Boosting WiFi-based Gesture Recognition via High-fidelity and Behaviorally Diverse Data Synthesis

Jianwei Liu*
Zhejiang University
China
jianweiliu@zju.edu.cn

Jiatong Chen*
Zhejiang University
China
chenjiatong@zju.edu.cn

Rui Xiao
School of Computing and Artificial Intelligence
Shanghai University of Finance and Economics
Shanghai, China
xiaorui1998@hotmail.com

Xing Fu
Ant Group
China
zicai.fx@antgroup.com

Xin-Wei Yao†
Zhejiang University of Technology
China
xwyao@zjut.edu.cn

Jinsong Han†
Zhejiang University
China
hanjinsong@zju.edu.cn

Abstract

Artificial-intelligence–driven WiFi-based human gesture recognition (HGR) critically depends on high-quality, well-labeled real-world datasets. However, collecting real WiFi data is labor-intensive and time-consuming. Data augmentation (DA) has emerged as an effective strategy to enhance the value of limited real data and improve system generalization. Existing wireless DA methods, however, either cannot generate samples for new gesture classes or fail to capture behavioral diversity sufficiently, fundamentally limiting their scalability to new sensing tasks and augmentation effectiveness. In this work, we propose *SynthFi*, a novel DA framework capable of synthesizing high-fidelity and diverse WiFi data, even for unseen classes with very few real samples. *SynthFi* incorporates two key components: a diversity-inducing factor quantification method that extracts rich behavioral variation cues from limited data, and a diffusion-based generative model that produces realistic samples conditioned on diverse behavioral features. Extensive real-world experiments show that *SynthFi* can improve recognition accuracy by 40% through DA, even when only two real samples of unseen classes are available, outperforming existing solutions.

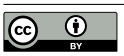
CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**.

Keywords

Gesture recognition, Human-computer interaction, WiFi sensing, Data augmentation, Diversity

*Jianwei Liu and Jiatong Chen are co-first authors.
†Jinsong Han and Xin-Wei Yao are corresponding authors.



UIST '26, Detroit, MI, USA

ACM ISBN 979-8-4007-2856-3/2026/11
https://doi.org/10.1145/3830398.3830548

ACM Reference Format:
Jianwei Liu, Jiatong Chen, Rui Xiao, Xing Fu, Xin-Wei Yao, and Jinsong Han. 2026. Boosting WiFi-based Gesture Recognition via High-fidelity and Behaviorally Diverse Data Synthesis. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 02–05, 2026, Detroit, MI, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3830398.3830548

1 Introduction

Human gesture recognition (HGR) serves as a core enabler for a broad spectrum of human–computer interaction (HCI) applications and gesture-based interfaces, such as virtual/augmented reality, smart home, and assistive systems [12, 38, 49, 68]. In recent years, WiFi has been increasingly adopted for arm-levek gesture recognition due to its advantages of privacy preservation and device-free interaction. More importantly, WiFi-based HGR can be realized via ubiquitous communication infrastructures, namely integrated sensing and communication (ISAC) [6, 18, 20, 21, 27–29]. To ensure high recognition accuracy, existing WiFi-based HGR systems typically rely on machine/deep learning models to map Channel State Information (CSI) to gesture classes. However, their performance heavily depends on large, well-annotated datasets, the collection and labeling of which are both costly and labor-intensive. A promising way to reduce this burden is data augmentation (DA), where novel samples are synthesized to enrich the training set [5, 14, 34, 55, 59].

Existing DA approaches for wireless sensing can be broadly grouped into two categories: sample transformation [39, 50, 65] and learning-based synthesis [3, 5, 59]. The former methods create new samples by adding noise to raw samples [50] or combining multiple samples [39]. However, they often struggle to satisfy the two essential requirements for synthesized data, namely high *fidelity* and *diversity* [11]. For example, OneSense [65] produces new samples by directly concatenating two real samples in the temporal domain, a process that disregards the physical propagation characteristics of wireless signals and therefore yields unrealistic results with low fidelity. Meanwhile, transformations applied to existing samples inherently underperform in expanding the underlying data distribution, which limits the diversity. In contrast, deep generative models such as Generative Adversarial Networks (GANs) [59]

