

Cross-Domain Multi-Person Human Activity Recognition via Near-Field Wi-Fi Sensing

Xin Li, *Member, IEEE*, Jingzhi Hu, *Member, IEEE*, Yinghui He, *Member, IEEE*,
Hongbo Wang, *Graduate Student Member, IEEE*, Jin Gan, and Jun Luo, *Fellow, IEEE*

Abstract—Wi-Fi-based human activity recognition (HAR) provides substantial convenience and has emerged as a thriving research field, yet the coarse spatial resolution inherent to Wi-Fi significantly hinders its ability to distinguish multiple subjects. By exploiting the near-field domination effect, establishing a dedicated sensing link for each subject through their personal Wi-Fi device offers a promising solution for multi-person HAR under native traffic. However, due to the subject-specific characteristics and irregular patterns of near-field signals, HAR neural network models require fine-tuning (FT) for cross-domain adaptation, which becomes particularly challenging with certain categories unavailable. In this paper, we propose *WiAnchor*, a novel training framework for efficient cross-domain adaptation in the presence of incomplete activity categories. This framework processes Wi-Fi signals embedded with irregular time information in three steps: during pre-training, we enlarge inter-class feature margins to enhance the separability of activities; in the FT stage, we innovate an *anchor matching* mechanism for cross-domain adaptation, filtering subject-specific interference informed by incomplete activity categories, rather than attempting to extract complete features from them; finally, the recognition of input samples is further improved based on their feature-level similarity with anchors. We construct a comprehensive dataset to thoroughly evaluate *WiAnchor*, achieving over 90% cross-domain accuracy with absent activity categories.

Index Terms—Wi-Fi sensing, multi-person sensing, human activity recognition, domain adaptation, imbalanced learning.

I. INTRODUCTION

WITH the ubiquitous deployment of its infrastructure, Wi-Fi has become an indispensable part of modern life [1]. The ubiquity of Wi-Fi, in turn, sparks significant interest in various research fields, prompting extensive exploration in multiple directions [2], [3]. Among these, Integrated Sensing and Communications (ISAC) [4], which seeks to harness Wi-Fi’s sensing capabilities rather than merely treating it as a convenient communication medium, has attracted considerable attention from both academia and industry due to its promising application potential [5]–[12]. In particular, Wi-Fi sensing refers to inferring environment conditions or human activities from variations in signal amplitude and phase

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible. This research is supported by The National Research Foundation Singapore and Infocomm Media Development Authority under its Future Communications Research & Development Programme, and MOE Tier 1 grant RG16/22.

X. Li, J. Hu, Y. He, H. Wang, J. Gan and J. Luo are with the College of Computing and Data Science, Nanyang Technological University, Singapore. (email: {xin, yinghui.he, hongbo001, jin010, junluo}@ntu.edu.sg, jingzhi.hu518@gmail.com).

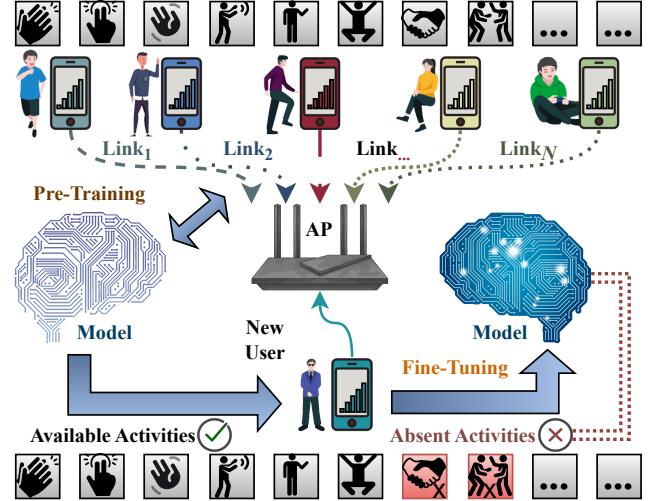


Fig. 1. Constructing dedicated links via smart devices holds promise for multi-person HAR, but the subject-specific characteristics necessitate model fine-tuning for cross-domain adaptation, which is hindered by the absence of certain activity categories.

during propagation [13]. As the pivotal enabler of Wi-Fi sensing, Channel State Information (CSI) [14] provides an easily accessible signal representation that propels the technology into a wide range of applications, including localization [5], [6], human activity recognition (HAR) [7]–[10], and vital sign monitoring [11], [12]. Among the various applications, HAR stands at the forefront, offering substantial practical values for diverse important scenarios, such as augmented/virtual reality (AR/VR) [15] and health emergency detection [16].

However, conventional HAR models of Wi-Fi sensing, such as Widar3.0 [8], only enable single-person rather than multi-person HAR, which cannot address the increasingly complex requirements of real-world scenarios. This is because multi-person HAR demands sufficient spatial resolution to distinguish different subjects, which is inherently constrained by the limited channel bandwidth of Wi-Fi systems. Moreover, since the primary task of Wi-Fi systems remains communication, excessive expansion of channel bandwidth is particularly restricted to reduce co-channel interference [17], [18]. In this regard, some studies explore makeshift approaches, including decomposing CSI into multiple source components [19] or employing deep neural networks to overfit CSI [9] for multi-person recognition; yet these approaches struggle to generalize beyond a small number of subjects due to the lack of *physical-layer diversity*. Another line of work collects signals across different antenna arrays [6], [20], [21] or channels [22]–[24]

to compensate the limited bandwidth with spatial or temporal diversity, though it often requires complex system modifications and may disrupt normal communications. Consequently, developing a realistic multi-person Wi-Fi HAR method is a essential and urgent step toward realizing the ISAC ambition.

Fortunately, the ubiquity of Wi-Fi-connected personal smart devices, such as smartphones, enables a promising framework for multi-person sensing through multi-link utilization, as shown in Fig. 1 (upper panel). While early studies [25], [26] have demonstrated that constructing multiple links using several fixed devices can marginally improve spatial resolution, they overlook a critical *near-field domination effect*: given the close proximity of each subject to its Wi-Fi-connected user equipment (UE), the activity-induced CSI impact on the Wi-Fi link is sufficiently strong to make interference from other subjects negligible [27]–[29]. This effect implies a unique correspondence between a subject and the link that its UE established with the access point (AP), making multi-person HAR feasible with commercial off-the-shelf (COTS) devices under prevalent communication configurations [30].

Despite their considerable promise, the HAR models that rely on the near-field domination effect still face several inherent challenges, as illustrated in Fig. 1 (lower panel). First, unlike existing Wi-Fi sensing systems that benefit from a high and regular CSI sampling rate (up to 1000 frames/s [5], [31]), the frame arrival rate per link in multi-user default communication scenarios is significantly lower and highly irregular due to the contention-based multi-access nature of Wi-Fi, which severely degrades sensing performance. Second, the strong subject-specific characteristics of near-field channel samples necessitate the calibration, i.e., fine-tuning (FT), of pre-trained models with the CSI samples for all potential activity classes to achieve cross-domain adaptation [32]. Nevertheless, in real-world scenarios, having users perform all classes of activities for extensive sample collection is impractical due to concerns about user experience and safety [33], while the use of a limited number of samples from an incomplete set of activity classes hinders effective FT. Last but not least, currently available datasets [8], [9], [34]–[46] have no support for the near-field multi-person HAR under default communication configurations, and are built with NICs that implement outdated Wi-Fi standards.

To address these challenges, we propose a novel framework, *WiAnchor*, which facilitates efficient cross-domain adaptation using only a small number of samples and under the complete absence of samples for certain activity categories, thereby enabling real-world deployment of Wi-Fi-based multi-person HAR. Specifically, we design a time information embedding algorithm that encodes the highly non-uniform frame arrival time into temporal features. During the pre-training (PT) stage, we introduce an inter-class margin enlarging strategy to encourage the HAR neural model to extract discriminative activity features. During the FT stage, features from subsampled portion of the PT dataset are used as anchors, and target domain features are encouraged to align with them to filter out subject-specific interference. In the inference phase, we adopt a composite strategy that combines model logits with feature similarity to the anchors to yield accurate activity

recognition. Finally, we construct a comprehensive dataset with approximately 65,000 samples and conduct a thorough evaluation of the proposed framework using it. In summary, our main contributions are:

- We present a practical method for multi-person HAR based on near-field domination effect, leveraging COTS Wi-Fi devices without requiring hardware modifications.
- We design a time information embedding algorithm to effectively capture and represent the highly non-uniform CSI sampling patterns.
- We propose WiAnchor framework, which facilitates efficient cross-domain adaptation in the absence of certain categories via a two-stage training strategy and a composite decision mechanism.
- We construct the first multi-person near-field sensing dataset, containing approximately 65,000 samples collected under default communication configurations, with a diverse set of subjects and environments.
- We conduct a comprehensive evaluation of WiAnchor framework, showing a 56.8% improvement in recognition accuracy for categories without FT samples and an overall accuracy exceeding 90%.

The rest of our paper is structured as follows: Section II presents theoretical and practical evidence for the near-field domination effect in multi-person HAR, along with associated challenges. Section III formulates our WiAnchor framework. Section IV details the experiment setup and dataset construction. Section V presents the evaluation results. The conclusion and discussion are presented in Section VI.

II. WI-FI SENSING UNDER NEAR-FIELD DOMINATION

In this section, we first introduce the fundamentals of Wi-Fi sensing and analyze existing studies. We then demonstrate the feasibility of multi-person sensing under the near-field domination effect. Finally, we present experiments that illustrate the challenges and potential solutions for fine-tuning HAR models in the absence of certain activity categories.

A. Wi-Fi Sensing Basics

We begin with a general Wi-Fi sensing system, comprising an AP-UE pair and multiple sensed subjects within the wireless network. The k -th path in this system at time t can be described by the tuple $(\tau_{k,t}, \theta_{k,t})$, where τ and θ are the *time of flight* (ToF) and *angle of arrival* (AoA), respectively. The CSI $[\mathbf{H}]_{n,m,t} = h_{n,m,t}$ received at AP can be modeled as:

$$\begin{aligned} h_{n,m,t} &= \sum_{k=1}^K \alpha_{n,m,k,t} \cdot h_{m,k,t}^{\text{ToF}} \cdot h_{n,k,t}^{\text{AoA}} + \zeta_t \\ &= \sum_{k=1}^K \alpha_{n,m,k,t} e^{-i2\pi(f_c \pm (m-1)f_b)\tau_{k,t}} e^{-i2\pi(n-1)d \cos(\theta_{k,t})\frac{f_c}{c}} + \zeta_t, \end{aligned} \quad (1)$$

where AP antennas are linearly arranged with a spacing of d , n and m respectively index the antenna and subcarrier, f_c and f_b respectively denote channel centre frequency and subcarrier bandwidth, α represents channel gain, c is the speed

of light, and ζ indicates noise introduced by the environment and hardware. For multi-person sensing, the multipath components in the Wi-Fi system need to be distinguished to extract subject-specific information, which requires sufficient spatial resolution. According to Eqn. (1), spatial resolution can be improved by enhancing the *range* (ToF) and/or *bearing* (AoA) resolutions. Based on [47], range resolution, $\Delta L = \frac{c}{W}$, increases linearly with the effective sensing bandwidth W . Given the impracticality of excessively expanding a single channel's bandwidth, previous works [22]–[24] fuse multiple channels to attain a larger effective sensing bandwidth, thereby enhancing ΔL . Additionally, as shown in [48], bearing resolution, $\Delta\theta = \frac{\lambda}{(N-1)d}$, increases linearly with the number of antennas N , a fact leveraged by previous works [6], [20], [21] to facilitate multi-person sensing.

