. This method can relieve the workload of collecting a large number of supervised samples and save great manpower and time. It is also suitable for the applications where supervised information is hard or impossible to acquire. In radar-based HAR applications, since annotating radar signals is difficult and time-consuming, most labelled radar datasets are too small-scale to train a DL model from scratch. Under this circumstance, FSL is an important mechanism to enable a trained model applicable to an new task with limited labelled samples.

There are several FSL methods that can potentially be used for radar-based HAR. Motiian *et al.* [113] adopted adversarial domain adaptation mechanism to FSL. To extract semantic-alignment features with this mechanism, complicated sampling and preprocessing on the training data are

required. Furthermore, Jones *et al.* [114] proposed a siamese network and employed different distance metrics to minimize inter-class differences and maximize intra-class differences. Nevertheless, the performance of this method is dependent on the distance metric, which difficult to determine. Feng *et al.* [115] proposed a model parameter transfer method for few-shot HAR with wearable sensor data. Rostami *et al.* [116] proposed a cross-domain few-shot learning method by transferring the knowledge from a known domain to the target domain, which alleviated the need for large-scale labeled data.

In this chapter, we propose a supervised few-shot adversarial domain adaptation (*FS-ADA*) method for radar-based HAR. This method does not require much radar data for training when applied to a new environment. Instead, HAR can be accomplished using the proposed method when only a few samples per class are used. Our main contributions are summarized as follows.

- We propose a multi-class discriminator network, which integrates the category classifier and the domain discriminator. The network can be used for both the domain-discrimination and activity-classification tasks. By sharing part of the discriminator network between the two tasks, the model complexity of *FS-ADA* is reduced.

- We propose a multitask generative adversarial training scheme. The proposed method optimizes the feature extractor and the discriminator in *FS-ADA* alternatively. The extracted features can be domain-invariant and category-discriminative by optimizing the discriminator with the combination of the domain discrimination loss and the activity classification loss.

- We provide extensive experimental results to verify the performance of our proposed method. The results show that the proposed *FS-ADA* method outperforms the state-of-the-art benchmarks on two few-shot learning tasks. We also demonstrate the effectiveness of *FS-ADA* with only limited training data.

The rest of this chapter is organized as follows. Section II introduces the related work about radar-based HAR and few-shot learning. The pipeline of the proposed *FS-ADA* method is briefly described in Section 5.2. Section 5.3 presents the key techniques of the *FS-ADA* method. Section 5.4 presents the experimental details and the experimental results. Finally, we conclude the paper in Section 5.5 and point out the direction for further work.

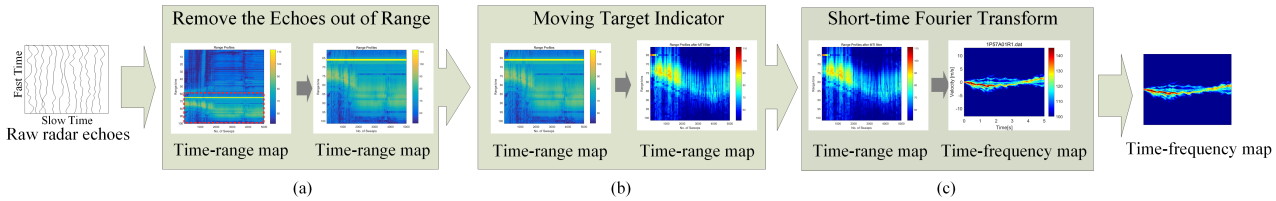


Figure 5.1: The preprocessing pipeline of radar echoes. (a) Filter out the echoes out of range. (b) Remove the static background clutter. (c) Short-time Fourier transform and normalization.

To extract features and classify the activities automatically, DL has been increasingly adopted for radar-based HAR [60], [75], [117], [118]. However, most of the existing DL-based methods, require a large volume of labelled radar data for training. When the training data is limited, the trained model tends to be overfitting, and cannot be adapted to a new environment.

5.2 Few-shot Adversarial Domain Adaptation

In this section, we introduce the problem setup of few-shot HAR, and describe the pipeline of the proposed *FS-ADA* briefly.

5.2.1 Problem Setup

We assume that there is a small-scale radar spectrogram dataset $\mathcal{D}_t = \{(x_t^i, y_t^i)\}_{i=1}^M$, where only a few labeled samples per class are available. Here, $x_t^i \in \mathbf{X}_t$ denotes the ith spectrogram in $\mathcal{D}_t$ with a label $y_t^i \in \mathbf{Y}_t$. There is another labeled radar spectrogram dataset $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^N$, where $x_s^i \in \mathbf{X}_s$ denotes the ith spectrogram in $\mathcal{D}_s$ with a label $y_s^i \in \mathbf{Y}_s$. The labeled samples in D_s are sufficient to learn a prediction function f_s: $\mathbf{X}_s \rightarrow \mathbf{Y}_s$. $\mathcal{D}_t$ and $\mathcal{D}_s$ share the same categories. Let $\mathcal{D}_s$ denote the source domain and $\mathcal{D}_t$ the target domain. We assume that there is a "domain shift" between the two domains, i.e., the different distributions between the source data $\mathcal{D}_s$ and the target data $\mathcal{D}_t$. Our goal under is to learn a target prediction function f^t: $\mathbf{X}_t \rightarrow \mathbf{Y}_t$ to classify data in $\mathbf{X}_t$. To achieve this goal, a supervised few-shot adversarial domain adaptation method is proposed.

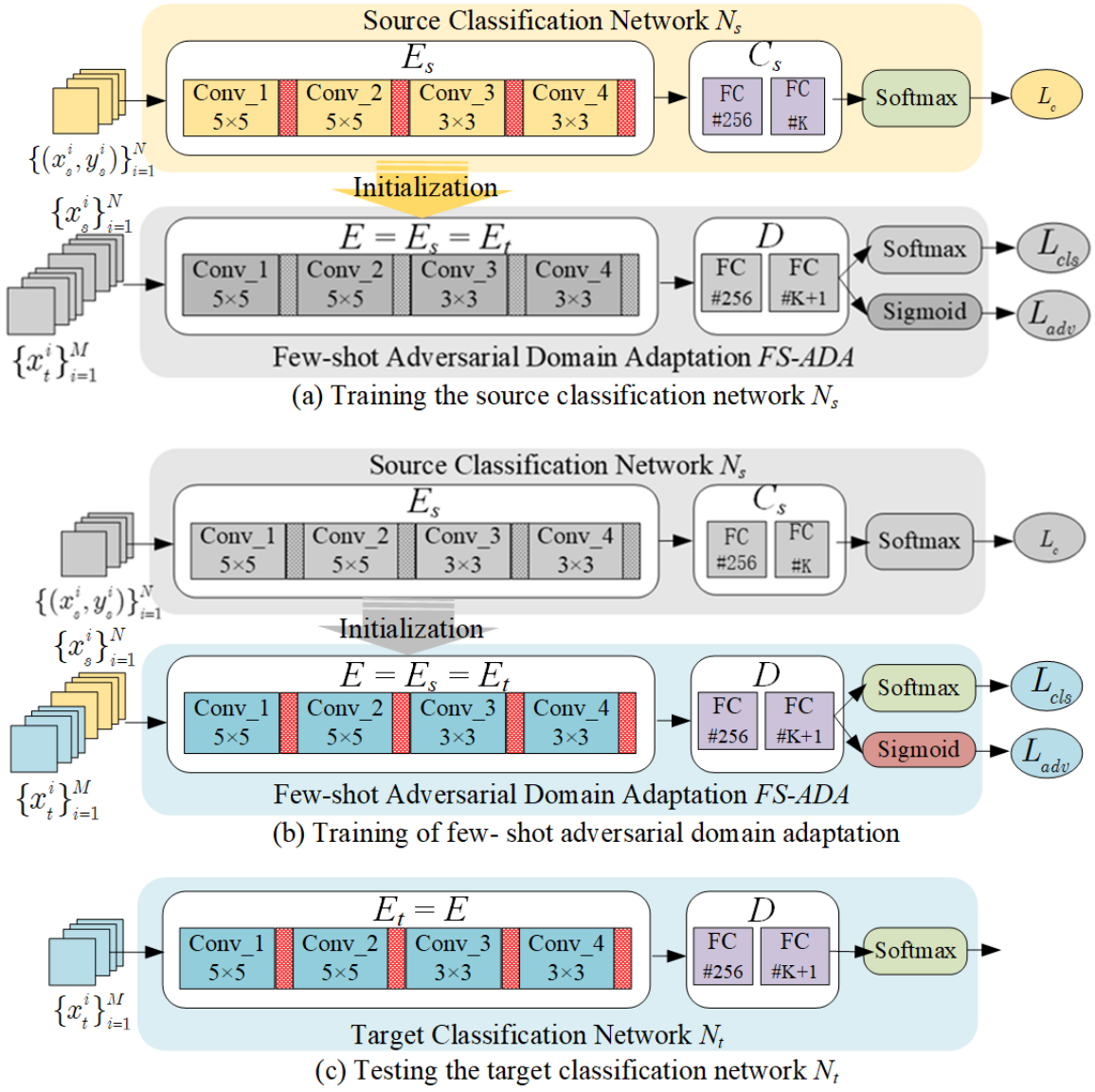


Figure 5.2: Main steps of the proposed *FS-ADA*. Note that the red rectangles denote "Max-pooling" operation, and the parameters of the gray parts in every substeps are fixed during training.

5.2.2 HAR Pipeline with *FS-ADA*

Here, we briefly describe the main steps in the HAR pipeline of *FS-ADA*. In the next section, we will elaborate on key techniques used in this method.

• Radar Data Preprocessing

The preprocessing pipeline of radar echoes is illustrated in Figure 5.1. Firstly, since persons are moving within the specified range ahead of the radar, we filter out the echoes reflected from objects outside the target range based on the time-range maps. Then, an MTI is adopted on the filtered echoes to remove the static background clutter. Next, the STFT is employed to transform the denoised radar signals into 2-D time-Doppler spectrograms. Finally, these spectrograms are normalized and input to the proposed classification model.