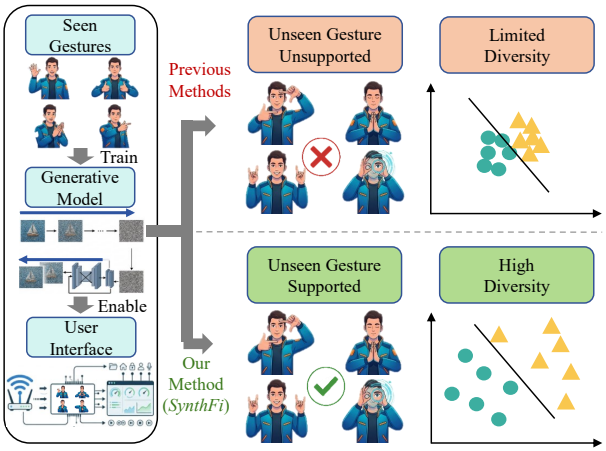


Figure 1: Comparison between prior methods of wireless sensing data synthesis and *SynthFi*.

and diffusion models [5] have demonstrated strong capabilities in producing high-fidelity data by automatically learning complex propagation patterns of wireless signals.

Nevertheless, as shown in Fig.1, current learning-based methods suffer from at least one of the following drawbacks:

(1) Restricted scalability. Previous data amplification methods [3, 5] typically rely on predefined categorical labels or auxiliary cues from other modalities (e.g., video and text [3]) to guide the generation process. This dependency prevents them from producing samples for unseen gesture classes. For example, RF-Diffusion [5] requires an activity label as a conditioning input, which helps avoid distribution shifts toward irrelevant classes but simultaneously limit generalization to new gesture categories. Consequently, these methods cannot generate samples for classes that lack sufficient real data to train the generative model. Given that gesture requirements often vary across users and applications, and collecting data for customized-yet-unseen classes is costly, the scalability of existing approaches remains inadequate for practical deployment.

(2) Limited diversity. Human motions naturally exhibit variations across repetitions due to inherent biomechanical and behavioral differences, which lead to substantial diversity in real gesture samples. Capturing this variability typically requires a large training set to expose the HGR model to rich distributions. However, existing wireless synthesis approaches [5, 55, 59] such as CsiGAN [59] primarily focus on coarse-grained cross-domain (e.g., cross-user/environment) attributes while overlooking fine-grained in-domain behavioral cues during synthesis. As a result, the generated samples lack sufficient diversity, a limitation that becomes particularly pronounced in HCI scenarios and significantly impairs the effectiveness of DA.

In this paper, we pose a fundamental question: *Is it possible to synthesize WiFi data with sufficient behavioral diversity to improve the DA effectiveness, even for unseen classes?* As a response to this challenge, we introduce *SynthFi*, a novel DA framework for WiFi-based HGR that generates high-fidelity and behaviorally diverse WiFi data using only a limited amount of real data, as shown in Fig. 1. With this capability, *SynthFi* enables the rapid development, evaluation, and deployment of data-efficient and accurate HCI systems. As a building block for gesture interfaces, *SynthFi* supports: i) rapid prototyping and expansion of gesture vocabularies by enabling low-cost training data generation for new gestures; ii) personalized gesture interfaces by adapting to user-specific interaction preferences with minimal data collection; and iii) scalable real-world WiFi interfaces by reducing the data collection, deployment, and maintenance overhead.

The foundation of *SynthFi* stems from a key observation: human gestures are inherently non-repeatable [45], and this non-repeatability triggers a cascade of behavioral variations that propagate into the WiFi signals. These variations manifest as differences among samples belonging to the same gesture class, referred to as *intra-class diversity*. Intuitively, factors such as how fast the motion unfolds over time and how the human limb is positioned relative to the transceivers play essential roles in shaping this diversity. To verify this intuition, we construct a signal propagation model tailored to the HGR scenario and conduct preliminary experiments. Both theoretical analysis and experimental results confirm that these physical factors (i.e., motion *speed* and *position*) are major contributors to intra-class diversity. This insight motivates us to explicitly incorporate them into the data synthesis process, enabling *SynthFi* to generate WiFi samples with substantially richer behavioral diversity, even for unseen classes.

To empower *SynthFi* toward fulfilling the aforementioned objectives, we confront the following challenges:

(1) How to quantify the diversity-inducing factors while ensuring consistency with signal propagation physics? Our theoretical and experimental analysis shows that motion speed and position are key contributors to intra-class diversity, referred to as diversity-inducing factors (DIFs). Incorporating various DIFs into data synthesis is essential for improving the diversity of generated samples. However, assigning arbitrary scalar values for DIFs cannot faithfully reflect real gesture execution, where these factors vary over time. The limited spatial resolution of WiFi signals further complicates accurate speed and position measurement. To address this challenge, we estimate the *relative shifts* in speed and position between gesture instances rather than their absolute values. By characterizing how these factors differ across instances, we obtain abundant and physically plausible diversity-inducing cues. This approach provides a reliable basis for downstream data generation.