As forward-looking prototypes, these methods require modifications to COTS devices; hence, current Wi-Fi-based HAR research [8], [34], [36]–[46] has primarily concentrated on single-person scenarios and corresponding dataset development. Even though multi-person HAR approaches, such as FallDeFi [35], which can detect the fall of one subject in a two-person environment by applying time-frequency analysis to CSIs, their scalability to more complex multi-person scenarios remains unvalidated. Besides, WiMANS [9] claims to support HAR for up to five subjects using CSI from a system with 20 MHz bandwidth and three antennas; however, its performance is heavily dependent on the neural network's fitting capacity due to the lack of additional physical-layer information to compensate for limited frequency diversity, thereby undermining generalization. Therefore, developing a practical Wi-Fi multi-person HAR system using COTS devices is critical for advancing its deployment in real-world scenarios.

B. Feasibility of Near-Field Sensing

Fortunately, the widespread availability of personal smart devices facilitates the construction of multi-link systems for multi-person HAR. In contrast to systems composed of multiple fixed devices that essentially function as a distributed multi-antenna array, this approach establishes a dedicated link for each subject, thereby enabling sensing based on the near-field domination effect. To demonstrate the feasibility of near-field sensing, we consider a scenario in which each subject is equipped with a UE connected to an AP. The multipath signal of a given link can then be decomposed into four components: target reflections $h_i(t)$ for the i -th ($i \in [1, Q]$) subject, non-target reflections $\sum_{j \neq i}^Q h_j(t)$ from other subjects, static components $h^S(t)$ due to the environment and the line-of-sight (LoS) path, and dynamic components $h^D(t)$ resulting from surrounding movements and hardware fluctuations. Thus, Eqn. (1) can be reformulated as:

$$h(t) = h_i(t) + \sum_{j \neq i}^Q h_j(t) + h^S(t) + h^D(t), \quad (2)$$

where indices n and m are omitted for brevity. Considering that both the channel gain α and phase depend on the propagation distance, we denote the distances from the subject to the

UE and AP as L^{U,S_i} and $L^{S_i,A}$, respectively. The component $h_i(t)$ is modeled as:

$$h_i(t) = \frac{\lambda^2 \sqrt{G_i} \exp(-i2\pi(L^{U,S_i}(t) + L^{S_i,A}(t))/\lambda)}{(4\pi)^2 (L^{U,S_i}(t)L^{S_i,A}(t))^{\sigma/2}}, \quad (3)$$

where wavelength $\lambda = c/f_c$, G denotes a coefficient determined by the antenna gain and the subject's reflection properties, and $\sigma \approx 4$ according to [49]. For the i -th subject located near or within the near-field region of its associated UE (approximately 0.2 m [30]), the variation in $h_i(t)$ is primarily determined by $h_i(t)$. This phenomenon, termed the **near-field domination effect**, facilitates practical multi-person HAR.

We first provide a theoretical demonstration to support the feasibility of sensing the near-field domination effect, i.e., near-field sensing. The variation in $h_i(t)$ is quantified using the power of channel variation \mathcal{P}_i , defined as the squared magnitude of its partial derivative w.r.t. time t :

$$\begin{aligned} \mathcal{P}_i &= \left| \frac{\partial h_i(t)}{\partial t} \right|^2 \\ &\approx \frac{G_i \lambda^4 v_i^2}{(4\pi)^4 (L^{U,S_i} L^{S_i,A})^\sigma} \left[\frac{\sigma^2}{4} \left(\frac{L^{U,S_i} + L^{S_i,A}}{L^{U,S_i} L^{S_i,A}} \right)^2 + \frac{16\pi^2}{\lambda^2} \right], \end{aligned} \quad (4)$$

where t is omitted for brevity, v_i , representing the velocity of the i -th subject's motion, is simplified as $v_i = \partial L^{U,S_i} / \partial t \approx \partial L^{S_i,A} / \partial t$. The first and second terms in the bracket correspond to amplitude and phase variations, respectively. In typical 5 GHz Wi-Fi near-field sensing systems, phase variations induced by the subject dominate, rendering the amplitude-related term negligible. As an illustrative example, consider $L^{U,S_i} = 0.2$ m, $L^{S_i,A} = 5$ m, and $\lambda = 0.06$ m; in this case, the second term is over 400 times larger than the first, further justifying its omission. Thus, Eqn. (4) can be simplified as:

$$\mathcal{P}_i \approx \frac{G_i \lambda^4 v_i^2}{(4\pi)^4 (L^{U,S_i} L^{S_i,A})^\sigma} \frac{16\pi^2}{\lambda^2} = \tilde{G}_i v_i^2 (L^{U,S_i} L^{S_i,A})^{-\sigma}, \quad (5)$$

where $\tilde{G}_i = G_i (\lambda/4\pi)^2$ is considered a constant. Similarly, the power of channel variation for the j -th subject can be in the same form as $\mathcal{P}_j = \tilde{G}_j v_j^2 (L^{U,S_j} L^{S_j,A})^{-\sigma}$. Since all subjects are generally far from the AP and move at similar speeds (i.e., $L^{S_i,A} \approx L^{S_j,A}$, $v_i \approx v_j$), and the i -th subject is in the near-field of its own UE (i.e., $L^{U,S_i} < L^{U,S_j}$), the near-field domination effect ($\propto (L^{U,S_i})^{-\sigma}$) leads to $\mathcal{P}_i \gg \mathcal{P}_j$, indicating that this UE-AP link is primarily dominated by the motion of the nearby i -th subject. Thus, by sniffing CSI from different links and associating each link with a subject via its MAC address, we can effectively distinguish multiple subjects for HAR, with the further advantage of mitigating the impact of environment factors.

To provide an intuitive insight, we conduct an experiment to validate the near-field domination effect. As shown in Fig. 2(a), four subjects are seated in a meeting room, each with a UE placed 20 cm in front of them; with a 60 cm spacing between body centers, the subjects are in close proximity, corresponding to typical adult body sizes. Each subject performs a sweeping motion in turn, while their respective UEs maintain communication with the AP by streaming video. We

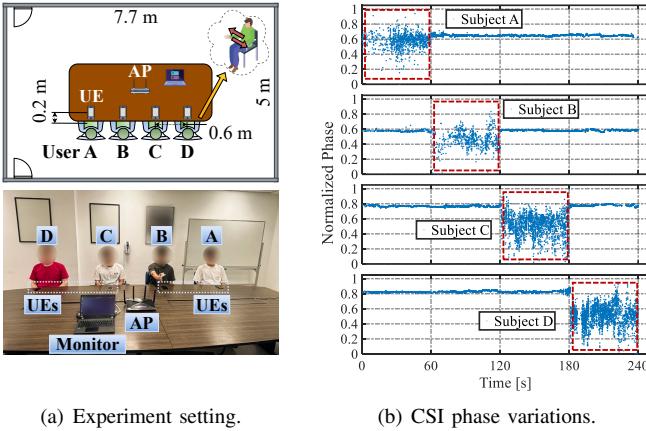


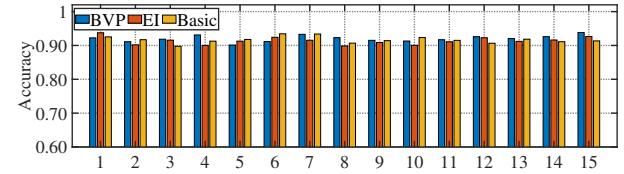
Fig. 2. Experiments on near-field sensing. The results indicate that the subject in the near-field of its corresponding UE has a dominant influence on the CSI.

sniff CSIs from all four links and show their phase variations in Fig. 2(b). The results illustrate that only the link corresponding to the active subject exhibits significant phase fluctuations, with minimal interference observed on the other links, demonstrating the practicality of near-field sensing for enabling multi-person HAR, as it effectively mitigates interference from other subjects and the environment. However, these results also indicate that, although default communication align with realistic ISAC scenarios, they introduce sample non-uniformity across links, which inevitably degrades sensing performance.

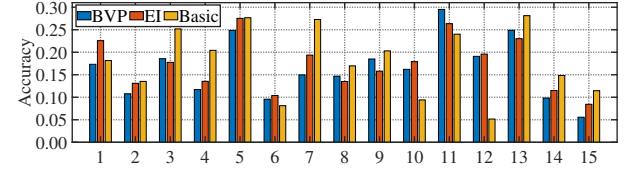
C. Fine-Tuning for Cross-Domain Adaptation

1) *Multi-person HAR via Near-field Sensing:* Due to its twofold nature, the near-field domination effect enhances CSI responsiveness to the intended subject and simplifies surrounding interference into a single-subject abstraction, while also increasing subject (domain) specificity in HAR. Conventional CSI-based HAR approaches [7], [8], [50]–[53], which are widely adopted, are initially employed in attempts to address cross-domain adaptation. Among these approaches, the first strategy [8], [50], [51] focuses on applying time-frequency transformations to CSIs in order to extract subject motion features such as speed and direction, which remain invariant across domains; for example, Widar3.0 [8] extracts a body-coordinate velocity profile (BVP) to serve this purpose. The second strategy [7], [52], [53] adopts adversarial learning to extract domain-invariant representations, as exemplified by the EI framework proposed by [7]. To evaluate their cross-domain adaptation in the context of the task considered in this work, we conduct further analyses.

For preliminary analysis, we extract data involving 2–4 concurrently active users from 15 subjects performing 10 types of activities, including gestures and body movements (see Section IV-A for details). We evaluate the models' cross-domain performance using the leave-one-out method [54]: data from one subject is used as the test set (target domain), while data from the remaining 14 subjects (source domain) is split into training and validation sets at a 9:1 ratio. In addition to the BVP and EI approaches, we also analyze the CSI using a simple GRU model (see Section III-B for details) as the basic approach; for all methods, the irregular CSI sequences are



(a) Accuracy on the source domain across different leave-one-out users.

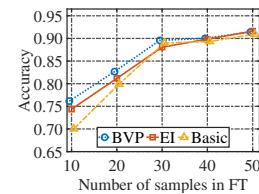


(b) Accuracy on the target domain across different leave-one-out users.