• Activity Recognition with *FS-ADA*

The main steps of the proposed *FS-ADA* are shown in Figure 5.2. As shown in Figure 5.2(a), in the first step, a source classification network N_s, which is composed of a source feature extractor E_s and a source classifier C_s, is constructed. Then, a pre-existing source dataset is employed to train N_s with the supervised cross-entropy loss function.

Next step is to train the target classification network N_t, as shown in Figure 5.2(b).Note that N_t shares the same architecture of N_s, and its parameters are initialized with the parameters of the trained N_s. Then, the proposed *FS-ADA* method is used to train the target classification model. Specifically, the adversarial domain adaptation (ADA) scheme is employed to map the input data into a common feature space. In this ADA scheme, symmetric transformation [119] is applied to make E_t fit into a similar output distribution to that of E_s with limited labeled samples, i.e., E_s=E_t=E. Different from unsupervised ADA utilizing a binary domain discriminator [119], we propose a multi-class discriminator D to take advantage of the limited supervised label information. The discriminator D integrates the category classifier and the domain discriminator into a network, and performs the HAR task and the domain discrimination task. As a result, the proposed *FS-ADA* can extract both domain-invariant and category-discriminative features from the input source/target data. Furthermore, a multitask generative adversarial loss, which is a combination of the activity classification loss and the domain discrimination loss, is presented to train E_t and D alternatively. In contrast with the

FSL method [120], which employed two independent networks for HAR and domain discrimination, we make the two tasks share part of the network D. Thus, the model complexity can be reduced.

The third step is to test the trained target classification network N_t, as shown in Figure 5.2(c). N_t is composed of the feature extractor E_t and the discriminator D. The function of D is classifying the input data and recognizing the corresponding activity. Therefore, when a piece of radar spectrogram is input into N_t, the domain discrimination result can be ignored.

5.3 Key Techniques of *FS-ADA*

As shown in Figure 5.2, the main steps of the proposed *FS-ADA* are composed of three modules: the source feature extractor (E_s), the target feature extractor (E_t), and the multi-class discriminator D. Details of the network architecture and training strategies are described as follows.

5.3.1 Feature Extractors with Shared Weights

Since there are sufficient source samples in $\mathcal{D}_s$, the classification network N_s can be trained from scratch for the source classification task. The supervised loss function L_c is formulated as

$$L_c = \mathbb{E}(\sum_{i=1}^{N} \ell(C_s(E_s(x_s^i)), y_s^i)), \tag{5.1}$$

where $\mathbb{E}[\cdot]$ denotes statistical expectation; $\ell(\cdot)$ is the cross-entropy loss function; E_s denotes the source feature extractor; and C_s denotes the source classifier. E_s and C_s make up the source classification network, i.e., $N_s = E_s \circ C_s$, where $\circ$ denotes the model concatenation.

After obtaining the trained E_s, we proceed to find a suitable E_t that can embed the target data to the same feature space as the source data. Typical DA methods fix the parameters of E_s and make E_t mimic the invariant distribution output by E_s. However, since there are few samples available in $\mathcal{D}_t$, E_t cannot fit into a similar output distribution to that for E_s. Under this circumstance, we adopt the symmetric transformation [119], which enables E_t to share the same structure and parameters as E_s, i.e., $E_t = E_s = E$.

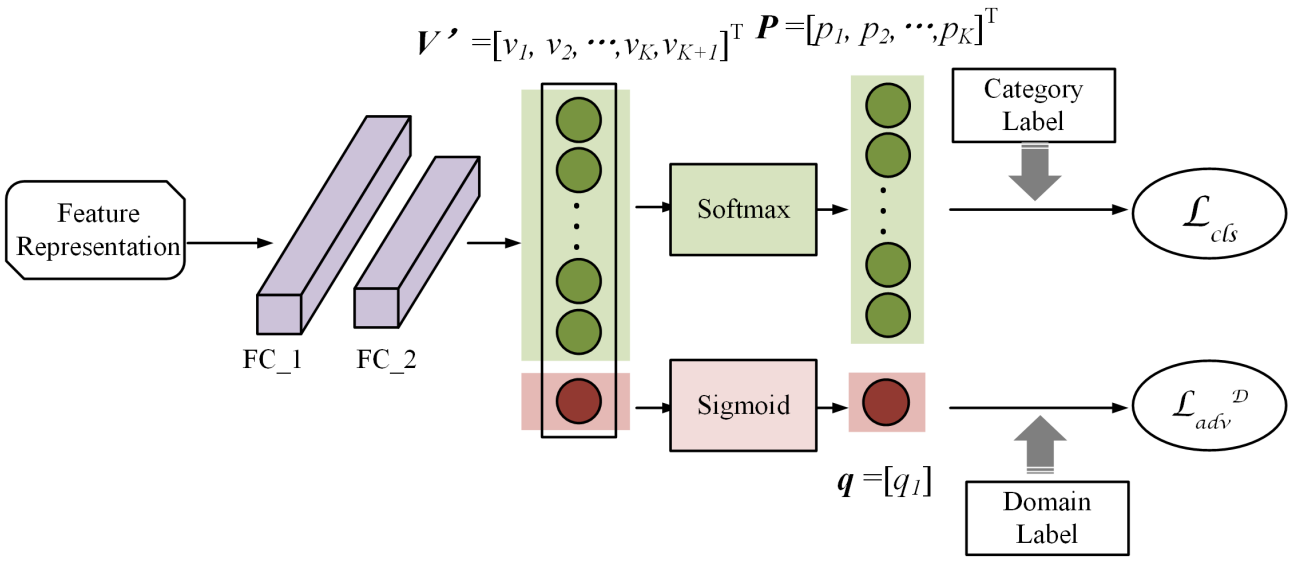


Figure 5.3: The structure of the proposed multi-class discriminator D, which is composed of two FC layers followed by a softmax layer and a sigmoid layer.

5.3.2 Multi-class Discriminator

To exploit the label information, we propose a multi-class discriminator D that combines the classifier and the discriminator to perform the supervised ADA. The structure of the proposed multi-class discriminator D is illustrated in Figure 5.3. The logit $V^{'}$ output by the second FC layer is first divided into two parts: classification nodes and domain node. Then, the two parts are fed into the softmax layer and the sigmoid layer for category classification and domain discrimination, respectively.

Generally, a standard classifier outputs a K-dimensional vector $\mathbf{v}$ to classify an input x into one of K possible classes, where $\mathbf{v} = [v_1, v_2, ..., v_K]^T$. Then, a softmax is applied to the output $\mathbf{v}$, and transforms $\mathbf{v}$ into the class probabilities $\mathbf{p} = [p_1, p_2, ..., p_K]^T$. The probability p_j of predicting x as being in the jth class is written as

$$p_j(y = j) = \frac{\exp(v_j)}{\sum_{k=1}^{K}(\exp(v_k))}. \tag{5.2}$$

In particular, the domain discriminator of the general ADA is a binary classifier with only one node in the last layer.

In this paper, we combine the activity classifier and the discriminator by designing a novel discriminator where there are $K + 1$ nodes in the last fully connected (FC) layer. The former K nodes referred to as "classification nodes", are used to classify the input into one of the K activity classes. The last node, called "domain node", is used for domain discrimination. When

a feature representation $\mathbf{r}$ is input to the multi-class discriminator, the output logit is changed from $\mathbf{v}$ into $\mathbf{v}^{'}$, where $\mathbf{v}^{'} = [v_1, v_2, ..., v_K, v_{K+1}]^T$.

Then, a softmax function is employed on the classification nodes. The probability that x belongs to the jth class is calculated as

$$p_j(y = j) = \frac{\exp(v_j^{'})}{\sum_{k=1}^{K}(\exp(v_k^{'}))}. \tag{5.3}$$

The cross-entropy function is applied to the optimization process of the HAR task. The loss of activity classification L_{cls} is formulated as

$$\mathcal{L}_{cls} = -\sum_{d\in\{s,t\}} \mathbb{E}_{x_d\sim X_d}[\mathbf{p}^{'}\log(\mathbf{p})], \tag{5.4}$$

where $\mathbf{p}^{'}$ is the one-hot label of x.

Meanwhile, to perform the domain discrimination task, a sigmoid function is applied to the domain node to predict the domain of x. The domain probability is calculated as

$$q = \frac{1}{1 + \exp(-v_{K+1}^{'})}. \tag{5.5}$$

A binary cross-entropy loss $\mathcal{L}_{dis}$ is employed for the domain discrimination loss. $\mathcal{L}_{dis}$ is formulated as

$$\mathcal{L}_{dis} = -\mathbb{E}_{x_s\sim X_s}[\log(q_s)] - \mathbb{E}_{x_t\sim X_t}[\log(1 - q_t)], \tag{5.6}$$

where q_s and q_t are the output domain probabilities of the source sample x_s and the target sample x_t, respectively.

5.3.3 Training Process of FS-ADA

Generative adversarial learning [119] is adopted to train the proposed FS-ADA method. Two training losses are also designed specifically to train E and D.

- Loss function of E: Generally, ADA methods utilize

the generative adversarial network (GAN) loss $\mathcal{L}_{adv}^{E}$ to train E, which is depicted as

$$\begin{aligned}&\min_{E} \mathcal{L}_{adv}^{E}(\mathbf{X}_s, \mathbf{X}_t, D)\\ =&-\mathbb{E}_{x_s \sim X_s}[log(1-q_s)] - \mathbb{E}_{x_t \sim X_t}[log(q_t)].\end{aligned} \tag{5.7}$$

This loss is typically used in the setting where the generator attempts to mimic another unchanging distribution. However, we note that with the weight sharing mechanism between E_s and E_t, applying this GAN loss will result in unstable training process, which makes the model hard to converge. This is because under this setting, the extracted $\mathbf{r}_s$ and $\mathbf{r}_t$ change simultaneously when the parameters of E are updated. To resolve this problem and improve stability, the domain confusion loss [119] is adopted. We extend it to our scenario, and the domain confusion loss $\mathcal{L}_{adv}^{E}$ for optimizing E can be formulated as

$$\begin{aligned}&\min_{E} \mathcal{L}_{adv}^{E}(\mathbf{X}_s, \mathbf{X}_t, D)\\ =&-\sum_{d\in\{s,t\}} \mathbb{E}_{x_d \sim X_d}[\frac{1}{2} log D^{'}(E(x_d)) + \frac{1}{2} log(1 - D^{'}(E(x_d)))]\\ =&-\sum_{d\in\{s,t\}} \mathbb{E}_{x_d \sim X_d}[\frac{1}{2} log(q_d) + \frac{1}{2} log(1 - q_d)],\\ &s.t. \quad \{E_s^l = E_t^l = E^l\}_{l\in\{1,2,\ldots,n\}}\end{aligned} \tag{5.8}$$

where $\mathbb{E}[\cdot]$ denotes the statistical expectation; $D^{'}$ represents the output by the domain node of D; and E^l represents the lth layer of E.