(2) How to synthesize sufficient and diverse data from extremely limited data of unseen classes? Existing learning-based DA methods rely on many instances per gesture class to learn the complex signal distribution. In practice, however, a target class may have only a few samples. Under such scarcity, conventional generative models face two key issues: they cannot generate data for unseen classes outside the predefined label space, and with limited samples they fail to capture the true distribution, often producing unrealistic or physically inconsistent results. Moreover, traditional methods ignore physical DIFs, limiting sample diversity and weakening augmentation effectiveness. To address these challenges, we propose a novel physically guided data synthesis approach. Unlike prior methods that reconstruct data from random noise, our approach learns a transformation from one instance to another conditioned on the DIF shift between them. Leveraging abundant DIF shifts extracted from seen-class samples, *SynthFi* can generate massive and diverse samples for unseen classes using a few available samples as the transformation starting points.

We prototype *SynthFi* using commercial off-the-shelf (COTS) WiFi devices and conduct extensive experiments involving ten participants. The results show that *SynthFi* can synthesize WiFi data with both high fidelity and high diversity. It boosts recognition accuracy by 40% through dataset augmentation even when each unseen class provides two real samples, outperforming existing DA frameworks such as WixUp [39] and RF-Diffusion [5]. In addition, *SynthFi* can also be extended to broader HCI sensing tasks such as activity recognition and user identification.

In summary, our contributions are as follows: (1) We present *SynthFi*, a novel DA framework for WiFi-based HGR that reduces data collection overhead while improving recognition performance. The source code for DIF shift quantification and the generative model has been released[1]. (2) We introduce a physics-informed data synthesis paradigm that generates high-fidelity and high-diversity WiFi data conditioned on human behavioral cues. This paradigm holds great promise for improving the generalization capability of data-driven wireless HCI. (3) We build a prototype of *SynthFi* and conduct extensive real-world experiments. The results indicate that *SynthFi* can effectively improve the sensing performance by augmenting the dataset. Moreover, our data synthesis method can be easily extended to other HCI applications.

2 Related Work

2.1 Gesture Recognition

HGR plays a vital role in many HCI applications. Conventional gesture recognition systems typically rely on cameras [10, 35, 56], sonars [31, 42, 62], or wearable devices [1, 58, 66] to capture gesture information. However, each of these modalities has inherent limitations. Cameras raise concerns about visual privacy, sonars operate within restricted sensing ranges, and wearable devices are often inconvenient for continuous or long-term use. By comparison, WiFi-based sensing [9, 16, 17, 19, 22, 24, 25, 36, 37, 51, 52, 60] offers several benefits because it preserves user privacy, provides a wide sensing range, and remains unobtrusive in daily environments. The main obstacle, however, is that current WiFi-based systems depend heavily on machine/deep learning but lack large-scale labeled datasets required for training.

To address this issue, two research directions have received increasing attention: few-shot learning (FSL) [60, 65] and DA [5, 50]. FSL aims to generalize from a very limited number of samples through meta-learning or transfer learning, whereas DA generates synthetic samples to enrich the diversity of the training set. Despite their effectiveness, FSL methods often depend on large backbone networks to extract discriminative features from scarce examples. For instance, OneSense [40] uses deep convolutional architectures and OneFi [60] adopts Transformer models. These designs introduce high computational and memory demands, which hinder deployment on resource-constrained edge devices. In contrast, generative DA does not impose such architectural requirements and therefore supports flexible deployment across diverse hardware platforms. Motivated by this advantage, we propose a new DA framework, named *SynthFi*, for WiFi-based HGR. Unlike prior approaches, *SynthFi* novelly incorporates physical behavioral conditions into the synthesis process, substantially enhancing the diversity of the generated data.

[1] https://github.com/Percilot/SynthFi

Table 1: Comparison with the most relevant methods.