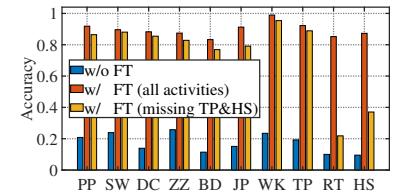
Fig. 3. Source domain accuracy vs. target domain accuracy.

interpolated to obtain uniformly structured data for processing. Fig. 3(a) illustrates the accuracy achieved in the source domain, showing that all three methods reliably exceed 90% recognition across the leave-one-user-out scenarios. However, as shown in Fig. 3(b), the accuracy in target domain drops sharply to below 30%, indicating that these approaches do not generalize well to near-field channel samples. Further analysis reveals that two factors contribute to the degradation: first, the near-field domination effect imparts subject-specific characteristics to the CSIs, causing signals from different domains to exhibit substantial physical variability; second, CSIs driven by native traffic are highly irregular and deviate significantly from the uniform traffic assumed in prior studies.

2) *Fine-Tuning with Categories Absence:* Since CSIs from various domains exhibit significant differences, fine-tuning a pre-trained model on a small subset of target domain data is an effective approach for cross-domain adaptation. As shown in Fig. 4(a), the recognition accuracy in the target domain improves steadily as the number of FT samples increases, and it saturates at around 30 samples per category, indicating that the model has acquired sufficient information. However, in practical scenarios, it is often unrealistic to assume the availability of data from every category, as the user experience burden of repeatedly performing a number of activities and other constraints may render the collection of FT data difficult or even infeasible. For example, the handshaking (HS) gesture is difficult to perform with only one person present, and the rotating (RT) action, which is often used to detect hazardous events for elderly people such as falls or medical emergencies, is not feasible or safe to collect for FT. To investigate the impact of missing category-specific data, we remove HS and



(a) Impact of sample quantity.



(b) FT performance with categories absence.

Fig. 4. FT for domain adaptation. The (a) limited number and (b) absence of samples from specific categories significantly degrade accuracy.

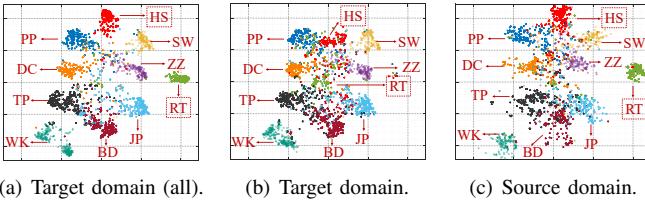


Fig. 5. Visualization with t-SNE. The absence of category-specific data negatively affects feature extraction for all categories.

RT samples from the FT process of the Basic model, while maintaining 30 samples per category for all other activities. As illustrated in Fig. 4(b), although HS and RT achieved an average recognition accuracy of 29.5%, reflecting a slight improvement over the results without FT, the performance remains substantially lower than when data from all categories are available. Nevertheless, these absent categories often correspond to activities that a HAR model must reliably recognize.

To gain deeper insights, we visualize the features using t-distributed Stochastic Neighbor Embedding (t-SNE). As shown in Fig. 5(a), the target domain features from the model fine-tuned with complete category data form well-defined clusters with distinct decision boundaries. In contrast, when RT and HS samples are excluded from FT, as shown in Fig. 5(b), these two categories become less distinguishable, and the inter-class separation of the remaining categories also diminishes, aligning with the slight drop in recognition accuracy observed in Fig. 4(b). A further examination of the source domain features under the same FT setting, as shown in Fig. 5(c), reveals that while these categories are identifiable, they remain densely packed (even compared to Fig. 5(a)), indicating that FT with incomplete categories reduces the inter-class margins. In addition, the feature distributions in Fig. 5(b) and Fig. 5(c) are not entirely consistent, suggesting that the target domain features may not be accurately extracted. Accordingly, two principal strategies for efficient FT with incomplete categories can be identified: enlarging the inter-class margins of features and shifting toward filtering subject-specific interference, rather than merely extracting incomplete features. In the following sections, we will design a training framework based on these two guiding principles.

III. METHODOLOGY

Our WiAnchor framework effectively improves the recognition accuracy of Wi-Fi-based HAR neural network models in the absence of FT samples for specific categories. As illustrated in Fig. 6, we first propose a time embedding algorithm to capture the irregular patterns of near-field CSIs. Building on this, the HAR model is processed through a three-step pipeline comprising PT, FT, and inference:

- In the PT stage, a strategy is proposed to reward the extraction of features with large inter-class margins, thereby enhancing category separability.
- In the FT stage, a small subset of source domain samples is used as anchors, and the target domain data are guided to learn subject-specific denoising characteristics driven by matched filtering.
- In the inference phase, a composite strategy combining the model logits with the similarity to anchors is introduced to further improve recognition accuracy.

The processing pipeline begins with embedding the time information for all data, as described in Section III-A. A large-scale source domain dataset is used to pre-train the neural network model, as detailed in Section III-B. A small subset of source domain data, previously used in PT stage, along with target domain data, is then employed to fine-tune the pre-trained model, as outlined in Section III-C. Finally, the fully trained model uses a composite decision strategy to recognize activity categories, as described in Section III-D. Detailed explanations are presented below.

A. Time Information Embedding

To address the temporal irregularity of near-field CSIs collected under native traffic, we propose a time embedding method to establish a solid foundation for accurate HAR. To effectively capture the temporal irregular patterns of sequences, the embedding process is divided into two components: time vector embedding and CSI data preprocessing.

We design an adaptive embedding scheme based on time differences, following this insight: in sparse regions of the sequence, long-term trends should be emphasized, while in dense regions, short-term fluctuations should be captured, rather than uniformly encoding all temporal information [55], [56]. Assuming $[\Delta t_1, \dots, \Delta t_i, \dots]$ is the time-difference vector

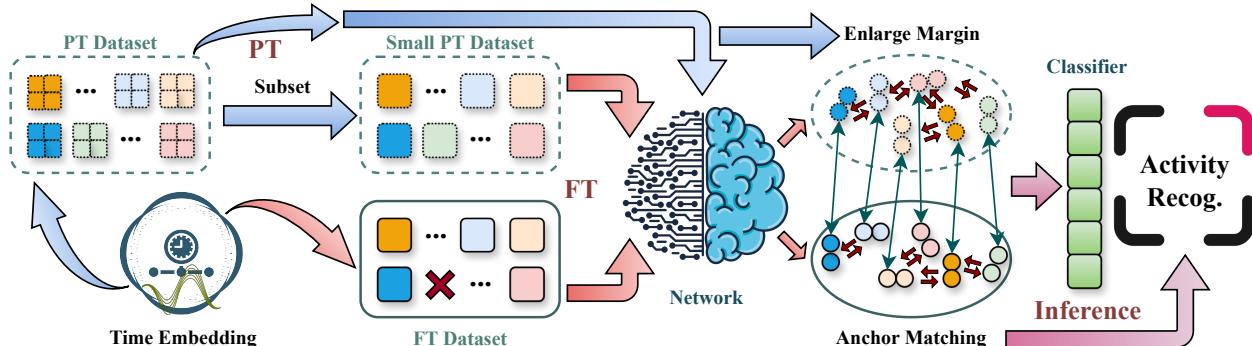


Fig. 6. WiAnchor framework overview.

obtained by differentiating the raw time vector of the received packets, the time embedding TE is then defined as:

$$\begin{cases} TE(i, 2j - 1) = \sin\left(\frac{\Delta t_i}{\mathcal{T}^{2j/D} \Delta t^{\text{Ref}}}\right) \\ TE(i, 2j) = \cos\left(\frac{\Delta t_i}{\mathcal{T}^{2j/D} \Delta t^{\text{Ref}}}\right), \end{cases} \quad (6)$$

where Δt^{Ref} denotes the reference interval, which can be obtained through statistical analysis of the data, \mathcal{T} represents the duration of the activity, and D is the embedding dimension. Given the analysis in Section II-B indicating that subject motions in the near field primarily induce variations in CSI phase, we apply a carefully designed procedure to enhance phase-related information extraction. Specifically, the signal received by the first antenna is used as a reference, and conjugate multiplication is applied to suppress interference: $\hat{h} = \bar{h}_{1,m,t} h_{n \neq 1,m,t}$. Since the CSI phase varies rapidly, leading to discontinuities due to wrapping around 0 and 2π , it is then mapped onto the continuous unit circle using sine and cosine representations, i.e., $\varphi = [\sin(\angle \hat{h}), \cos(\angle \hat{h})]$. Finally, to prevent information loss, TE , the processed phase information φ , the normalized CSI amplitude $\text{norm}(|\hat{h}|)$, and the normalized Received Signal Strength Indicator (RSSI) $\text{norm}(\text{RSSI})$ are concatenated to form a unified input representation for the neural network model:

$$x = \{TE, \text{norm}(\text{RSSI}), \text{norm}(|\hat{h}|), \varphi\}. \quad (7)$$

In addition, the input representation is formatted to a consistent shape of $T \times S$ by padding -1 values at the end, where T corresponds to the maximum number of packets collected during the activity and S is the total dimension of all previously mentioned parameters. Therefore, given $x \in \mathcal{X}$ and its ground truth label $y \in \mathcal{Y}$, the dataset is defined as $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$.

B. Pre-training Strategy

We begin with a basic network model composed of three simple components: sequence condenser, feature projection, and a classifier¹, as shown in Fig. 7. Specifically, the sequence condenser consists of two Multi-Layer Perceptron (MLP) modules, a Gated Recurrent Unit (GRU) module, and a condenser operator. The two MLPs form a lightweight encoder-decoder (ED) structure that reconstructs input information and adjusts the feature dimension, denoted as $x \rightarrow x^{\text{MLP}}$. The GRU then captures the contextual features of CSIs across continuous motions, yielding $x^{\text{MLP}} \rightarrow x^{\text{GRU}}$. To address the irregularity of CSI data under native traffic, where each sample contains a varying number of valid entries, a condenser operator extracts the final non-padded element (i.e., the last value not set to -1) from each GRU output: $x^{\text{SC}} = x^{\text{GRU}}[\tilde{t}, :]$, enabling effective temporal aggregation and reduced redundancy to avoid overfitting. Subsequently, the feature projection module built with a simple MLP compresses the features x^{SC} into a low-dimensional space. These modules jointly form the feature extractor, effectively defining the mapping

¹This basic model is empirically designed to extract a compact feature representation, and Section V-D4 further demonstrates that our training framework remains effective across diverse network architectures.

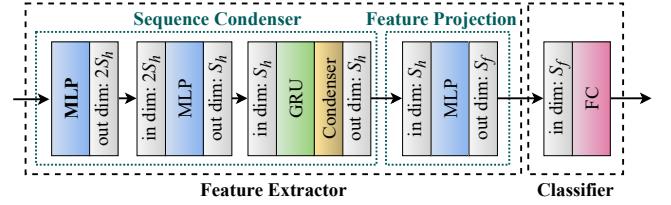


Fig. 7. Basic neural network architecture.

$\mathbb{N} = \phi^{\text{FP}}(\phi^{\text{SC}}(x))$ that preserves the compact representation of the input. Finally, a fully connected (FC) layer serves as a classifier, producing the output $y = \phi^{\text{CLS}}(\mathbb{N})$.