- Loss function of D: $\mathcal{L}_{adv}^{D}$ is the sum of $\mathcal{L}_{cls}$ in Equation (5.4) and $\mathcal{L}_{dis}$ in Equation (5.6),

which can be formulated as

$$\begin{aligned}&\min_{E} \mathcal{L}_{adv}^{D}(\mathbf{X}_s, \mathbf{X}_t, \mathbf{Y}_s, \mathbf{Y}_t, E)\\ =&\mathcal{L}_{dis}(\mathbf{X}_s, \mathbf{X}_t, E) + \lambda \mathcal{L}_{cls}(\mathbf{X}_s, \mathbf{X}_t, \mathbf{Y}_s, \mathbf{Y}_t, E).\end{aligned} \tag{5.9}$$

where λ is a hyper-parameter that adjusts the weights of the two loss functions $\mathcal{L}_{dis}$ and $\mathcal{L}_{cls}$, and is set to 1.0.

It is noted that when the trained model is used for testing, E is employed to extract feature representation from the input data. The classification nodes of D can be used as a classifier, while the domain node of D is abandoned.

5.3.4 Complexity Analysis of *FS-ADA*

The complexity of the proposed *FS-ADA* method is analyzed as follows. The time complexity of DL models can be measured by floating point of operations (FLOPs). The space complexity can be measured by the volume of model. Since the activity classification network is composed of the CNN-based feature extractor and the fully-connected classifier, its complexity is the sum of the complexities of the feature extractor and the classifier.

In detail, the time complexity Com_t^{CNN} and space complexity Com_s^{CNN} of the CNNs are derived as

$$Com_t^{CNN} = O(\sum_{l=1}^{n} N_{l-1}^{CNN} \cdot S_l^2 \cdot N_l^{CNN} \cdot M_l^2), \tag{5.10}$$

$$Com_s^{CNN} = O(\sum_{l=1}^{n} N_{l-1}^{CNN} \cdot S_l^2 \cdot N_l^{CNN} + \sum_{l=1}^{n} N_l^{CNN} \cdot M_l^2), \tag{5.11}$$

where l is the index of a convolutional layer; n is the number of convolutional layers; N_{l-1}^{CNN} is the number of output channels of the (l-1)-*th* layer, and is also the number of input channels of the l-*th* layer; N_l^{CNN} is the number of filters in the l-*th* layer; s_l is the spatial size of the filter in the l-*th* layer; and M_l is the spatial size of the feature map output from the l-*th* layer, which is calculated as

$$M_l = (X - S_l + 2 \cdot Padding)/Stride + 1, \tag{5.12}$$

where X is the input data; and *Padding* and *Stride* are the parameters set in the covolution operation.

The time complexity Com_t^{FC} and space complexity Com_s^{FC} of the fully-connected classifiers are calculated as

$$Com_t^{FC} = O(\sum_{l=1}^{m} N_{l-1}^{FC} \cdot N_l^{FC}), \tag{5.13}$$

$$Com_s^{FC} = O(\sum_{l=1}^{m} (N_{l-1}^{FC} \cdot N_l^{FC} + N_l^{FC})), \tag{5.14}$$

where m is the index of a FC layer; N_{l-1}^{FC} is the number of input nodes of the l-*th* layer; and N_l^{FC} is the number of output nodes in the l-*th* layer.

5.4 Experimental Results

In this section, we perform experiments on the two HAR tasks to validate the performance of the proposed *FS-ADA* method.

5.4.1 Dataset Description

We evaluate the proposed *FS-ADA* method on two few-shot HAR tasks. The utilized datasets are described briefly as follow. Several typical spectrograms are shown in Figure 5.4.

• Measured BUPT-5 & Simulated Mocap-5

BUPT-5 is a measured radar MD dataset with five human activities, i.e., walking forward, running forward, jumping forward, boxing in place, and running in a circle. This dataset is composed of the measured data in Chapter 4, but the data of "sitting on a chair" is not adopted due to the lack of simulated radar data for this activity. A UWB radar *PulsON* 440 with a centre frequency of 4.0 GHz and a bandwidth of 1.7 GHz is employed to collect radar data. Six human individuals perform the five activities along the line of sight (LOS) of the radar with an aspect angle of 0 degree.

Mocap-5 is a simulated radar MD dataset with the same five activities as the *BUPT-5*. *Mocap-5* is simulated with the Motion Capture Database from Carnegie Mellon University. The simulated radar parameters are the same as those of *PulsON* 440. An ellipsoid-based human motion model [121] is constructed for the simulation.

There is a domain shift between *BUPT-5* and *Mocap-5*. The first few-shot HAR task ($\mathcal{M} \rightarrow \mathcal{B}$) is to train a DL model with a few measured radar data in *BUPT-5* ($\mathcal{D}_t$). *Mocap-5* is used as the source dataset $\mathcal{D}_s$.

• Measured Glasgow-Young-5 & Glasgow-Old-5

Glasgow-6 [122] is a public radar dataset of five indoor human activities, i.e., walking back and forth, sitting down on a chair, standing up, bending to pick up an object, and drinking from

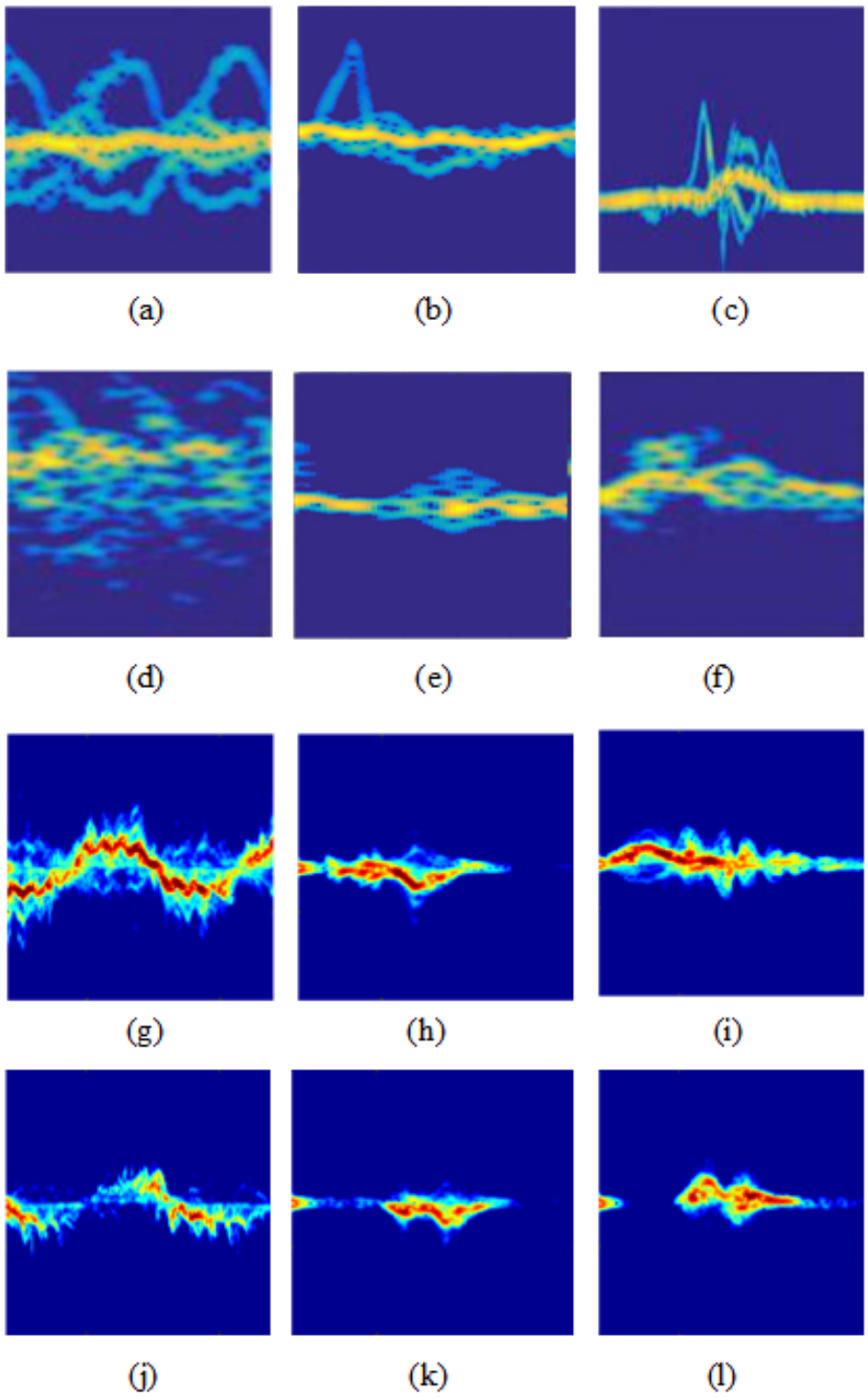


Figure 5.4: Several typical radar spectrograms used in the two few-shot HAR tasks. (a)-(c) are the spectrograms in *Mocap-5*; (d)-(f) are the spectrograms in *BUPT-5*; (g)-(i) are in *Glasgow-Young-5*; and (j)-(f) are in *Glasgow-old-5*. (a) and (d) represent 'running'; (b) and (e) represent 'boxing'; (c) and (f) represent 'jumping'; (g) and (j) represent 'walking'; (h) and (k) represent 'sitting down'; and (i) and (l) represent 'standing up'.

a cup. Volunteers ranging in age from 20 to 100 perform the five activities. We select two subsets, i.e., *Glasgow-Young-5* and *Glasgow-Old-5*, from *Glasgow-6* to evaluate the proposed method. *Glasgow-Young-5* consists of the data of five activities performed by the volunteers aging from 20 to 40, while *Glasgow-Old-5* consists of the data of five activities performed by the volunteers aging from 80 to 100 in *Glasgow-Old-5*.