Framework	Within Modality	High Fidelity	Behavioral Diversity	Unseen Classes
RF Genesis [3]	✗	✓	✗	✓
RF-Diffusion [5]	✓	✓	✗	✗
SynthFi (Ours)	✓	✓	✓	✓

2.2 Data Synthesis for Wireless Sensing

Data synthesis has been widely adopted in wireless sensing to enable domain transfer [2, 54, 55, 59, 67] and dataset enrichment [3, 5, 34, 39, 50]. Current wireless data synthesis techniques fall broadly into two categories: physics-informed generation and data-driven generation. Physics-informed approaches exploit the underlying physical properties of wireless signals and emulate propagation behavior to create new samples through in-place transformations or ray-tracing-based simulation. For example, CSI-Net [50] augments data by injecting carefully designed perturbations into original CSI measurements. WixUp [39] constructs a custom Gaussian mixture model together with a probability-driven transformation strategy to produce synthetic samples by blending multiple real instances. mmFace [48] reconstructs signal propagation paths using physical modeling, thereby generating virtual measurements based on modeled reflections. Although these methods offer strong interpretability, they struggle to achieve high realism because multipath-shaped signal trajectories [46] are exceedingly difficult to model with adequate accuracy.

In contrast, data-driven approaches leverage deep neural networks to learn the complex and nonlinear relationships embedded in wireless signals, enabling the synthesis of data that better aligns with real-world characteristics. CsiGAN [59], for instance, employs GANs to learn user-specific signal attributes and achieves cross-user transfer. RF Genesis [3] adopts a diffusion-based generative framework that incorporates auxiliary information such as video cues to guide virtual sample generation. RF-Diffusion [5] also uses iterative denoising to learn the distribution of wireless measurements and produce synthetic CSI samples. Despite their effectiveness, existing data-driven methods either cannot synthesize data for unseen classes or fail to capture sufficient intra-class diversity, as shown in Tab. 1. Although RF Genesis [3] enables unseen-gesture data synthesis for mmWave using auxiliary modalities, its design may not directly transfer to WiFi due to the substantially different propagation characteristics. In contrast, *SynthFi explicitly incorporates quantified in-domain behavioral factor shifts into the generation process, enabling highly realistic and diverse synthesis for unseen classes without requiring additional modalities.* We believe that this framework opens a promising direction for enriching data augmentation across a wider range of wireless sensing modalities.

3 Motivation

Intuitively, the motion speed and the relative position of human limbs with respect to the transceivers vary across gesture repetitions. These factors may play a crucial role in shaping the intra-class

diversity of WiFi sensing data. If so, incorporating these DIFs into data synthesis is essential for expanding diversity of generated samples. In this section, we first introduce the basics of WiFi-based HGR. We then develop a theoretical model of human reflection and conduct preliminary experiments to validate the necessity of incorporating diverse DIFs into data synthesis.

3.1 Basics of WiFi-based HGR

To implement a WiFi-based HGR system, users first need to define a set of gesture classes to be recognized. Subsequently, multiple signal transceivers are deployed to collect WiFi signals, and a set of samples is gathered for each class to construct a dataset. During data acquisition, the transmitted signals are reflected by the human body, thereby capturing human dynamics, and are then received by the receiver. The receiver extracts CSI [53] from the received signals to characterize environmental changes, including human motions. After a series of preprocessing steps, the processed data are used to train a machine/deep learning-based gesture recognition model. Once well-trained, the model can be deployed to infer the gesture class of newly received samples in real time.

It is evident that data collection incurs the highest cost in this pipeline. Moreover, the collected data often fail to adequately cover the underlying distribution, leading to limited generalization performance. Additionally, when new gesture classes (i.e., unseen classes) beyond the existing class set are required to recognize, additional data collection costs arise. With the growing demand for customized gesture classes across users and applications, these limitations become more pronounced. To address these issues, existing studies [5, 39] prefer to employ DA techniques to improve performance without incurring additional data collection overhead.

3.2 How DIF Shapes WiFi Signals?

To theoretically examine whether speed and position indeed contribute to intra-class diversity in HCI scenarios, we develop a propagation model comprising a transmitter (Tx) and a receiver (Rx) located at positions $\mathbf{P}_T, \mathbf{P}_R \in \mathbb{R}^3$, as shown in Fig. 2. Without loss of generality, we focus on a single representative point on the moving limb, denoted $\mathbf{P}_b(t) \in \mathbb{R}^3$, which acts as a time-varying scatterer. In practice, the received signal is the superposition of reflections from many such scatterers.