During the PT stage, we leverage the insight of enlarging inter-class margins, ensuring that the extracted features remain distinguishable even if these margins shrink during the subsequent FT stage. Cross-entropy (CE) loss between the one-hot encoded HAR prediction \hat{y} and ground truth label y is firstly employed to ensure effective sensing performance:

$$\mathcal{L}^{\text{CE}} = - \sum_{i=1}^C p_i \log(\hat{p}_i), \quad (8)$$

where $p_i = \frac{\exp(y_i)}{\sum_{j=1}^C \exp(y_j)}$ is the softmax output representing the predicted probability distribution and C denotes the number of activity categories. To enlarge the inter-class margins, we adopt a two-pronged strategy: extracting intrinsic and robust features of activities from the CSIs, and projecting them into a space that maximizes inter-class separability. Considering that penalizing features contributing to misclassification helps the model avoid overconfidence in specific information while promoting a more comprehensive representation of CSI features, the optimization objective is formulated as: $\sum_{i=1}^C (p_i - \hat{p}_i)^2 \cdot \|\mathbb{N}\|_2^2$. Meanwhile, with the cluster centers of features \mathbb{N}_i^C for each class, excessive inter-class similarity is penalized based on Euclidean distance, yielding to the following optimization objective: $-\frac{1}{S_h} \cdot \frac{1}{C(C-1)} \sum_{i \neq j} \|\mathbb{N}_i^C - \mathbb{N}_j^C\|_2$, where S_h is the feature dimension. Accordingly, the loss \mathcal{L}^{FE} aimed at enlarging the inter-class margins is formulated as:

$$\mathcal{L}^{\text{FE}} = \lambda_{11} \sum_{i=1}^C (p_i - \hat{p}_i)^2 \cdot \|\mathbb{N}\|_2^2 - \frac{\lambda_{12}}{S_h C(C-1)} \sum_{i \neq j} \|\mathbb{N}_i^C - \mathbb{N}_j^C\|_2, \quad (9)$$

where λ_{11} and λ_{12} are weighting parameters. The final loss \mathcal{L}^{PT} in the PT stage is defined as:

$$\mathcal{L}^{\text{PT}}(\phi^{\text{FE}}, \phi^{\text{CLS}}) = \mathcal{L}^{\text{CE}}(\phi^{\text{FE}}, \phi^{\text{CLS}}) + \mathcal{L}^{\text{FE}}(\phi^{\text{FE}}), \quad (10)$$

where $\phi^{\text{FE}} = \phi^{\text{FP}}(\phi^{\text{SC}}(\cdot))$. With the dataset \mathcal{D}^{PT} in this stage, the optimization problem is formulated as $\min_{\phi^{\text{FE}}, \phi^{\text{CLS}}} \mathbb{E}_{(x,y) \sim \mathcal{D}^{\text{PT}}} [\mathcal{L}^{\text{PT}}(\phi^{\text{FE}}, \phi^{\text{CLS}})]$.

C. Fine-tuning Strategy

During the FT stage, the training strategy aims to achieve high-performance cross-domain adaptation with the absence of certain categories. Since the gradient w.r.t. the absent categories remain at $\nabla_{\hat{p}} \mathcal{L} = 0$ during training, the model parameters cannot be updated to facilitate feature extraction in the target domain for these categories. Fortunately, since

the model has already learned to extract complete activity features in the PT stage, the FT stage can focus on filtering unseen target-domain interference, as even incomplete categories suffice to capture the subject-specific interference invariant to activities. Based on this insight, we decompose the problem into two steps: preventing catastrophic forgetting and learning additional filtering characteristics.

Catastrophic forgetting [57] is an inevitable issue in cross-domain adaptation, as the data distribution in the target domain affects the network weights, causing those related to the source domain to shift and disrupting the model's intrinsic characteristics. To address this issue, we employ a straightforward approach that leverages a subset of source domain data to ensure that the model retains its learned parameters during adaptation to the target domain. To ensure training efficiency and alleviate storage pressure, a small subset of the PT dataset, denoted as $\tilde{\mathcal{D}}^{\text{PT}}$ and comparable in size to FT dataset \mathcal{D}^{FT} , is used during the FT stage. Specifically, to avoid potential gradient conflicts from uncertain batches generated by randomly sampling the mixed dataset of $\tilde{\mathcal{D}}^{\text{PT}}$ and \mathcal{D}^{FT} , which may hinder training stability and convergence, we compute their losses $\tilde{\mathcal{L}}^{\text{PT}}$ and $\tilde{\mathcal{L}}^{\text{FT}}$ separately and then combine them into the final loss for the FT stage:

$$\mathcal{L}^{\text{FT}} = \lambda_{21}\tilde{\mathcal{L}}^{\text{PT}} + \lambda_{22}\tilde{\mathcal{L}}^{\text{FT}}, \quad (11)$$

where λ_{21} and λ_{22} are weighting parameters. We now introduce the design of each sub-loss function in detail

Since the pre-trained model already possesses the ability to extract complete activity features, we continue to apply \mathcal{L}^{PT} as the base loss function for $\tilde{\mathcal{D}}^{\text{PT}}$ during the FT stage, thereby preserving the original memory as much as possible. However, since the pre-trained model has converged to a non-trivial decision boundary, fine-tuning all parameters using limited data often results in overfitting, particularly in the final classifier layer, which is highly sensitive to distribution shifts. To mitigate this issue, we adopt a strategy that freezes the classifier while fine-tuning only the feature extractor with small learning rates. Accordingly, the loss $\tilde{\mathcal{L}}^{\text{PT}}$ is adjusted as:

$$\tilde{\mathcal{L}}^{\text{PT}}(\phi^{\text{FE}}) = \mathcal{L}^{\text{CE}}(\phi^{\text{FE}}) + \mathcal{L}^{\text{FE}}(\phi^{\text{FE}}). \quad (12)$$

Therefore, the sub-optimization problem is formulated as $\min_{\phi^{\text{FE}}} \mathbb{E}_{(\tilde{x}, \tilde{y}) \sim \tilde{\mathcal{D}}^{\text{PT}}} [\tilde{\mathcal{L}}^{\text{PT}}(\phi^{\text{FE}})]$. This design enables the neural network model to preserve its original decision boundary and extracted activity features, while gradually adapting its filtering characteristics to the target domain distribution under controlled FT.

To facilitate cross-domain adaptation, the model is designed to learn the matched filtering characteristics inherent to the target domain dataset \mathcal{D}^{FT} , essentially equivalent to suppressing interference in the extracted feature \mathbb{N} to yield an ideal representation $\tilde{\mathbb{N}}$, thereby leading to the minimization objective $\|\mathbb{N} - \tilde{\mathbb{N}}\|_2^2$. In conventional neural network training strategies, the filtering behavior can only be shaped indirectly through label supervision, limiting explicit control and ultimately hindering cross-domain adaptation. Fortunately, to mitigate catastrophic forgetting, we have intentionally introduced $\tilde{\mathcal{D}}^{\text{PT}}$, from which ideal features can be extracted to

Algorithm 1: WiAnchor framework in the FT stage for cross-domain adaptation

Input: Pre-trained model ϕ_0 , datasets $\tilde{\mathcal{D}}^{\text{PT}}$ and \mathcal{D}^{FT} , learning rate $\eta(\varphi)$, number of available activity categories C^{FT} , and training epochs \mathcal{E} .

Output: Fine-tuned model $\phi_{\mathcal{E}}$.

```

1 for  $\epsilon = 1, \dots, \mathcal{E}$  do
2   Sample batches  $(\tilde{x}, \tilde{y}) \sim \tilde{\mathcal{D}}^{\text{PT}}$  and  $(x, y) \sim \mathcal{D}^{\text{FT}}$ ;
3   Compute activity features:  $\tilde{\mathbb{N}} = \phi^{\text{FE}}(\tilde{x})$  and  $\mathbb{N} = \phi^{\text{FE}}(x)$ ;
4   Compute category-wise cluster centers:  $\tilde{\mathbb{N}}_i^C \leftarrow \tilde{\mathbb{N}}$  and  $\mathbb{N}_i^C \leftarrow \mathbb{N}$ ,  $\forall i \leq C^{\text{FT}}$  with valid category  $i$ ;
5    $\mathcal{L}^{\text{AC}} \leftarrow \text{Loss}(\mathbb{N}_i^C, \tilde{\mathbb{N}}_i^C)$  based on Eqn. (13);
6    $\tilde{\mathcal{L}}^{\text{FT}}(\phi^{\text{FE}}) \leftarrow \text{Loss}(\phi_{\epsilon}(x), y) + \mathcal{L}^{\text{AC}}$  based on Eqn. (14);
7    $\tilde{\mathcal{L}}^{\text{PT}} \leftarrow \text{Loss}(\phi_{\epsilon}(\tilde{x}), \tilde{y})$  based on Eqn. (12);
8   Update  $\phi_{\epsilon}$  based on Eqn. (15).
9 end
```

serve as anchors for learning well-behaved filtering characteristics. Nevertheless, directly matching a large number of target domain features with those from the source domain may lead to overfitting and misalignment of structural patterns across domains; therefore, we use the cluster centers of features as anchors to improve generalization. Let the activity features from the source domain be denoted as $\tilde{\mathbb{N}} = \phi^{\text{FE}}(\tilde{x})$ ($\tilde{x} \in \tilde{\mathcal{D}}^{\text{PT}}$), which are clustered by category to obtain the cluster centers $\tilde{\mathbb{N}}_i^C$ ($i \in \{1, \dots, C\}$). Similarly, the target domain activity features $\mathbb{N} = \phi^{\text{FE}}(x)$ ($x \in \mathcal{D}^{\text{FT}}$) yield cluster centers \mathbb{N}_i^C ($i \in \{1, \dots, C^{\text{FT}}\}$), where $\{1, \dots, C^{\text{FT}}\}$ and $\{C^{\text{FT}}+1, \dots, C\}$ correspond to present and absent categories, respectively. Cosine similarity is employed to measure the discrepancy between them, leading to the anchor matching loss function $\tilde{\mathcal{L}}^{\text{FT}}$ defined as:

$$\mathcal{L}^{\text{AC}} = \lambda_{23} \sum_{i=1}^{C^{\text{FT}}} \left(1 - \cos(\mathbb{N}_i^C, \tilde{\mathbb{N}}_i^C) \right), \quad (13)$$

where λ_{23} is a weighting parameter. In addition, to ensure accurate recognition of activities in the target domain, we further compute the \mathcal{L}^{CE} and \mathcal{L}^{FE} losses on the dataset \mathcal{D}^{FT} . The overall loss $\tilde{\mathcal{L}}^{\text{FT}}$ is then defined as:

$$\tilde{\mathcal{L}}^{\text{FT}}(\phi^{\text{FE}}) = \mathcal{L}^{\text{AC}}(\phi^{\text{FE}}) + \mathcal{L}^{\text{CE}}(\phi^{\text{FE}}) + \mathcal{L}^{\text{FE}}(\phi^{\text{FE}}). \quad (14)$$

We continue to adopt the strategy of freezing the classifier; accordingly, the sub-optimization problem is denoted as $\min_{\phi^{\text{FE}}} \mathbb{E}_{(x, y) \sim \mathcal{D}^{\text{FT}}} [\tilde{\mathcal{L}}^{\text{FT}}(\phi^{\text{FE}})]$. Finally, by substituting Eqns. (12) and (14) into Eqn. (11), we obtain the complete loss function for the FT stage. To finely control the training dynamics, module-specific learning rates $\eta(\varphi)$ ($\varphi \in \{\phi^{\text{SC}}, \phi^{\text{FP}}\}$) are introduced. Thus, the parameter update is expressed as:

$$\mathcal{G} = \nabla_{\phi} \left(\lambda_{21} \mathbb{E}_{(\tilde{x}, \tilde{y}) \sim \tilde{\mathcal{D}}^{\text{PT}}} \tilde{\mathcal{L}}^{\text{PT}} + \lambda_{22} \mathbb{E}_{(x, y) \sim \mathcal{D}^{\text{FT}}} \tilde{\mathcal{L}}^{\text{FT}} \right)$$

$$\phi_{(\epsilon+1)} \leftarrow \phi_{\epsilon} - \eta(\varphi) \circ \mathcal{G}, \quad (15)$$

where \circ denotes Hadamard product and ϵ indicates the iteration index. The training algorithm for the FT stage is detailed in Algorithm 1.