Since the young and the old move differently, there are some differences between the motion data of young persons and the data of old persons. The second few-shot task ($\mathcal{G}y \rightarrow \mathcal{G}o$) is used to train a DL model to recognize old persons' activities. In this task, *Glasgow-Young-5* is employed as $\mathcal{D}_s$, while *Glasgow-Old-5* is $\mathcal{D}_t$.

5.4.2 Classification Results of FS-ADA

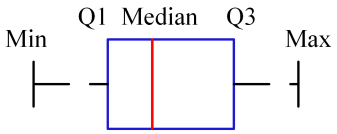
To evaluate the performance of the proposed *FS-ADA* method on the few-shot HAR tasks, we select different numbers n (n=1, 5, 10, 15, 20) of labeled samples per class from $\mathcal{D}_t$ as the target training data, and use them to train the *FS-ADA* model. The accuracies of *FS-ADA* with various numbers of labeled target samples are shown in Figure 5.5. A baseline method is used for comparison. In the baseline, the classification model is trained with $\mathcal{D}_s$, and directly used to classify target data in $\mathcal{D}_t$. No labeled target sample is utilized. The classification accuracies of the baseline on the two tasks are presented in the table of Figure 5.5(a).

Specifically, for the $\mathcal{M} \rightarrow \mathcal{B}$ task, as shown in Figure 5.5(b), the performance of *FS-ADA* is better than that of the baseline, and continuously improves with increasing the number of target samples. When there are 20 samples per class, the proposed *FS-ADA* achieves an accuracy of 84.5%, outperforming the baseline by 14.2%. Especially, when there is only one labeled training sample per class, the proposed method can still yield a classification accuracy of 73.3%, outperforming the baseline by 3.0%. For the $\mathcal{G} \rightarrow \mathcal{G}$ task, as illustrated in Figure 5.5(c), the accuracy of *FS-ADA* also exhibits an upward trend as the number of labeled target samples increases. In particular, when there is one sample per class, *FS-ADA* achieves an average accuracy of 84.3%, outperforming the baseline with a margin of 3.8%.

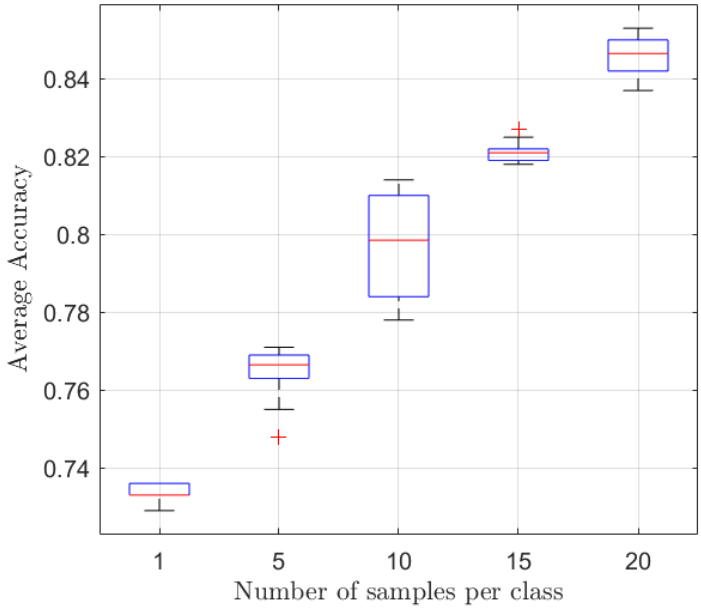
5.4.3 Comparison with State-of-the-art few-shot Methods

We also compare our proposed *FS-ADA* method with the following state-of-the-art supervised FSL methods.

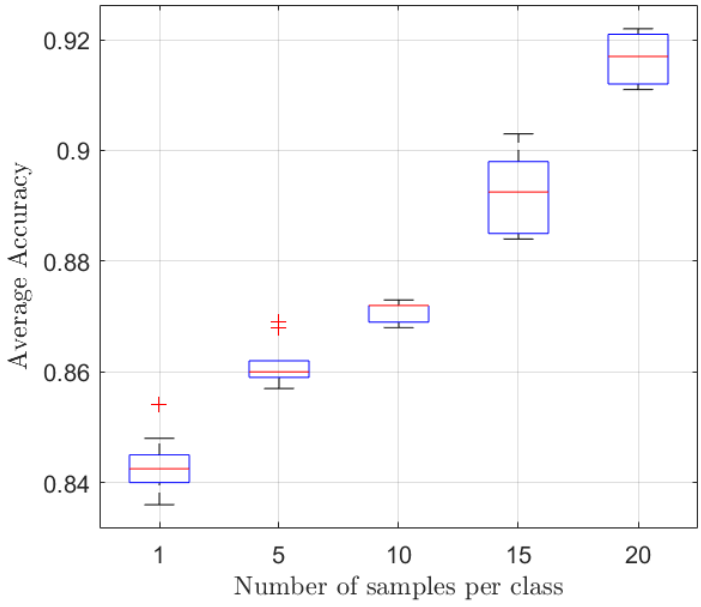
	Baseline
$\mathcal{M} \rightarrow \mathcal{B}$	0.703
$\mathcal{G}y \rightarrow \mathcal{G}o$	0.805



(a) The classification accuracies of the baseline model on the two HAR tasks (*left*), and the description of box plot (right).



(b) Classification results of *FS-ADA* on the $\mathcal{M} \rightarrow \mathcal{B}$ task with varying numbers n of target samples per class.



(c) Classification results of *FS-ADA* on the $\mathcal{G}\dagger \rightarrow \mathcal{G}\wr$ task with varying numbers n of target samples per class.

Figure 5.5: Performance of the proposed *FS-ADA* on the two HAR tasks.

Table 5.1: Performance Comparison of *FS-ADA* with Several Few-Shot Methods

Number of labeled samples n		1	5	10	15	20
$\mathcal{M} \rightarrow \mathcal{B}$	FT [34]	0.718	0.741	0.769	0.797	0.830
	DTDA [120]	0.731	0.747	0.765	0.789	0.838
	FADA [113]	0.724	0.750	0.784	0.820	0.841
	CCSA [114]	0.727	0.756	0.782	**0.828**	**0.849**
	Ours	**0.733**	**0.764**	**0.797**	0.821	0.845
$\mathcal{G}\dagger \rightarrow \mathcal{G}\wr$	FT [34]	0.810	0.826	0.842	0.882	**0.918**
	DTDA [120]	0.834	0.821	0.854	0.875	0.902
	FADA [113]	0.829	0.844	0.868	0.885	0.906
	CCSA [114]	0.836	0.857	**0.874**	0.889	0.914
	Ours	**0.843**	**0.861**	0.871	**0.892**	0.916

- *FT* [34] is a typical transfer learning methods where a DL model is first trained with the source data, and then fine-tuned with a small number of labeled target samples.
- *DTDA* [120] is a supervised domain adaptation method that uses a domain invariance optimization algorithm and a distribution matching loss for inter-domain transfer.
- *FADA* [113] is a supervised ADA method that augments the typical binary adversarial discriminator to distinguish four different classes.
- *CCSA* [114] integrates the classification loss and the contrastive semantic-alignment loss into a DL method for supervised domain adaptation.

The classification accuracies of these methods on the two tasks are shown in Table 5.1. It can be seen that, for both of the two HAR tasks, our proposed *FS-ADA* can achieve the best or nearly best performance. In particular, it shows excellent performance in term of the classification accuracy where there are only a small number of labeled training data, *e.g.*, n=1, 5, and 10. When the number of labeled samples increases, the performance of all the five methods improves. However, for the $\mathcal{M} \rightarrow \mathcal{B}$ task, *CCSA* achieves the best performance when n is larger than 10. It is indicated that compared with the adopted training scheme in *FS-ADA*, combining

Table 5.2: Complexity Comparison of *FS-ADA* with Several Few-Shot Methods

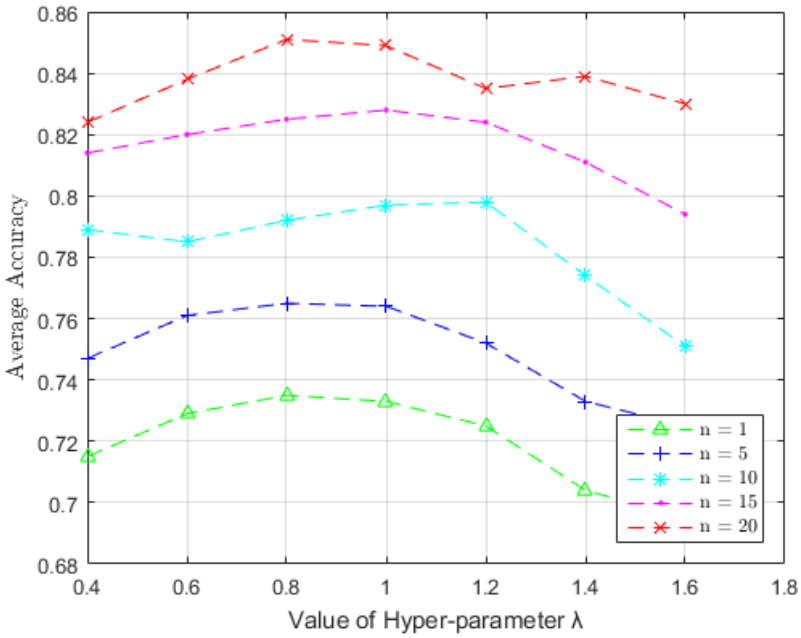
Method	FLOPs	Params
FT [34]	8.3G	20M
DTDA [120]	2.41G	270M
FADA [113]	10.26M	2.13M
CCSA [114]	9.24M	1.11M
Ours	**9.21M**	**1.08M**

the classification loss and the contrastive semantic-alignment loss in *CCSA* can extract more discriminative features for HAR when there are more training samples. For the $\mathcal{G} \rightarrow \mathcal{G}$ task, *FT* and the proposed *FS-ADA* achieve almost the same classification performance when $n = 20$, showing that fine-tuning a trained model with enough labeled samples can also achieve good HAR performance.