Under the first-order Born approximation [4], the received complex signal can be expressed as $s(t) = s_{\text{LOS}}(t) + s_{\text{scat}}(t)$, where $s_{\text{LOS}}(t)$ is the line-of-sight (LOS) component between Tx and Rx, and $s_{\text{scat}}(t)$ is the single-bounce scattered contribution from $\mathbf{P}_b(t)$. The LOS term can be modeled as:

$$s_{\text{LOS}}(t) = C_0\, \mathbf{g}(\mathbf{P}_T, \mathbf{P}_R), \tag{1}$$

where C_0 incorporates the transmit power and antenna gains. The scattered term can be expressed as:

$$s_{\text{scat}}(t) = A_b\, G_b(t)\, \mathbf{g}(\mathbf{P}_T, \mathbf{P}_b(t))\, \mathbf{g}(\mathbf{P}_b(t), \mathbf{P}_R), \tag{2}$$

where A_b is an amplitude coefficient (we set $A_b = 1$ for a point), and $G_b(t)$ is a directional gain factor accounting for the orientation-dependent reflectivity. The function $\mathbf{g}(\mathbf{X}, \mathbf{Y})$ is the free-space Green's function [32] $\mathbf{g}(\mathbf{X}, \mathbf{Y}) = \frac{\exp\left(j\frac{2\pi}{\lambda}\|\mathbf{X}-\mathbf{Y}\|\right)}{4\pi\|\mathbf{X}-\mathbf{Y}\|}$, where λ is the carrier wavelength.

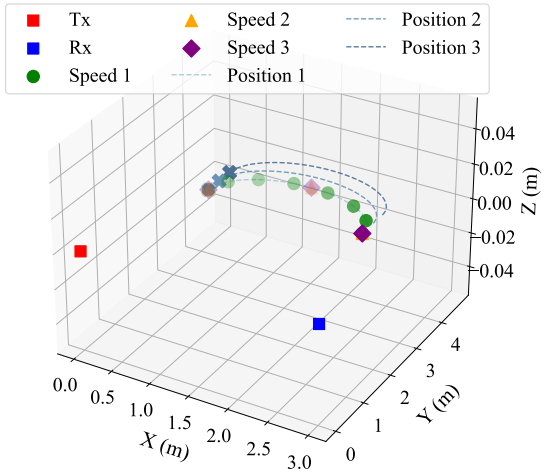


Figure 2: Sensing scenario with one Tx and one Rx.

As human body is best modeled as a quasi-specular reflector that reflects the signal into many directions with different amplitudes [61], we have $\mathbf{u} = \mathbf{P}_b(t) - \mathbf{P}_T$ and $\mathbf{r}_b(t) = \frac{\mathbf{u} - 2(\mathbf{u}^\top \mathbf{n}_b)\mathbf{n}_b}{\|\mathbf{u}\|}$, where $\mathbf{n}_b$ is the local surface normal at $\mathbf{P}_b(t)$. The strongest component is reflected into the $\mathbf{r}_b(t)$ direction. Moreover, according to [32], the directional mask $G_b(t)$ attenuates reflected amplitude away from $\mathbf{r}_b(t)$ via a Gaussian function of the angular deviation:

$$G_b(t) = \exp\left(-\frac{\left(\arccos\frac{(\mathbf{P}_R - \mathbf{P}_b(t))^\top \mathbf{r}_b(t)}{\|\mathbf{P}_R - \mathbf{P}_b(t)\|\,\|\mathbf{r}_b(t)\|}\right)^2}{2\epsilon^2}\right), \tag{3}$$

where ϵ controls the lobe width.

Next, we incorporate motion into the model to analyze the resulting dynamic reflections. For analytical tractability, we model the motion of $\mathbf{P}_b(t)$ as a semicircular trajectory in a horizontal plane (as shown in Fig. 2), representing a simplified arm-swing gesture:

$$\mathbf{P}_b(t) = \mathbf{C} + R\begin{bmatrix}\cos\theta(t)\\ \sin\theta(t)\\ 0\end{bmatrix}, \quad \theta(t) = \theta_0 + \omega t, \tag{4}$$

where $\mathbf{C}$ is the trajectory center, R is the radius, θ_0 is the initial angle, and $\omega = v/R$ is the angular speed determined by the speed v.

By substituting $\mathbf{P}_b(t)$ into Eq. (2), we can derive an explicit dependency of the scattered signal strength on the motion speed v and the initial position $\mathbf{P}_b(0)$:

$$\begin{aligned}|s_{\text{scat}}(t; v, \mathbf{P}_b(0))| = A_b\, G_b(t; v, \mathbf{P}_b(0))\, &\frac{1}{\|\mathbf{P}_T - \mathbf{P}_b(t; v, \mathbf{P}_b(0))\|}\\ \times &\frac{1}{\|\mathbf{P}_b(t; v, \mathbf{P}_b(0)) - \mathbf{P}_R\|},\end{aligned} \tag{5}$$

where the inverse-distance factors come from the amplitude term of the free-space Green's function, and

$$\mathbf{P}_b(t; v, \mathbf{P}_b(0)) = \mathbf{C} + R\begin{bmatrix}\cos\left(\theta_0(\mathbf{P}_b(0)) + \frac{v}{R}t\right)\\ \sin\left(\theta_0(\mathbf{P}_b(0)) + \frac{v}{R}t\right)\\ 0\end{bmatrix}, \tag{6}$$

with $\theta_0(\mathbf{P}_b(0))$ denoting the initial angular coordinate determined by $\mathbf{P}_b(0)$ on the trajectory.