D. Inference Strategy

During the inference phase, a composite strategy integrating feature similarity and the predicted probability distribution is designed to enhance HAR performance. The probability distribution predicted by the softmax function essentially reflects the similarity between activity features and classifier weights [58]. However, such implicit modeling overlooks the clustering structure and geometric organization of features in the embedding space, potentially leading to decision boundaries with limited generalizability. To mitigate this issue, an explicit similarity-based mechanism is introduced to assist classification by identifying the most similar category center in the feature space, fully benefiting from the designs of both the PT and FT stages. Specifically, the fine-tuned model $\phi_{\mathcal{E}}$ first processes the deliberately constructed dataset $\tilde{\mathcal{D}}^{\text{PT}}$ and performs category-wise clustering to obtain the cluster centers \mathbb{N}_i^c ($i \in \{1, \dots, C\}$). For each test sample $\check{x} \in \mathcal{X}^{\text{Test}}$, both the predicted probability distribution \check{p}_i and the normalized similarity $\check{q}_i = \cos(\phi_{\mathcal{E}}^{\text{FE}}(\check{x}), \mathbb{N}_i^c)$ between its feature and the cluster centers from $\tilde{\mathcal{D}}^{\text{PT}}$ are computed. The final decision result is then denoted as:

$$\check{y} = \arg \max_{i \in \{1, \dots, C\}} (\check{p}_i + \lambda_3 \check{q}_i). \quad (16)$$

where λ_3 is a weighting parameter. This inference strategy captures both discriminative decision boundaries and the semantic consistency of features, thereby further improving the accuracy of HAR in the target domain.

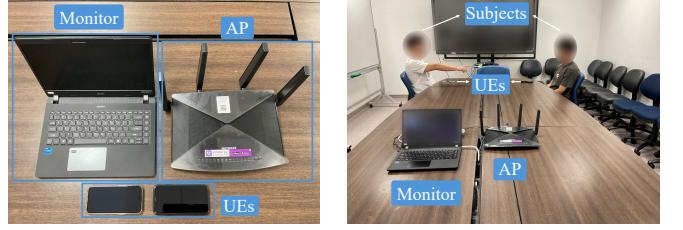
IV. NFS-FI DATASET

In this section, we first construct a Wi-Fi Near-Field Sensing (NFS-Fi) dataset², and provide a brief statistical analysis.

A. Dataset Collection

To advance Wi-Fi sensing towards practical multi-person sensing and ISAC development, we build a multi-person HAR dataset, NFS-Fi, consisting of near-field channel samples generated under native traffic, leveraging up-to-date NICs. We begin by setting up the data collection system, followed by a detailed description of the experiment setup.

Our data collection system consists of an AP and several UEs. The AP is a Netgear Nighthawk X10 router compliant with the IEEE 802.11ac standard, operating on a 5260 MHz carrier frequency with a 40 MHz channel bandwidth. The UEs are smartphones running Android or iOS, equipped with NICs compliant with the IEEE 802.11ax standard, and placed approximately 20 cm in front of the subjects to induce the near-field domination effect. During the experiment, the UEs connect to the AP and generate uplink traffic through video meetings, while the subjects engage in various activities, as shown in Fig. 8. A laptop equipped with an Intel AX210 NIC, which also adheres to the IEEE 802.11ax standard, serves as the monitor, and the PicoScenes tool [14] is employed to capture the Wi-Fi signals. Among these signals, the QoS Data packets are extracted and parsed to obtain the required sensing



(a) Hardware components. (b) Experiment setup.

Fig. 8. Data collection system.

information, including timestamp, RSSI, and CSI data. The raw CSI structure is a 2×117 complex matrix, representing the number of receiving antennas and subcarriers, respectively. Owing to its versatility, this data collection system finds applicability across diverse near-field sensing applications.

To build our dataset, we recruit 56 participants, including 36 males and 20 females, aged between 20 and 55 years, with heights ranging from 155 cm to 185 cm. Our experiment involves six different environments: meeting room (MR), lecture room (LR), discussion room (DR), classroom (CR), office room (OR), and self-study room (SR), as shown in Fig. 9. Each subject performs activities in two distinct environments: Subjects 1–16 in MR and LR, Subjects 17–36 in DR and CR, and Subjects 37–56 in OR and SR. They execute 10 activities in total, including 4 hand gestures, 2 interactive gestures, and 4 body activities. Specifically, these are push&pull (PP), sweeping (SW), drawing circle (DC), zig&zag (ZZ), typing on a phone (TP), handshaking (HS), bending (BD), jumping (JP), rotating (RT), and walking (WK). Each experiment involves 2 to 4 concurrent participants who perform the activities at their own pace, while being instructed to complete each activity within 2 seconds, followed by a short 1-second pause before starting the next round to facilitate data segmentation. In evaluations, we extract only the first 2 seconds of data for HAR. These experiments have strictly followed the IRB of our institute. Informed consent was obtained from all participants.

B. Analysis of the Dataset

Our NFS-Fi dataset is the first practical Wi-Fi multi-person sensing dataset, offering three key advantages over existing

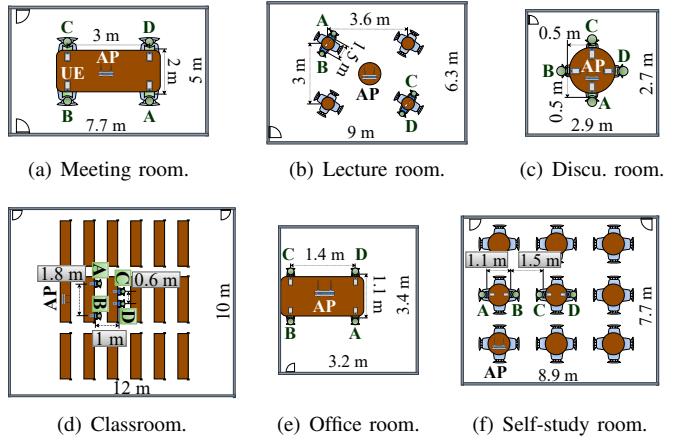


Fig. 9. Environment layouts.

²The dataset is available via <https://github.com/DeepWiSe888/NFS-Fi>

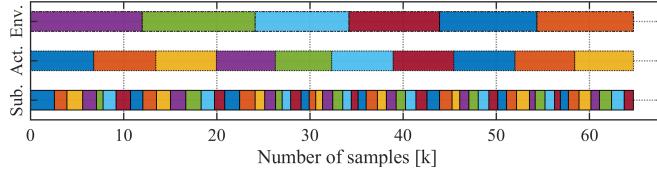


Fig. 10. Statistics of samples across subjects, activities, and environments.

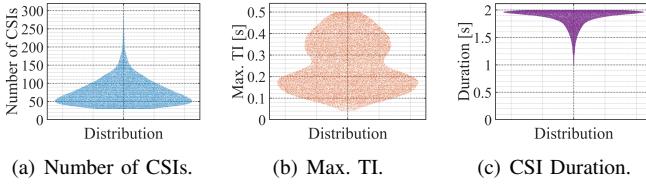


Fig. 11. Statistics of CSI entries across samples.

datasets [8], [9], [34]–[46], as summarized in Table I. First, leveraging diverse physical information, the multi-link near-field sensing strategy enables practical multi-person sensing beyond simple scenarios [35] or mere reliance on neural network fitting [9]. Second, the dataset contains native traffic from the normal operation of smart devices, without injecting evenly spaced sensing packets, thus avoiding interference with default communication and reflecting realistic conditions. Third, the dataset is built using up-to-date NICs compliant with IEEE 802.11ac/ax standards, keeping Wi-Fi sensing aligned with the latest technological developments.

Beyond those advantages, our NFS-Fi dataset provides suf-

ficient diversity to capture real-world scenarios. It comprises 64,823 samples, with Subject 1 contributing the most valid activities (2,563) and Subject 49 the fewest (624). Among all activities, PP has the most samples (6,780), while TP has the fewest (6,075). Across the six environments, LR contains the most samples (12,121) and CR the fewest (9,735). The detailed distribution is shown in Fig. 10. Furthermore, on average, each sample contains 77 CSI entries, with a maximum time interval (Max. TI) of approximately 0.25 s and an average data collection duration of 1.86 s. The CSI entry statistics for each sample are presented in Fig. 11. These results confirm that the dataset collected under native traffic conditions can effectively capture activity cycles, ensuring its usability.

V. EVALUATIONS

In this section, we conduct a comprehensive evaluation of WiAnchor framework using NFS-Fi dataset, beginning with the evaluation setup, followed by a micro-benchmark study, comparison to baselines, and analysis of impact factors.