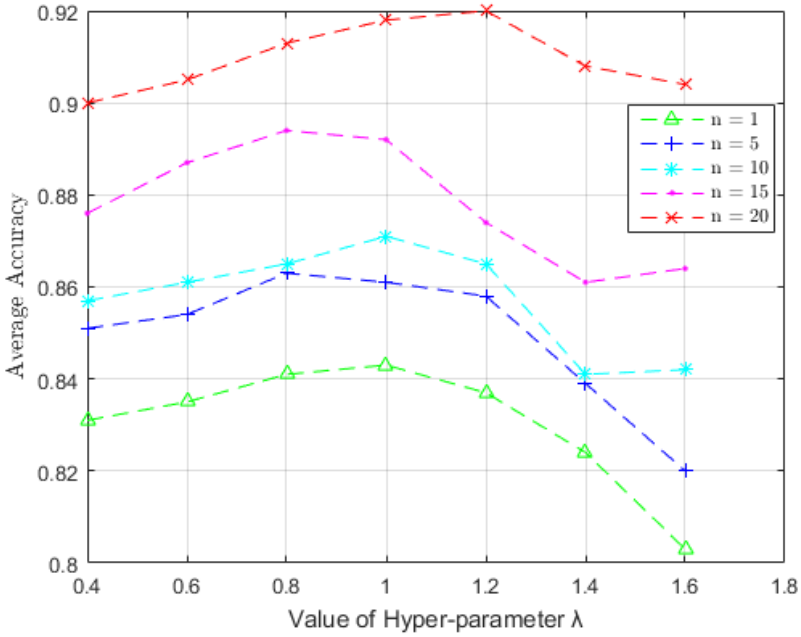
Furthermore, to compare the time complexity and the space complexity of these methods, we calculate the FLOPs and the number of parameters of the models, as listed in Table 5.2. It can be seen that compared with the other methods, the proposed *FS-ADA* has lower FLOPs, indicating *FS-ADA* has a lower time complexity. Besides, *FS-ADA* has the lowest number of parameters, which can significantly save storage space.

5.4.4 Sensitivity of Hyper-parameter

To show the impact of the hyper-parameter λ on the HAR performance of *FS-ADA*, we perform the sensitivity analysis experiment on λ. The experimental results are shown in Figure 5.6. It can be seen that when the value of λ increases from 0.4 to 1.6, the classification accuracy of *FS-ADA* first goes up and then goes down on both of the two tasks. A better HAR performance is achieved when λ is set to 0.8 to 1.2, compared to the other λ values. Furthermore, on the two tasks, the classification accuracy of *FS-ADA* fluctuates more as the value of λ changes when n is smaller. It is indicated that the smaller the value of n is, the more sensitive the performance of *FS-ADA* is to the variation of λ. Besides, when λ is set to a value between 1.4 and 1.6, *FS-ADA* cannot achieve satisfactory performance. This is because when λ is large, the HAR task dominates the training process and impacts model parameters more than the



(a) Classification results of *FS-ADA* on the $\mathcal{M} \rightarrow \mathcal{B}$ task with different values of λ.



(b) Classification results of *FS-ADA* on the $\mathcal{G}\dagger \rightarrow \mathcal{G}\wr$ task with different values of λ.

Figure 5.6: Performance variation of the proposed *FS-ADA* on the two HAR tasks with different values of λ. n refers to as the number of samples per class.

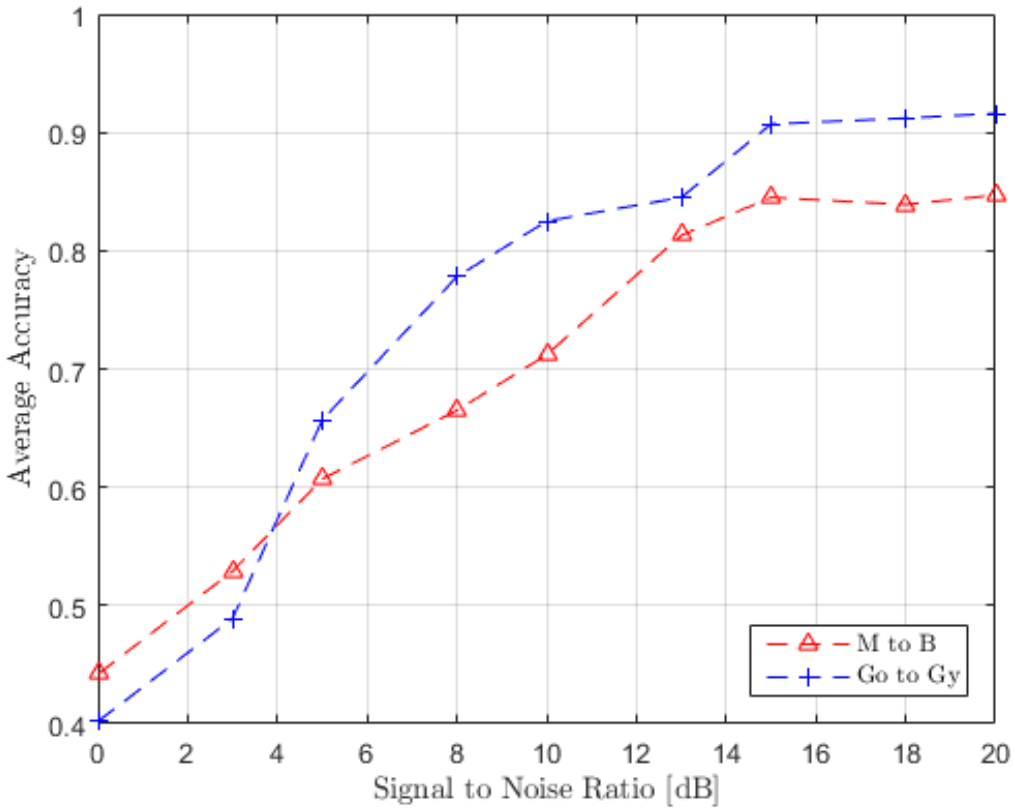


Figure 5.7: Performance variation of the proposed *FS-ADA* on the two tasks with different levels of SNR.

domain discrimination task. However, when there are fewer target samples for training, fewer features could be extracted from the HAR task than from the domain discrimination task.

5.4.5 Impact of SNR

To investigate the impact of noise on the performance of *FS-ADA*, we add different intensities of additive white Gaussian noise (AWGN) to the target training data and use the noisy target data to perform experiments. Figure 5.7 shows the performance variation of the proposed *FS-ADA* for the two tasks when different levels of AWGN are added. In Figure 5.7, the accuracies for the two tasks increase with the improvement of SNR of the noisy target data. When the SNR is 0 dB, the model fails to classify the radar spectrograms due to the strong noise interference. When the SNR rises to 10 dB, an obvious improvement of performance is shown, with an accuracy of approximately 82.5% for $\mathcal{G} \rightarrow \mathcal{G}$ and 71.2% for $\mathcal{M} \rightarrow \mathcal{B}$, respectively. When the SNR reaches 20 dB, the spectrograms are so clear that the model can classify the spectrograms in the two tasks with high accuracies.

5.5 Summary

In this chapter, a supervised domain adaptation method for few-shot radar-based human activity recognition has been proposed. The technique consists of two feature extractors that share weights and a multi-class discriminator. The multi-class discriminator network combines the activity classifier and the domain discriminator for extracting domain-invariant and category-discriminative features. A multitask generative adversarial loss has also been proposed to optimize the extractors together with the discriminator.

We conducted experiments on two radar-based HAR tasks. Experimental results demonstrated the superiority of the proposed method for few-shot HAR. We also compared the proposed method with several state-of-the-art FSL methods. Comparison results showed the proposed *FS-ADA* could achieve better HAR performance than the state-of-the-art. Finally, the analysis of hyper-parameter sensitivity and the impact of SNR on *FS-ADA* was performed. In the future, we will focus on the few-shot HAR problem in more complicated environments, such as multiple-person and through-wall scenarios.

Chapter 6

Conclusions and Future Work

6.1 Summary

This dissertation studied contact-free human activity sensing with wireless signals, including Doppler speeds estimation of moving human target with WiFi CSI, cross-target HAR with limited radar MD spectrograms, and supervised domain adaptation for few-shot radar-based HAR. The relevant literature review can be found in Chapter 2, and the innovative research results achieved in this thesis are summarised as follows.

In Chapter 3, we mainly proposed three Doppler frequency estimation algorithms based on the CSI ratio across antennas for applications involving sensing of moving targets, such as human activity recognition and mobile tracking. Among them, the signal difference-based algorithm has the best and the most robust performance for Doppler frequency estimation, while the periodicity-based algorithm is the easiest to implement. Furthermore, the Mobius-based method can estimate both the sign and value of Doppler frequencies. Experiments demonstrated that the best solution might be combining the strengths of the three algorithms.

In Chapter 4, we introduced an instance-based TL approach *ITL* for radar-based cross-target activity recognition. *ITL* is composed of three interconnected and necessary parts (*MNet* pre-training, CSDS and ACFT) rather than a collection of three distinct pieces. Experimental results showed that the proposed *ITL* could scale well to recognize different persons' activities. When it is trained to recognize a new person's activities, it can still achieve good performance on the previous HAR task, effectively alleviating the catastrophic forgetting problem.

Chapter 5 proposed a supervised domain adaptation method for few-shot radar-based human activity recognition. The technique consists of two feature extractors that share weights and a multi-class discriminator. A multitask generative adversarial loss was also proposed for optimizing the extractors and the discriminator. Experimental results demonstrated the superiority of the proposed method for few-shot HAR. We also compared the proposed method with several state-of-the-art FSL methods. Comparison results showed the proposed *FS-ADA* could achieve better HAR performance than the state-of-the-art.