A. Evaluation Setup

The architecture of our basic GRU model is illustrated in Fig. 7, with a hidden size of $S_h = 64$ and a single layer. In each evaluation round, one subject's data is used as the target domain, and data from 6 randomly selected subjects, excluding the target subject and the environment where that subject is recorded, serve as the source domain. A batch size

TABLE I
COMPARISON WITH PUBLIC WI-FI SENSING DATASETS FOR HUMAN ACTIVITY RECOGNITION

Dataset	Dataset Size	Concurrent Users	No. Activities	No. Participants	No. Environments	Sampling	Bandwidth (MHz)	Wi-Fi Band (GHz)	Standard
UT-HAR [34]	5173	1	7	6	1	1000 Hz	20	5	802.11n
FallDeFi [35]	1070	1–2	28	3	5	1000 Hz	20	5	802.11n
SignFi [36]	14280	1	276	5	2	200 Hz	20	5	802.11n
WiAR [37]	4800	1	16	10	3	30 Hz	20	5	802.11n
Brinke et al. [38]	4199	1	6	9	1	20 Hz	20	2.4	802.11n
Widar3.0 [8]	258575	1	16	16	3	1000 Hz	20	5	802.11n
Baha et al. [39]	9000	1	12	30	3	320 Hz	20	2.4	802.11n
CSIDA [40]	3000	1	6	5	2	1000 Hz	40	5	802.11n
OPERAnet [41]	6235	1	6	6	2	1600 Hz	20	5	802.11n
NTU-HAR [42]	2400	1	6	20	1	500 Hz	40	5	802.11n
MM-Fi [43]	1080	1	27	40	4	1000 Hz	40	5	802.11n
CSI-BERT [44]	3360	1	7	8	1	100 Hz	20	2.4	802.11n
XRF55 [45]	429000	1	55	39	4	200 Hz	20	5	802.11n
WiMANS [9]	11286	0–5	9	6	3	1000 Hz	20	2.4/5	802.11n
XRF V2 [46]	853	1	45	16	3	200 Hz	20	5	802.11n
NFS-Fi	64823	2–4	10	56	6	<i>Native Traffic</i>	40	5	802.11ac/ax

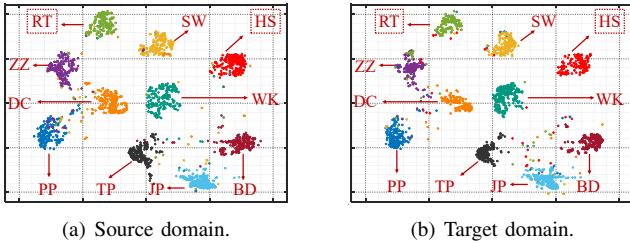


Fig. 12. t-SNE visualization. Large inter-class margin and feature alignment between the target and source domains demonstrate WiAnchor's effectiveness.

of 64 is used throughout the entire training process. During the PT stage, the learning rate is set to 1^{-3} for 50 epochs. In the FT stage, the FT dataset \mathcal{D}^{FT} contains 10 samples for each available activity, while absent categories contain 0 samples; the anchor dataset $\bar{\mathcal{D}}^{\text{PT}}$ includes all classes with 30 samples per category. The SC and FP modules use learning rates of 7^{-4} and 5^{-4} , respectively, while the other modules are frozen to prevent overfitting, and the model is fine-tuned for 200 epochs.

B. Micro-benchmark Study

To analyze the effectiveness of our WiAnchor framework, we use t-SNE to visualize the extracted features. As shown in Fig. 12(a), benefiting from the inter-class margin enlarging strategy, the features in the source domain present well-defined clustering patterns, where samples from the same category are tightly grouped and different categories are clearly separated. In Fig. 12(b), leveraging a small subset of data from source domain as anchors during the PT stage results in target domain features closely aligning with the source domain features, exhibiting a consistent distribution without noticeable shift. Thanks to these strategies and the matched filter-driven mechanism, subject-specific interference in the RT and HS categories of the target domain is effectively eliminated during the FT stage, even without samples, achieving feature separation nearly comparable to that of categories with sufficient samples in the source domain. These results demonstrate that our WiAnchor framework, proposed in Section III, is consistent with the insights discussed in Section II-C2, achieving satisfactory recognition accuracy with certain categories absent.

C. Overall Performance

To evaluate the overall performance of our WiAnchor framework, we analyze activity recognition accuracy and compare it with representative baselines. Since no existing approaches can be directly applied to this novel Wi-Fi sensing task and dataset, we adopt widely used methods from three aspects, namely class-sensitive learning, data augmentation, and module optimization, as baselines [59]. Specifically:

- **Class-sensitive Learning:** During both the PT and FT stages, the softmax loss is reweighted across categories to balance uneven gradients, while label smoothing is applied to mitigate overconfident predictions, thereby improving recognition of activities without FT samples.
 - **Data Augmentation:** A generative model is trained on cluster centers derived from abundant source domain data

during the PT stage, and subsequently generates absent-category samples from limited target domain data in the FT stage to enhance HAR performance.

- **Module Optimization:** A scale-invariant cosine classifier [60] is employed in both PT and FT stages to eliminate the effect of feature and weight scales by constraining vectors on a hypersphere. During the FT stage, \mathcal{D}^{FT} and $\hat{\mathcal{D}}^{\text{PT}}$ are jointly sampled to promote intra-class similarity and inter-class dissimilarity within each batch, thereby enhancing the feature extractor.

To ensure a fair comparison, all baselines are trained on the data processed according to Section III-A, using the basic neural network model presented in Section III-B. RT and HS, two activities inherently difficult to collect, are treated as categories without available samples during the FT stage.

We sequentially designate the data from 56 different subjects as the target domain and compute their overall HAR performance, as shown in Fig. 13. It can be observed from Fig. 13(a) that after FT with our WiAnchor framework, the overall recognition accuracy reaches approximately 90.4%. The categories with only a few FT samples achieve an average accuracy of about 91.4%, while the categories without FT samples, namely RT and HS, attain an average accuracy of approximately 86.3%. The RT and HS exhibit an improvement of about 56.8% over the approximately 29.5% accuracy shown in Fig. 4(b), demonstrating the feasibility of our WiAnchor.

In contrast, Fig. 13(b) shows that the class-sensitive learning framework yields an overall accuracy of 77.7%, while the average accuracy of RT and HS is only about 33%. Since there are no RT and HS samples from the target domain for FT, the framework can only adjust the loss of source domain data to emphasize certain categories; consequently, this strategy still fails to capture the subject-specific features of the target domain effectively. Fig. 13(c) shows that the data augmentation framework achieves an overall accuracy of

	PP	SW	DC	ZZ	BD	JP	WK	TP	RT	HS
PP	.92	.02	.02	.01	.01	.00	.00	.00	.00	.00
SW	.02	.92	.02	.01	.00	.00	.00	.00	.00	.00
DC	.02	.02	.91	.04	.01	.01	.00	.00	.00	.00
ZZ	.01	.02	.02	.91	.01	.01	.00	.00	.00	.00
BD	.01	.02	.01	.02	.90	.01	.00	.01	.01	.01
JP	.01	.01	.02	.02	.01	.89	.01	.00	.01	.01
WK	.00	.00	.00	.00	.00	.00	.89	.01	.04	.05
TP	.00	.00	.00	.00	.00	.00	.01	.97	.01	.00
RT	.00	.00	.00	.00	.00	.00	.05	.01	.85	.07
HS	.00	.00	.00	.00	.00	.00	.05	.01	.06	.87

(a) WiAnchor.

(b) Class-sensitive learning.

PP	.90	.03	.02	.02	.01	.01	.01	.00	.00	.00
SW	.02	.91	.02	.02	.00	.01	.00	.00	.01	.00
DC	.02	.02	.91	.03	.00	.00	.00	.00	.00	.00
ZZ	.02	.02	.04	.89	.00	.00	.00	.01	.01	.01
BD	.00	.00	.00	.00	.84	.07	.06	.01	.00	.00
JP	.00	.01	.00	.01	.07	.83	.06	.01	.00	.00
WK	.00	.00	.00	.00	.06	.07	.82	.02	.00	.00
TP	.00	.00	.00	.00	.01	.01	.01	.96	.00	.00
RT	.08	.07	.09	.09	.04	.04	.05	.02	.50	.02
HS	.05	.06	.04	.08	.03	.04	.05	.04	.02	.66
PP	PP	SW	DC	ZZ	BD	JP	WK	TP	RT	HS

(c) Data augmentation

PP	.90	.03	.02	.02	.01	.00	.01	.01	.01	.00
SW	.02	.91	.02	.02	.00	.00	.01	.00	.00	.00
DC	.01	.02	.92	.03	.00	.00	.00	.00	.01	.00
ZZ	.01	.02	.03	.92	.01	.01	.01	.00	.00	.00
BD	.00	.00	.00	.00	.84	.07	.06	.01	.00	.00
JP	.00	.00	.00	.00	.07	.85	.05	.01	.01	.00
WK	.00	.00	.00	.00	.07	.06	.83	.01	.00	.00
TP	.00	.00	.00	.01	.01	.01	.96	.00	.00	.00
RT	.09	.08	.12	.13	.08	.05	.06	.03	.35	.01
HS	.06	.06	.08	.09	.05	.06	.05	.06	.02	.47
PP	PP	SW	DC	ZZ	BD	JP	WK	TP	RT	HS

(d) Module optimization

Fig. 13. Overall HAR performance of (a) WiAnchor, (b) class-sensitive learning, (c) data augmentation, and (d) module optimization frameworks.

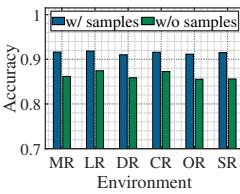
81.6%. Although RT and HS show noticeable improvement, their average accuracy remains limited to about 55%. This limitation arises from the inherent complexity and ambiguity of Wi-Fi signals, which inevitably introduce discrepancies between real and generated data, thereby restricting recognition to partially similar samples. Moreover, since the absent categories are not fixed, such methods require maintaining multiple additional generative models, which further increases the overall system complexity. In Fig. 13(d), the module optimization framework yields an overall accuracy of 79.4%, with RT and HS achieving an average accuracy of approximately 40%, which falls between the results of the previous two baselines. This is primarily because, although such methods can promote target domain feature extraction to some extent, they capture only the local sample distributions of \mathcal{D}^{FT} and $\tilde{\mathcal{D}}^{\text{PT}}$ within small batches; these limitations, compounded by the reliance on complex classifier, often lead to gradient conflicts and ultimately result in unstable optimization. These results fully demonstrate the superiority of our WiAnchor.

D. Impact Factors

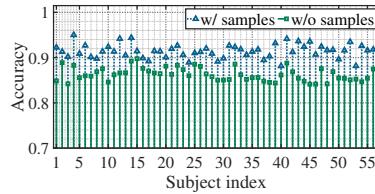
In this section, we first evaluate the potential impact factors to demonstrate the generalization capability of our WiAnchor framework. For conciseness, the metrics are defined as the average accuracies in the target domain for categories with FT samples and for those without FT samples. Finally, we conduct an ablation study to assess the contribution of each algorithm module within the framework.

1) *Environment and Subject*: To evaluate the impact of the environment on HAR performance, we analyze the average accuracy of all subjects across different environments, as shown in Fig. 14(a). The results indicate that the accuracies of categories with and without FT samples vary only slightly across environments. A closer examination shows that accuracies in the DR and OR environments are relatively lower than in other scenarios. This is primarily due to the small and crowded nature of these rooms, which severely complicates multipath propagation and increases the likelihood of interference during activity execution, thereby negatively impacting recognition performance.

To assess the influence of subjects, we analyze the recognition accuracies of all 56 subjects, as shown in Fig. 14(b). For activity categories with FT samples, the recognition accuracy for all subjects remains around 90%, with Subject 4 achieving the highest accuracy of 95.0%, while Subjects 40 and 50 record relatively lower accuracies of 88.1% and 89.6%, respectively. For activity categories without FT samples, most subjects achieve recognition accuracies around 85%, with Subject

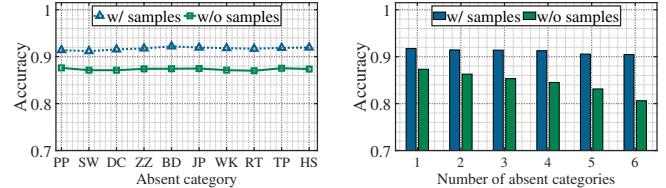


(a) Impact of environment.