6.2 Future Work

The work on human sensing with wireless signals can be potentially enriched in, but not limited to, the following various aspects.

1) Multi-person sensing: When multiple moving persons are in the sensing area, identifying the target of interest or recognizing the activities of multiple persons with wireless signals is a challenging problem. The general idea is to extract the reflected signals of each person and then identify the corresponding human activity with the separated signals. This can be typically realized in two strategies: separation via physical location and moving speed [88], or separation via signal statistics [21]. In the first strategy, the signals for different targets may be separated from the spatial dimension by using the range, angle, and/or moving speed information, which requires high resolutions in these domains. In the second strategy, signals from multiple humans may be modeled as a linear sum of statistically-independent signals, and the separation can be cast as a blind source separation problem.

2) Through-the-wall HAR: Sensing through walls is also a challenging but essential task in HAR. RF signals generally experience unpredictable reflection and absorption as they pass through walls, significantly weakening the receiving signals and hence reducing the HAR information. Furthermore, the characteristics of human activities in the received signals can be overwhelmed by environmental noise, affecting the subsequent feature extraction, especially for similar activities with nuanced differences. In this case, approaches that can classify human activities behind the walls with less signal processing and human intervention are required [123]. On the other hand, model-based algorithms can be explored to characterize the mathematical relationship between the received sensing signals and the human activities in the non-line-of-sight through-the-wall environment.

3) HAR robustness and generalization: HAR with wireless signals is sensitive to many factors such as the sensing environment, network settings, relative location of the human target, geometry, and mobility situations. For instance, since different moving directions and orientations of the person concerning the transceivers can result in various Doppler/micro-Doppler frequencies, improving the system generalization on recognizing human activities from diverse directions is challenging. Additionally, due to the unique behavior of each individual, it is essential to generalize the trained HAR algorithms when new persons or new environments emerge.

4) Distributed sensing: With the potentially significant improvement in coverage and HAR accuracy, sensing based on a distributed topology is the general trend. However, research on sensing with off-the-shelf wireless devices under a distributed topology is still minimal. The challenges for distributed HAR mainly lie in deployment and cooperation between transceivers, fusion strategies on data from diverse receivers, and scheduling issues with target return. In addition, there is almost no discussion on the HAR performance bound for distributed sensing networks yet, which is also a promising research direction.

Bibliography

[1] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "Wifi csi based passive human activity recognition using attention based blstm," *IEEE Transactions on Mobile Computing*, vol. 18, no. 11, pp. 2714–2724, 2019.

[2] Z. Wang, Z. Huang, C. Zhang, W. Dou, Y. Guo, and D. Chen, "CSI-based human sensing using model-based approaches: a survey," *Journal of Computational Design and Engineering*, vol. 8, no. 2, pp. 510–523, Feb. 2021.

[3] Y. Zeng, D. Wu, J. Xiong, E. Yi, R. Gao, and D. Zhang, "Farsense: Pushing the range limit of wifi-based respiration sensing with csi ratio of two antennas," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–26, Sep. 2019.

[4] J. A. Zhang, K. Wu, X. Huang, Y. J. Guo, D. Zhang, and R. W. Heath Jr, "Integration of radar sensing into communications with asynchronous transceivers," *arXiv preprint arXiv:2203.16043*, 2022.

[5] M. S. Seyfioğlu, A. M. Özbayoğlu, and S. Z. Gürbüz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1709–1723, Aug. 2018.

[6] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "Dnn transfer learning from diversified micro-doppler for motion classification," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 5, pp. 2164–2180, 2019.

[7] A. Shrestha, C. Murphy, I. Johnson, A. Anbulselvam, F. Fioranelli, J. Le Kernec, and S. Z. Gurbuz, "Cross-frequency classification of indoor activities with dnn transfer learning," in *2019 IEEE Radar Conference (RadarConf)*, Boston, MA, USA, Apr. 2019, pp. 1–6.

[8] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[9] L. Wang, X. Bai, C. Gong, and F. Zhou, "Hybrid inference network for few-shot SAR automatic target recognition," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9257–9269, 2021.

[10] L. Wang, X. Bai, R. Xue, and F. Zhou, "Few-shot SAR automatic target recognition based on Conv-BiLSTM prototypical network," *Neurocomputing*, vol. 443, pp. 235–246, 2021.

[11] Y. Ma, G. Zhou, and S. Wang, "Wifi sensing with channel state information: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.

[12] Z. Shi, J. A. Zhang, R. Y. Xu, and Q. Cheng, "Environment-robust device-free human activity recognition with channel-state-information enhancement and one-shot learning," *IEEE Transactions on Mobile Computing*, vol. 21, no. 2, pp. 540–554, Feb. 2022. DOI: 10.1109/TMC.2020.3012433.

[13] L. Chen, I. Ahriz, and D. Le Ruyet, "Aoa-aware probabilistic indoor location fingerprinting using channel state information," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 10 868–10 883, 2020.

[14] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," 2015.

[15] D. Wu, D. Zhang, C. Xu, H. Wang, and X. Li, "Device-free wifi human sensing: From pattern-based to model-based approaches," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 91–97, 2017. DOI: 10.1109/MCOM.2017.1700143.

[16] H. Wang, D. Zhang, K. Niu, Q. Lv, Y. Liu, D. Wu, R. Gao, and B. Xie, "Mfdl: A multicarrier fresnel penetration model based device-free localization system leveraging commodity wi-fi cards," *arXiv preprint arXiv:1707.07514*, 2017.

[17] Y. Zeng, D. Wu, J. Xiong, and D. Zhang, "Boosting wifi sensing performance via csi ratio," *IEEE Pervasive Computing*, vol. 20, no. 1, pp. 62–70, 2021. DOI: 10.1109/MPRV.2020.3041024.

[18] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei, "Indotrack: Device-free indoor human tracking with commodity wi-fi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–22, 2017.

[19] Z. Ni, J. A. Zhang, X. Huang, K. Yang, and J. Yuan, "Uplink sensing in perceptive mobile networks with asynchronous transceivers," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1287–1300, 2021.

[20] F. Zhang, Z. Chang, K. Niu, J. Xiong, B. Jin, Q. Lv, and D. Zhang, "Exploring LoRa for long-range through-wall sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–27, Jun. 2020.

[21] Y. Zeng, D. Wu, J. Xiong, J. Liu, Z. Liu, and D. Zhang, "Multisense: Enabling multi-person respiration sensing with commodity wifi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–29, Sep. 2020.

[22] V. C. Chen, D. Tahmoush, and W. J. Miceli, *Radar micro-Doppler signatures*. Institution of Engineering and Technology, 2014.

[23] Y. Kim and H. Ling, "Human activity classification based on micro-doppler signatures using a support vector machine," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 5, pp. 1328–1337, 2009.

[24] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–19, 2016.

[25] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *ACM Symposium on User Interface Software and Technology*, Tokyo, Japan, 2016, pp. 851–860.

[26] M. S. Seyfioğlu and S. Z. Gürbüz, "Deep neural network initialization methods for micro-doppler classification with low training sample support," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2462–2466, 2017.

[27] Y. Shao, Y. Dai, L. Yuan, and W. Chen, "Deep learning methods for personnel recognition based on micro-doppler features," in *9th International Conference on Signal Processing Systems*, AUT, Auckland, New Zealand, 2017, pp. 94–98.

[28] G Klarenbeek, R. I. A. Harmanny, and L Cifola, "Multi-target human gait classification using lstm recurrent neural networks applied to micro-doppler," in *European Radar Conference*, Nuremberg, Germany, 2017, pp. 167–170.

[29] B. Jokanovic, M. Amin, and F. Ahmad, "Radar fall motion detection using deep learning," in *IEEE Radar Conference (RadarConf)*, Philadelphia, PA, USA, 2016, pp. 1–6.

[30] M. S. Seyfioğlu, A. M. Özbayğglu, and S. Z. Gurbuz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1709–1723, Aug. 2018.

[31] H. T. Le, S. L. Phung, A. Bouzerdoum, and F. H. C. Tivive, "Human motion classification with micro-doppler radar and bayesian-optimized convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 2961–2965.

[32] M. Wang, Y. D. Zhang, and G. Cui, "Human motion recognition exploiting radar with stacked recurrent neural network," *Digital Signal Processing*, vol. 87, pp. 125–131, 2019.

[33] Kim, Youngwook and Toomajian, Brian, "Application of doppler radar for the recognition of hand gestures using optimized deep convolutional neural networks," in *European Conference on Antennas and Propagation*, Paris, France, 2017, pp. 1258–1260.

[34] H. Du, Y. He, and T. Jin, "Transfer learning for human activities classification using micro-doppler spectrograms," in *2018 IEEE International Conference on Computational Electromagnetics (ICCEM)*, Chengdu, China, Mar. 2018, pp. 1–3.

[35] P. Cao, W. Xia, M. Ye, J. Zhang, and J. Zhou, "Radar-id: Human identification based on radar micro-doppler signatures using deep convolutional neural networks," *IET Radar, Sonar & Navigation*, vol. 12, no. 7, pp. 729–734, 2018.

[36] Y. Shao, S. Guo, L. Sun, and W. Chen, "Human motion classification based on range information with deep convolutional neural network," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, 2017, pp. 1519–1523.

[37] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor," *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3278–3289, 2018.

[38] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, Ljubljana, Slovenia, 2015, pp. 1–8.

[39] B. Jokanović and M. Amin, "Fall detection using deep learning in range-doppler radars," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 1, pp. 180–189, 2018.

[40] B. Jokanovic, M. Amin, and B. Erol, "Multiple joint-variable domains recognition of human motion," in *2017 IEEE Radar Conference (RadarConf)*, IEEE, 2017, pp. 0948–0952.

[41] Y. He, F. Le Chevalier, and A. G. Yarovoy, "Range-doppler processing for indoor human tracking by multistatic ultra-wideband radar," in *13th International Radar Symposium (IRS)*, Warsaw, Poland, 2012, pp. 250–253.

[42] Z. Peng and C. Li, "Portable microwave radar systems for short-range localization and life tracking: A review," *Sensors*, vol. 19, no. 5, p. 1136, 2019.

[43] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[44] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1–7.