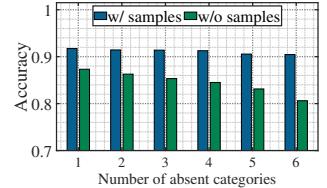


(b) Impacts of subject.

Fig. 14. Impacts of environment and subject.



(a) Absence activity category.



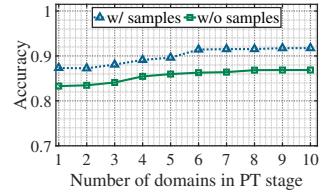
(b) Absent category number.

Fig. 15. Impact of absent activity.

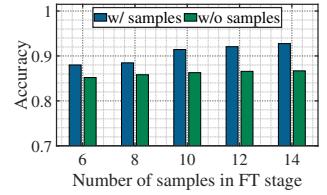
48 having the lowest accuracy of 84.2% and Subject 15 achieving the highest accuracy of 89.8%. Based on the experiment observations, this discrepancy may be attributed to differences in inter-class and intra-class similarity caused by variations in motion amplitudes. Overall, satisfactory recognition results are achieved regardless of variations in environment or subjects, demonstrating the generalization capability of our WiAnchor framework.

2) *Activity Category*: To evaluate the impact of activity category without FT samples, we analyze each activity individually, as shown in Fig. 15(a). The results indicate that the accuracy of categories with FT samples remains above 90%, while the accuracy of the absent activity improves to approximately 87%, demonstrating that our WiAnchor framework can handle scenarios with various absent categories. We further analyze the effect of the number of activity categories without FT samples, as shown in Fig. 15(b). The results reveal that the recognition accuracy of categories with FT samples fluctuates only slightly. However, as the number of absent categories increases, the recognition accuracy of these activities gradually decreases, dropping to approximately 80% when six categories are absent. This decline is primarily due to the limited available samples, which do not provide sufficient information for the anchor matching algorithm introduced in Section III-C to learn subject-specific filtering characteristics. Nevertheless, our WiAnchor framework consistently demonstrates significant performance in categories without FT samples, in comparison with Fig. 4(b), while ensuring high recognition accuracy for categories with FT samples.

3) *Training Data Size*: To evaluate the impact of the data size used in the PT stage, we analyze the recognition results under different numbers of source domains (subjects). As shown in Fig. 16(a), the recognition accuracy of all activities increases with more PT data; however, once the number of source domains reaches six, the improvement becomes negligible. This indicates that the model requires sufficient information to capture the distribution of activity features for better handling of unseen subjects. Nevertheless, once the training data diversity reaches a certain scale, simply



(a) Impact of PT data size.



(b) Impact of FT data size.

Fig. 16. Impact of training data size.

increasing the data volume no longer provides additional cross-domain recognition benefits. Therefore, we select data from six subjects as source domains for training our model.

We further evaluate the impact of the number of available FT samples per category on model performance, as shown in Fig. 16(b), where RT and HS remain absent categories. The results demonstrate that the recognition accuracy of activities with FT samples is strongly affected by their data size, but the performance gains saturate beyond 10 samples per category, indicating that the HAR model has already learned stable activity feature distributions. In contrast, the number of available FT samples does not significantly affect the accuracy of activities without samples, since even a small and diverse set drawn from multiple categories provides sufficient information for the HAR model to filter subject-specific interference. Overall, selecting 10 samples per available category for FT is adequate to achieve satisfactory performance. Moreover, compared with Fig. 4(a), where 30 FT samples per category are needed to reach saturation, our WiAnchor framework substantially reduces the sample requirement for FT.

4) Model Structure and Architecture: To evaluate the impact of the model structure, we first configure the GRU with one layer and vary the hidden size as 32, 64, and 128, denoted as Cases 1–3, and then increase the number of layers to two for these hidden sizes, denoted as Cases 4–6. As shown in Fig. 17(a), the accuracy differences are marginal, indicating the generalizability of our WiAnchor framework to different neural network structures. In addition, the recognition accuracy of activities without FT samples is slightly lower in the two cases with a hidden size of 32. Considering both performance and model simplicity, Case 2, i.e., one layer with a hidden size of 64, is selected as our configuration.

To further assess the influence of the model architecture, we replace the GRU units with a module built on 1D CNNs. We first set the number of convolutional layers to one with kernel sizes of 3 and 5, denoted as Cases 1 and 2, and then increase the number of layers to two and three, denoted as Cases 4–6. As shown in Fig. 17(b), the recognition accuracy is also insensitive to variations in the CNN structure, and the overall performance does not exhibit a significant difference compared with the GRU-based model. These results demonstrate that our WiAnchor framework can effectively adapt to diverse network architectures. It is worth emphasizing that the core of this work focuses on the design of the training framework rather than the neural network architecture, while the development of high-performance models remains an open avenue.

5) Ablation Study: To evaluate the importance of each algorithmic component in our WiAnchor framework, we perform

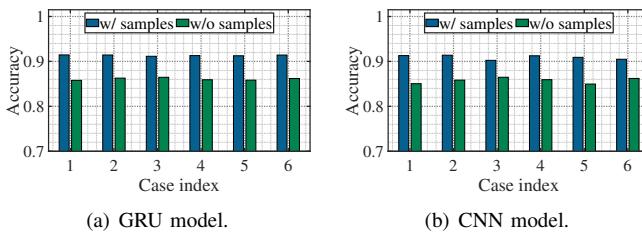


Fig. 17. Impact of model structure and architecture.

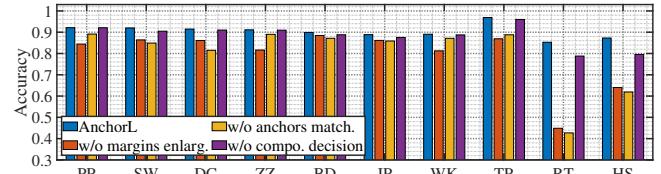


Fig. 18. Impact of algorithms in WiAnchor.

HAR analysis by removing them individually, with the results shown in Fig. 18, where RT and HS remain as categories without FT samples.

First, we remove the inter-class margin enlarging applied throughout both PT and FT stages, with the results indicated in red. The average recognition accuracy of activities with FT samples drops to 85.2%, while that of activities without FT samples decreases significantly to 54.4%. This can be intuitively explained by the t-SNE visualization of features in Fig. 5: the decision boundaries between categories are relatively blurred, which inevitably leads to performance degradation when FT samples are scarce, and makes activities with no FT samples even harder to distinguish.

Second, we remove the anchor matching algorithm used in the PT stage, with the results shown in yellow. The average recognition accuracy of activities with FT samples is 86.7%, slightly higher than the previous case but still lower than the full WiAnchor framework; for activities without FT samples, the average accuracy drops to 52.3%, even lower than the previous case. This suggests that, despite clear decision boundaries between categories, using only feature extraction without feature-matching filtering hinders accurate recognition of RT and HS, which have no FT samples.

Finally, we remove the composite inference strategy, with results indicated in purple. The average accuracy of activities with FT samples is 90.7%, while that of activities without FT samples decreases to 79.2%. Although the performance drop is less pronounced than in the first two cases, the accuracy of RT and HS falls below 80%. This highlights the benefit of the composite decision, which effectively leverages the advantages of both the PT and FT stages, yielding superior generalization compared with relying solely on the softmax function.

VI. CONCLUSIONS AND DISCUSSIONS

We have introduced a realistic Wi-Fi multi-subject HAR method based on the near-field domination effect, which requires FT to recognize unseen subjects due to its subject-specific nature. To address the challenge posed by the absence of FT samples in certain categories, we develop the WiAnchor framework. Our WiAnchor first captures temporal irregularity patterns in CSI data through time information embedding. HAR model training is then divided into two stages: during the PT stage, WiAnchor enlarges inter-class margins to improve category separability, while in the FT stage, it learns subject-specific filtering characteristics through an anchor matching mechanism. In the inference phase, a composite decision strategy is employed to further enhance recognition performance. Due to the lack of publicly available datasets, we construct a unique dataset comprising approximately 65,000

multi-subject near-field sensing samples to evaluate our WiAnchor framework. Extensive evaluation shows that WiAnchor achieves 91.4% accuracy for categories with FT samples and 86.3% for those without, while also demonstrating robust generalization to various impact factors. Although improving recognition performance, WiAnchor framework introduces potential privacy risks, as Wi-Fi APs could infer unauthorized activities. To mitigate this, we have proposed a poisoning-based approach [61] to protect user privacy and plan to explore additional efficient strategies in future work.

REFERENCES

- [1] C. Wu, X. Huang, J. Huang, and G. Xing, "Enabling Ubiquitous Wi-Fi Sensing with Beamforming Reports," in *Proc. of 37th ACM SIGCOMM*, 2023, pp. 20–32.
- [2] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 3, pp. 1629–1645, 2020.
- [3] 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.
- [4] Z. Chen, T. Zheng, C. Hu, H. Cao, Y. Yang, H. Jiang, and J. Luo, "ISACoT: Integrating Sensing with Data Traffic for Ubiquitous IoT Devices," *IEEE Communications Magazine*, vol. 61, no. 5, pp. 98–104, 2023.
- [5] 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 *Proc. of the 16th ACM MobiSys*, 2018, pp. 350–361.
- [6] Y. Xie, J. Xiong, M. Li, and K. Jamieson, "mD-Track: Leveraging Multi-Dimensionality for Passive Indoor Wi-Fi Tracking," in *Proc. of the 25th ACM MobiCom*, 2019, pp. 8:1–16.
- [7] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsoukolas, W. Xu, and L. Su, "Towards Environment Independent Device Free Human Activity Recognition," in *Proc. of the 24th ACM MobiCom*, 2018, pp. 289–304.
- [8] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Widar3.0: Zero-effort Cross-domain Gesture Recognition with WiFi," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8671–8688, 2021.
- [9] S. Huang, K. Li, D. You, Y. Chen, A. Lin, S. Liu, X. Li, and J. A. McCann, "WiMANS: A Benchmark Dataset for WiFi-based Multi-user Activity Sensing," in *Proc. of the 18th ECCV*, 2024, pp. 72–91.
- [10] M. Torun and Y. Mostofi, "Wi-Flex: Reflex Detection with Commodity WiFi," *Proc. of the ACM IMWUT*, vol. 7, no. 3, pp. 1–27, 2023.
- [11] X. Wang, C. Yang, and S. Mao, "PhaseBeat: Exploiting CSI Phase Data for Vital Sign Monitoring with Commodity WiFi Devices," in *Proc. of the 37th IEEE ICDCS*, 2017, pp. 1230–1239.
- [12] P. Hillyard, A. Luong, A. S. Abrar, N. Patwari, K. Sundar, R. Farney, J. Burch, C. Poroznik, and S. H. Pollard, "Experience: Cross-technology Radio Respiratory Monitoring Performance Study," in *Proc. of the 24th ACM MobiCom*, 2018, pp. 487–496.
- [13] 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," *Proc. of the ACM IMWUT*, vol. 3, no. 3, pp. 1–26, 2019.
- [14] Z. Jiang, T. H. Luan, X. Ren, D. Lv, H. Hao, J. Wang, K. Zhao, W. Xi, Y. Xu, and R. Li, "Eliminating the Barriers: Demystifying WiFi Baseband Design and Introducing the PicoScenes WiFi Sensing Platform," *IEEE Internet of Things Journal*, pp. 1–21, 2021.
- [15] H. Wang, X. Li, J. Li, H. Zhu, and J. Luo, "Vr-fi: Positioning and recognizing hand gestures via vr-embedded wi-fi sensing," *IEEE Transactions on Mobile Computing*, 2025.
- [16] H. Wang, D. Zhang, Y. Wang, J. Ma, Y. Wang, and S. Li, "RT-Fall: A Real-Time and Contactless Fall Detection System with Commodity WiFi Devices," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 511–526, 2016.
- [17] K. Chintalapudi, B. Radunovic, V. Balan, M. Buettner, S. Yerramalli, V. Navda, and R. Ramjee, "WiFi-NC: WiFi Over Narrow Channels," in *Proc. of the 9th USENIX NSDI*, 2012, pp. 43–56.
- [18] Y. Luo and K.-W. Chin, "An Energy Efficient Channel Bonding and Transmit Power Control Approach for WiFi Networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 8251–8263, 2021.
- [19] Zeng, Youwei and Wu, Dan and Xiong, Jie and Liu, Jinyi and Liu, Zhaopeng and Zhang, Daqing, "MultiSense: Enabling Multi-Person Respiration Sensing with Commodity WiFi," in *Proc. of the 22nd ACM UbiComp*, 2020, pp. 102:1–29.
- [20] C. R. Karanam, B. Korany, and Y. Mostofi, "Tracking from One Side: Multi-person Passive Tracking with WiFi Magnitude Measurements," in *Proc. of the 18th ACM/IEEE IPSN*, 2019, pp. 181–192.
- [21] K. Song, Q. Wang, S. Zhang, and H. Zeng, "SiWiS: Fine-grained Human Detection Using Single WiFi Device," in *Proc. of the 30th ACM MobiCom*, 2024, pp. 1439–1454.
- [22] Y. Xie, Z. Li, and M. Li, "Precise Power Delay Profiling with Commodity WiFi," in *Proc. of the 21st ACM MobiCom*, 2015, pp. 53–64.
- [23] D. Vasisht, S. Kumar, and D. Katabi, "Decimeter-Level Localization with a Single WiFi Access Point," in *Proc. of the 13th USENIX NSDI*, 2016, pp. 165–178.
- [24] X. Li, H. Wang, Z. Chen, Z. Jiang, and J. Luo, "UWB-Fi: Pushing WiFi towards Ultra-wideband for Fine-Granularity Sensing," in *Proc. of the 22nd ACM MobiSys*, 2024, pp. 42–55.
- [25] S. Tan, L. Zhang, Z. Wang, and J. Yang, "MultiTrack: Multi-user Tracking and Activity Recognition using Commodity WiFi," in *Proc. of the 37th ACM CHI*, 2019, pp. 1–12.
- [26] Y. Ren, Z. Wang, Y. Wang, S. Tan, Y. Chen, and J. Yang, "GoPose: 3D Human Pose Estimation Using WiFi," *Proc. of the 24th ACM UbiComp*, vol. 6, no. 2, pp. 1–25, 2022.
- [27] J. Hu, H. Wang, T. Zheng, J. Hu, Z. Chen, H. Jiang, and J. Luo, "Password-stealing Without Hacking: WiFi Enabled Practical Keystroke Eavesdropping," in *Proc. of the 30th ACM CCS*, 2023, pp. 239–252.
- [28] H. Wang, J. Hu, T. Zheng, J. Hu, Z. Chen, H. Jiang, Y. Zheng, and J. Luo, "MuKI-Fi: Multi-person Keystroke Inference with BFI-enabled WiFi Sensing," *IEEE Transactions on Mobile Computing*, 2024.
- [29] J. Cong, C. You, J. Li, L. Chen, B. Zheng, Y. Liu, W. Wu, Y. Gong, S. Jin, and R. Zhang, "Near-field Integrated Sensing and Communication: Opportunities and Challenges," *IEEE Wireless Communications*, 2024.
- [30] J. Hu, T. Zheng, Z. Chen, H. Wang, and J. Luo, "MUSE-Fi: Contactless MUti-person SEnsing Exploiting Near-field WiFi Channel Variation," in *Proc. of the 29th ACM MobiCom*, 2023, pp. 75:1–15.
- [31] W. Jiang, H. Xue, C. Miao, W. Shiyang, L. Sen, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3D Human Pose Construction Using WiFi," in *Proc. of the 26th ACM MobiCom*, 2020, pp. 23:1–14.
- [32] J. Hu, X. Li, Z. Su, and J. Luo, "Cross-Domain Continual Learning for Edge Intelligence in Wireless ISAC Networks," *IEEE Trans. Wireless Commun.*, 2025, early access.
- [33] J. Hu, X. Li, J. Gan, and J. Luo, "Poison to Cure: Privacy-preserving WiFi Multi-User Sensing via Data Poisoning," in *Proc. of the 31st ACM MobiCom*, 2025.
- [34] 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.
- [35] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, "FallDefi: Ubiquitous Fall Detection Using Commodity WiFi Devices," *Proc. of the ACM IMWUT*, vol. 1, no. 4, pp. 1–25, 2018.
- [36] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "SignFi: Sign Language Recognition Using WiFi," *Proc. of the ACM IMWUT*, vol. 2, no. 1, pp. 1–21, 2018.
- [37] L. Guo, L. Wang, C. Lin, J. Liu, B. Lu, J. Fang, Z. Liu, Z. Shan, J. Yang, and S. Guo, "Wiar: A Public Dataset for WiFi-based Activity Recognition," *IEEE Access*, vol. 7, pp. 154935–154945, 2019.
- [38] J. K. Brinke and N. Meratnia, "Dataset: Channel State Information for Different Activities, Participants and Days," in *Proc. of the 2nd Workshop on Data Acquisition to Analysis*, 2019, pp. 61–64.
- [39] A. Baha'A, M. M. Almazari, R. Alazrai, and M. I. Daoud, "A Dataset for WiFi-based Human Activity Recognition in Line-of-Sight and Non-Line-of-Sight Indoor Environments," *Data in Brief*, vol. 33, p. 106534, 2020.
- [40] P. Hu, C. Tang, K. Yin, and X. Zhang, "WiGR: A Practical WiFi-based Gesture Recognition System with A Lightweight Few-shot Network," *Applied Sciences*, vol. 11, no. 8, p. 3329, 2021.
- [41] M. J. Bocus, W. Li, S. Vishwakarma, R. Kou, C. Tang, K. Woodbridge, I. Craddock, R. McConville, R. Santos-Rodriguez, K. Chetty *et al.*, "OPERAnet, A Multimodal Activity Recognition Dataset Acquired from Radio Frequency and Vision-based Sensors," *Scientific Data*, vol. 9, no. 1, p. 474, 2022.
- [42] J. Yang, X. Chen, H. Zou, D. Wang, Q. Xu, and L. Xie, "EfficientFi: Toward Large-scale Lightweight WiFi Sensing via CSI Compression," *IEEE Internet of Things Journal*, vol. 9, no. 15, pp. 13086–13095, 2022.

- [43] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "MM-Fi: Multi-modal Non-intrusive 4D Human Dataset for Versatile Wireless Sensing," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 756–18 768, 2023.
- [44] Z. Zhao, T. Chen, F. Meng, H. Li, X. Li, and G. Zhu, "Finding the Missing Data: A BERT-inspired Approach Against Package Loss in Wireless Sensing," in *Proc. of the 43rd IEEE INFOCOM WKSHPS*, 2024, pp. 1–6.
- [45] F. Wang, Y. Lv, M. Zhu, H. Ding, and J. Han, "XRF55: A Radio Frequency Dataset for Human Indoor Action Analysis," *Proc. of the ACM IMWUT*, vol. 8, no. 1, pp. 1–34, 2024.
- [46] B. Lan, P. Li, J. Yin, Y. Song, G. Wang, H. Ding, J. Han, and F. Wang, "XRF V2: A Dataset for Action Summarization with Wi-Fi Signals, and IMUs in Phones, Watches, Earbuds, and Glasses," *arXiv preprint arXiv:2501.19034*, 2025.
- [47] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3D Tracking via Body Radio Reflections," in *Proc. of the 11th USENIX NSDI*, 2014, pp. 317–329.
- [48] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Simon & Schuster, Inc., 1992.
- [49] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Cambridge University Press, 2024.
- [50] R. Gao, W. Li, Y. Xie, E. Yi, L. Wang, D. Wu, and D. Zhang, "Towards Robust Gesture Recognition by Characterizing the Sensing Quality of WiFi Signals," *Proc. of the ACM IMWUT*, vol. 6, no. 1, pp. 1–26, 2022.
- [51] K. Niu, F. Zhang, X. Wang, Q. Lv, H. Luo, and D. Zhang, "Understanding WiFi Signal Frequency Features for Position-independent Gesture Sensing," *IEEE Transactions on Mobile Computing*, vol. 21, no. 11, pp. 4156–4171, 2021.
- [52] S. Liu, Z. Chen, M. Wu, C. Liu, and L. Chen, "WiSR: Wireless Domain Generalization based on Style Randomization," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 4520–4532, 2023.
- [53] D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "AirFi: Empowering WiFi-based Passive Human Gesture Recognition to Unseen Environment via Domain Generalization," *IEEE Transactions on Mobile Computing*, vol. 23, no. 2, pp. 1156–1168, 2022.
- [54] T.-T. Wong, "Performance Evaluation of Classification Algorithms by k-fold and Leave-one-out Cross Validation," *Pattern Recognition*, vol. 48, no. 9, pp. 2839–2846, 2015.
- [55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proc. of the 31st ACM NIPS*, 2017, p. 6000–6010.
- [56] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional Sequence to Sequence Learning," in *Proc. of the 34th ICML*, 2017, pp. 1243–1252.
- [57] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An Empirical Investigation of Catastrophic Forgetting in Gradient-based Neural Networks," *arXiv preprint arXiv:1312.6211*, 2013.
- [58] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep Hypersphere Embedding for Face Recognition," in *Proc. of the 30th IEEE/CVF CVPR*, 2017, pp. 212–220.
- [59] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep Long-tailed Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 795–10 816, 2023.
- [60] T. Wu, Z. Liu, Q. Huang, Y. Wang, and D. Lin, "Adversarial Robustness under Long-tailed Distribution," in *Proc. of the 34th IEEE/CVF CVPR*, 2021, pp. 8659–8668.
- [61] J. Hu, X. Li, J. Gan, and J. Luo, "Poison to Cure: Privacy-preserving Wi-Fi Multi-User Sensing via Data Poisoning," in *Proc. of the 31st ACM MobiCom*, 2025.