[45] Y. Sang, L. Shi, and Y. Liu, "Micro hand gesture recognition system using ultrasonic active sensing," *IEEE Access*, vol. 6, pp. 49 339–49 347, 2017.

[46] Z. Zhou, Z. Cao, and Y. Pi, "Dynamic gesture recognition with a terahertz radar based on range profile sequences and doppler signatures.," *Sensors*, vol. 18, no. 1, p. 10, Jan. 2017.

[47] D. Tahmoush, "Review of micro-doppler signatures," *IET Radar, Sonar and Navigation*, vol. 9, no. 9, pp. 1140–1146, 2015.

[48] V. C. Chen and S. Qian, "Joint time-frequency transform for radar range-doppler imaging," *IEEE Transactions on Aerospace & Electronic Systems*, vol. 34, no. 2, pp. 486–499, 1998.

[49] R. P. Trommel, R. I. A. Harmanny, L. Cifola, and J. N. Driessen, "Multi-target human gait classification using deep convolutional neural networks on micro-doppler spectrograms," in *European Radar Conference*, London, UK, 2016, pp. 81–84.

[50] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[51] Y. Kim and T. Moon, "Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks," *IEEE Geoscience Remote Sensing Letters*, vol. 13, no. 1, pp. 8–12, 2016.

[52] J. Zhang, J. Tao, and Z. Shi, "Doppler-radar based hand gesture recognition system using convolutional neural networks," in *IEEE International Conference in Communications, Signal Processing, and Systems*, Karunya Nagar, Coimbatore, India, 2017, pp. 1096–1113.

[53] H. T. Le, S. L. Phung, and A. Bouzerdoum, "Human gait recognition with micro-doppler radar and deep autoencoder," in *24th International Conference on Pattern Recognition (ICPR)*, Beijing, China, 2018, pp. 3347–3352.

[54] Y. He, P. Molchanov, T. Sakamoto, P. Aubry, F. L. Chevalier, and A. Yarovoy, "Range-doppler surface: A tool to analyse human target in ultra-wideband radar," *IET Radar Sonar & Navigation*, vol. 9, no. 9, pp. 1240–1250, 2015.

[55] B. Erol, M. Amin, Z. Zhou, and J. Zhang, "Range information for reducing fall false alarms in assisted living," in *IEEE Radar Conference*, Philadelphia, PA, USA, 2016, pp. 1–6.

[56] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Short-range fmcw monopulse radar for hand-gesture sensing," in *IEEE Radar Conference*, Washington, DC, USA, 2015, pp. 1491–1496.

[57] B. Erol and M. G. Amin, "Fall motion detection using combined range and doppler features," in *24th European Signal Processing Conference (EUSIPCO)*, IEEE, Budapest, Hungary, 2016, pp. 2075–2080.

[58] Y. Wang and A. E. Fathy, "Uwb micro-doppler radar for human gait analysis using joint range-time-frequency representation," *Proceedings of SPIE*, vol. 8734, pp. 17–25, 2013.

[59] Z. A Cammenga, G. E Smith, and C. J Baker, "Combined high range resolution and micro-doppler analysis of human gait," in *2015 IEEE Radar Conference (RadarConf)*, Arlington, USA, 2015, pp. 1038–1043.

[60] H. Chen and W. Ye, "Classification of human activity based on radar signal using 1-d convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 7, pp. 1178–1182, 2019.

[61] R. Zhao, X. Ma, X. Liu, and J. Liu, "An end-to-end network for continuous human motion recognition via radar radios," *IEEE Sensors Journal*, vol. 21, no. 5, pp. 6487–6496, 2020.

[62] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 6645–6649.

[63] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 2012, pp. 4277–4280.

[64] Y. He, X. Li, and X. Jing, "A mutiscale residual attention network for multitask learning of human activity using radar micro-doppler signatures," *Remote Sensing*, vol. 11, no. 21, p. 2584, 2019.

[65] Y. Yang, C. Hou, Y. Lang, D. Guan, D. Huang, and J. Xu, "Open-set human activity recognition based on micro-doppler signatures," *Pattern Recognition*, vol. 85, pp. 60–69, 2019.

[66] H. Du, T. Jin, Y. He, Y. Song, and Y. Dai, "Segmented convolutional gated recurrent neural networks for human activity recognition in ultra-wideband radar," *Neurocomputing*, vol. 396, pp. 451 –464, 2019.

[67] C. Ding, H. Hong, Y. Zou, H. Chu, X. Zhu, F. Fioranelli, J. Le Kernec, and C. Li, "Continuous human motion recognition with a dynamic range-doppler trajectory method based on fmcw radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6821–6831, 2019.

[68] Y. Yang, C. Hou, Y. Lang, T. Sakamoto, Y. He, and W. Xiang, "Omnidirectional motion classification with monostatic radar system using micro-doppler signatures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3574–3587, 2020.

[69] Z. Lin, K. Ji, M. Kang, X. Leng, and H. Zou, "Deep convolutional highway unit network for sar target classification with limited labeled training data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 7, pp. 1091–1095, 2017.

[70] F. Deng, S. Pu, X. Chen, Y. Shi, T. Yuan, and S. Pu, "Hyperspectral image classification with capsule network using limited training samples," *Sensors*, vol. 18, no. 9, p. 3153, 2018.

[71] S. Tian, C. Wang, H. Zhang, and B. Bhanu, "Sar object classification using the dae with a modified triplet restriction," *IET Radar, Sonar & Navigation*, vol. 13, no. 7, pp. 1081–1091, 2019.

[72] X. Zhang, Z. Wang, D. Liu, and Q. Ling, "Dada: Deep adversarial data augmentation for extremely low data regime classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 2807–2811.

[73] A. Davari, H. C. Özkan, A. Maier, and C. Riess, "Fast and efficient limited data hyperspectral remote sensing image classification via gmm-based synthetic samples," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2107–2120, 2019.

[74] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Motion classification using kinematically sifted acgan-synthesized radar micro-doppler signatures," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 4, pp. 3197–3213, 2020.

[75] W. Jiang, K. Huang, J. Geng, and X. Deng, "Multi-scale metric learning for few-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1091–1102, 2020.

[76] I. Alnujaim, D. Oh, I. Park, and Y. Kim, "Classification of micro-doppler signatures measured by doppler radar through transfer learning," in *2019 13th European Conference on Antennas and Propagation (EuCAP)*, 2019, pp. 1–3.

[77] H. Du, T. Jin, Y. Song, Y. Dai, and M. Li, "Efficient human activity classification via sparsity-driven transfer learning," *IET Radar, Sonar & Navigation*, vol. 13, no. 10, pp. 1741–1746, 2019.

[78] J. Park, R. J. Javier, T. Moon, and Y. Kim, "Micro-doppler based classification of human aquatic activities via transfer learning of convolutional neural networks," *Sensors*, vol. 16, no. 12, p. 1990, 2016.

[79] M. Abdullah Jamal, H. Li, and B. Gong, "Deep face detector adaptation without negative transfer or catastrophic forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5608–5618.

[80] A. Mallya and S. Lazebnik, "Packnet: Adding multiple tasks to a single network by iterative pruning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7765–7773.

[81] H. Du, T. Jin, Y. Song, and Y. Dai, "Unsupervised adversarial domain adaptation for micro-doppler based human activity classification," *IEEE Geoscience and Remote Sensing letters*, vol. 17, no. 1, pp. 62–66, 2019.

[82] Y. Lang, Q. Wang, Y. Yang, C. Hou, D. Huang, and W. Xiang, "Unsupervised domain adaptation for micro-doppler human motion classification via feature fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 3, pp. 392–396, 2018.

[83] Q. Chen, Y. Liu, F. Fioranelli, M. Ritchie, and K. Chetty, "Eliminate aspect angle variations for human activity recognition using unsupervised deep adaptation network," in *2019 IEEE Radar Conference (RadarConf)*, 2019, pp. 1–6.

[84] B. Tan, Q. Chen, K. Chetty, K. Woodbridge, W. Li, and R. Piechocki, "Exploiting WiFi channel state information for residential healthcare informatics," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 130–137, May 2018.

[85] Y. Cui, F. Liu, X. Jing, and J. Mu, "Integrating sensing and communications for ubiquitous IoT: Applications, trends, and challenges," *IEEE Network*, vol. 35, no. 5, pp. 158–167, Sep. 2021. DOI: 10.1109/MNET.010.2100152.

[86] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Towards dual-functional wireless networks for 6G and beyond," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2022. DOI: 10.1109/JSAC.2022.3156632.

[87] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial WiFi devices," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1118–1131, May 2017.

[88] J. Pegoraro, F. Meneghello, and M. Rossi, "Multiperson continuous tracking and identification from mm-wave micro-Doppler signatures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 2994–3009, April 2021. DOI: 10.1109/TGRS.2020.3019915.

[89] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using WiFi channel state information," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 98–104, 2017. DOI: 10.1109/MCOM.2017.1700082.

[90] J. A. Zhang, M. L. Rahman, K. Wu, X. Huang, Y. J. Guo, S. Chen, and J. Yuan, "Enabling joint communication and radar sensing in mobile networks - a survey," *IEEE*

Communications Surveys & Tutorials, early access, 2021. DOI: 10.1109/COMST.2021.3122519.

[91] D. Bleh, M. Rösch, M. Kuri, A. Dyck, A. Tessmann, A. Leuther, S. Wagner, B. Weismann-Thaden, H.-P. Stulz, M. Zink, M. Rießle, R. Sommer, J. Wilcke, M. Schlechtweg, B. Yang, and O. Ambacher, "*W* -band time-domain multiplexing FMCW MIMO radar for far-field 3-D imaging," *IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 9, pp. 3474–3484, 2017. DOI: 10.1109/TMTT.2017.2661742.

[92] S. H. Javadi and A. Farina, "Radar networks: A review of features and challenges," *Information Fusion*, vol. 61, pp. 48–55, 2020, ISSN: 1566-2535.

[93] Q. Chen, B. Tan, K. Chetty, and K. Woodbridge, "Activity recognition based on micro-doppler signature with in-home wi-fi," in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2016, pp. 1–6. DOI: 10.1109/HealthCom.2016.7749457.

[94] C. Tang, W. Li, S. Vishwakarma, F. Shi, S. Julier, and K. Chetty, "Mdpose: Human skeletal motion reconstruction using wifi micro-doppler signatures," *arXiv preprint arXiv:2201.04212*, 2022.

[95] J. A. Zhang, F. Liu, C. Masouros, R. W. Heath, Z. Feng, L. Zheng, and A. Petropulu, "An overview of signal processing techniques for joint communication and radar sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 6, pp. 1295–1315, 2021.

[96] J. Olsen, "The geometry of möbius transformations," *Rochester: University of Rochester*, 2010.

[97] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu, "Widar2. 0: Passive human tracking with a single wi-fi link," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, 2018, pp. 350–361.

[98] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in neural information processing systems*, vol. 29, Barcelona, Spain, Dec. 2016, pp. 901–909.

[99] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[100] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Utah, USA, June 2018, pp. 7132–7141.

[101] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[102] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, Lake Tahoe, USA,Dec. 2012, pp. 1097–1105.

[103] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[104] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, GA, USA, Nov. 2016, pp. 265–283.

[105] J. Ngiam, D. Peng, V. Vasudevan, S. Kornblith, Q. V. Le, and R. Pang, "Domain adaptive transfer learning with specialist models.," *arXiv preprint arXiv:1811.07056*, 2018.

[106] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, July 2017, pp. 10–19.

[107] A. Asgarian, P. Sobhani, J. C. Zhang, M. Mihailescu, A. Sibilia, A. B. Ashraf, and B. Taati, "A hybrid instance-based transfer learning method," *arXiv preprint arXiv:1812.01063*, 2018.

[108] Y. Kim, J. Park, and T. Moon, "Classification of micro-doppler signatures of human aquatic activity through simulation and measurement using transferred learning," *Radar Sensor Technology XXI*, vol. 10188, pp. 324 –329, 2017.

[109] X. Shi, Y. Li, F. Zhou, and L. Liu, "Human activity recognition based on deep learning method," in *2018 International Conference on Radar (RADAR)*, Brisbane, Australia, Aug. 2018, pp. 1–5.

[110] Y. He, Y. Yang, Y. Lang, D. Huang, X. Jing, and C. Hou, "Deep learning based human activity classification in radar micro-doppler image," in *2018 15th European Radar Conference (EuRAD)*, Madrid, Spain, Sep. 2018, pp. 230–233.

[111] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "Diversified radar micro-doppler simulations as training data for deep residual neural networks," in *2018 IEEE radar Conference (radarConf18)*, Oklahoma City, OK, USA, Apr. 2018, pp. 0612–0617.

[112] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[113] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto, "Few-shot adversarial domain adaptation," in *Advances in Neural Information Processing Systems (NIPS)*, Long Beach CA, USA, Dec. 2017, pp. 6670–6680.

[114] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, Venice, Italy, Oct. 2017, pp. 5715–5725.

[115] S. Feng and M. F. Duarte, "Few-shot learning-based human activity recognition," *Expert Syst. Appl.*, vol. 138, p. 112 782, 2019.

[116] M. Rostami, S. Kolouri, E. Eaton, and K. Kim, "SAR image classification using few-shot cross-domain transfer learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Long Beach, CA, Jun. 2019.

[117] Y. Kim, I. Alnujaim, and D. Oh, "Human activity classification based on point clouds measured by millimeter wave mimo radar with deep recurrent neural networks," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13 522–13 529, Mar. 2021.

[118] X. Bai, Y. Hui, L. Wang, and F. Zhou, "Radar-based human gait recognition using dual-channel deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9767–9778, 2019.

[119] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, Hawaii, USA, Jul. 2017, pp. 7167–7176.

[120] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, Santiago, Chile, Dec. 2015, pp. 4068–4076.

[121] Y. Lang, C. Hou, Y. Yang, D. Huang, and Y. He, "Convolutional neural network for human micro-doppler classification," in *Proceedings of the European Microwave Conference*, Nuremberg, Germany, Oct. 2017.

[122] F. Fioranelli, S. A. Shah, H. Li, A. Shrestha, S. Yang, and J. Le Kernec, "Radar sensing for healthcare," *Electronics Letters*, vol. 55, no. 19, pp. 1022–1024, Sep. 2019.

[123] Z. Wang, K. Jiang, Y. Hou, Z. Huang, W. Dou, C. Zhang, and Y. Guo, "A survey on CSI-based human behavior recognition in through-the-wall scenario," *IEEE Access*, vol. 7, pp. 78 772–78 793, 2019. DOI: 10.1109/ACCESS.2019.2922244.

Appendix A

Appendix

Based on the Mobius transform, when there is no noise, the CSI-ratio samples $R(t_k)_{k=1}^{n} = \{R(t_1), R(t_2), ..., R(t_n)\}$ are on a circle of the complex plane. Let the coordinates of the center be (A, B) and the radius be r.

Then, the distance d_k between the kth CSI-ratio sample $R(t_k) = (x_k, y_k)$ $(k = 1, 2,..., n)$ and the center C_0 can be denoted as

$$d_k^2 = (x_k - A)^2 + (y_k - B)^2. \tag{A.1}$$

The square deviation δ_k between d_k and r can be calculated as

$$\delta_k = d_k^2 - r^2 = (x_k - A)^2 + (y_k - B)^2 - r^2, \tag{A.2}$$

which should be zero for all samples on the circle.

To accurately estimate C_0, we seek the value of A, B, and r that can minimize the sum of deviation Q(A, B, r). The derivatives can be computed as

$$\begin{aligned} Q(A, B, r) &= \sum_{k=1}^{n} \delta_k^2 \\ &= \sum_{k=1}^{n} \left((x_k - A)^2 + (y_k - B)^2 - r^2\right)^2 \\ &= \sum_{k=1}^{n} (x_k^2 - 2Ax_k + A^2 + y_k^2 - 2By_k + B^2 - r^2)^2. \end{aligned} \tag{A.3}$$

Let a, b and c equal to -2A, -2B, and $A^2 + B^2 - r^2$, respectively, then

$$Q(A, B, r) = Q(a, b, c) = \sum_{k=1}^{n} (x_k^2 + y_k^2 + ax_k + by_k + c)^2. \tag{A.4}$$

It can be seen from Eq. (A.4), Q(a, b, c) reaches its minimum when the partial derivatives of this function with respect to a, b and c are all equal to 0, which are given by

$$\frac{\partial Q(a, b, c)}{\partial A} = \sum_{k=1}^{n} 2(x_k^2 + y_k^2 + ax_k + by_k + c)x_k = 0, \tag{A.5}$$

$$\frac{\partial Q(a, b, c)}{\partial B} = \sum_{k=1}^{n} 2(x_k^2 + y_k^2 + ax_k + by_k + c)y_k = 0, \tag{A.6}$$

and

$$\frac{\partial Q(a, b, c)}{\partial r} = \sum_{k=1}^{n} 2(x_k^2 + y_k^2 + ax_k + by_k + c) = 0 \tag{A.7}$$

We construct $n\times$ Eq. (A.5) - $\sum_{k=1}^{n} x_k \times$ Eq. (A.7) to cancel out c, and obtain

$$\begin{aligned}(n\sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} x_k)a + (n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k)b \\ + n\sum_{k=1}^{n} x_k^3 + n\sum_{k=1}^{n} x_k y_k^2 - \sum_{k=1}^{n}(x_k^2 + y_k^2)\sum_{k=1}^{n} x_k = 0.\end{aligned} \tag{A.8}$$

Similarly, computing $n\times$ Eq. (A.6) - $\sum_{k=1}^{n} y_k \times$ Eq. (A.7) leads to

$$\begin{aligned}(n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{N} y_k)a + (n\sum_{k=1}^{n} y_k^2 - \sum_{k=1}^{n} y_k \sum_{k=1}^{n} y_k)b \\ + n\sum_{k=1}^{n} x_k^2 y_k + n\sum_{k=1}^{n} y_k^3 - \sum_{k=1}^{n}(x_k^2 + y_k^2)\sum_{k=1}^{n} y_k = 0.\end{aligned} \tag{A.9}$$

Rewrite Eq. (A.8) and (A.9) as

$$\zeta_1 a + \zeta_2 b + \zeta_3 = 0, \tag{A.10}$$

and

$$\zeta_2 a + \zeta_4 b + \zeta_5 = 0, \tag{A.11}$$

where

$$\begin{aligned}
\zeta_1 &\triangleq n\sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k;\\
\zeta_2 &\triangleq n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k;\\
\zeta_3 &\triangleq n\sum_{k=1}^{n} x_k^3 + n\sum_{k=1}^{n} x_k y_k^2 - \sum_{k=1}^{n}(x_k^2+y_k^2)\sum_{k=1}^{n} x_k;\\
\zeta_4 &\triangleq n\sum_{k=1}^{n} y_k^2 - \sum_{k=1}^{n} y_k y_k;\\
\zeta_5 &\triangleq n\sum_{k=1}^{n} x_k^2 y_k + n\sum_{k=1}^{n} y_k^3 - \sum_{k=1}^{n}(x_k^2+y_k^2)\sum_{k=1}^{n} y_k.
\end{aligned} \tag{A.12}$$

Based on (A.10) and (A.11), a and b can be computed as

$$a = \frac{\zeta_5\zeta_2 - \zeta_3\zeta_4}{\zeta_1\zeta_4 - \zeta_2^2}, \tag{A.13}$$

$$b = \frac{\zeta_5\zeta_1 - \zeta_3\zeta_2}{\zeta_2^2 - \zeta_4\zeta_1}. \tag{A.14}$$

Then,

$$c = -\frac{\sum_{k=1}^{n}(x_k^2+y_k^2) + a\sum_{k=1}^{n} x_k + b\sum_{k=1}^{n} y_k}{n} \tag{A.15}$$

Finally, the coordinate [A, B] of C_0 and the radius r can be estimated as

$$\begin{aligned}
A &= -a/2,\\
B &= -b/2,\\
r &= \frac{1}{2}\sqrt{a^2+b^2-4c}.
\end{aligned} \tag{A.16}$$