Combining Eq. (5) and (6), we can observe that:

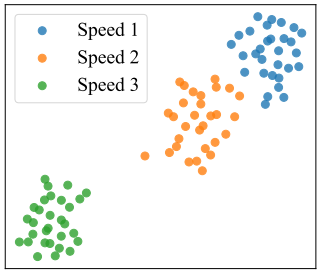


Figure 3: Signal distributions under three motion speeds.

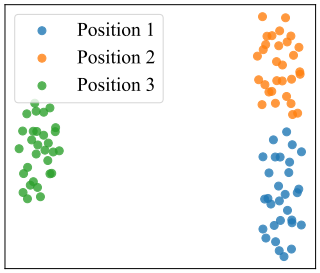


Figure 4: Signal distributions under three motion positions.

- the speed v controls the temporal rate at which the Tx–point and point–Rx distances change, thereby affecting the variation trend of received signal;
- the position $\mathbf{P}_b(0)$ sets the initial Tx–point and point–Rx distances, as well as the starting value of $G_b(t)$, thereby determining the temporal offset of the signal profile.

By varying v with $\mathbf{P}_b(0)$ fixed or vice versa, we can isolate the respective effects of motion speed and position on the signal dynamics. The model predicts that both factors alter the temporal modulation patterns of $s(t)$, thereby shifting the statistical distribution of it. *Therefore, even for a fixed gesture class, differences in speed or position naturally and significantly give rise to intra-class diversity in the measured WiFi signals.*

3.3 Preliminary Experiment

To intuitively investigate how speed and position influence signal distributions, we conduct two types of controlled experiments, each varying a single factor.

In the first experiment, we fix the initial position $\mathbf{P}_b(0)$ and vary the speed v across three levels (slow, medium, and fast), as illustrated in Fig. 2. For each configuration, we generate noisy realizations of $s(t)$ by injecting complex Gaussian noise to emulate path perturbations, collecting 30 instances per speed–position pair. We then apply Uniform Manifold Approximation and Projection (UMAP) [41] for dimensionality reduction and visualization. As shown in Fig. 3, instances with different values of v (under the same initial position) form clearly separated clusters, demonstrating speed-induced distributional diversity. To further demonstrate the importance of incorporating speed diversity in sensing model training, we introduce another trajectory in which the reflection point moves along the tangent direction of the semicircle. For each trajectory, we collect 100 instances under three speed levels. A support vector machine [23] is then trained to perform two-class recognition. The classifier achieves only 64% accuracy when trained solely on 100 slow-speed instances and evaluated on a test set composed of 100 randomly selected instances that span all three speeds. Adding 50 additional slow-speed instances to the training set increases the accuracy only to 67%, and further enlarging the slow-speed training set provides no additional gain. In contrast, when the additional 50 instances are randomly drawn from all three speeds, the accuracy rises to 80%. This comparison indicates that enriching the training set with diverse speeds is essential for improving gesture recognition performance. In the second experiment, we fix v and vary $\mathbf{P}_b(0)$ across three distinct positions (near, medium, and far), as illustrated in Fig. 2. Using the same procedure as above, we obtain the clustering results shown in Fig. 4. Evidently, instances with different values of $\mathbf{P}_b(0)$ (under the same speed) again form separate clusters, revealing position-induced diversity. Moreover, we conduct an additional classification experiment following a setting similar to that of the first experiment. The results reveal the same phenomenon as that observed under distinct speeds. We also conduct two real-world experiments using a LeKiwi robotic arm [47]. The robotic arm is programmed to execute gestures with varied yet controlled speeds and positions. The results are consistent with the simulation observations: data collected at different speeds or positions form distinct clusters.

These experiments further confirm that *motion speed and position are crucial DIFs that shape signal distributions. Therefore, it is essential to incorporate them into the data synthesis process to expand data diversity and improve HGR model's generalization performance.*

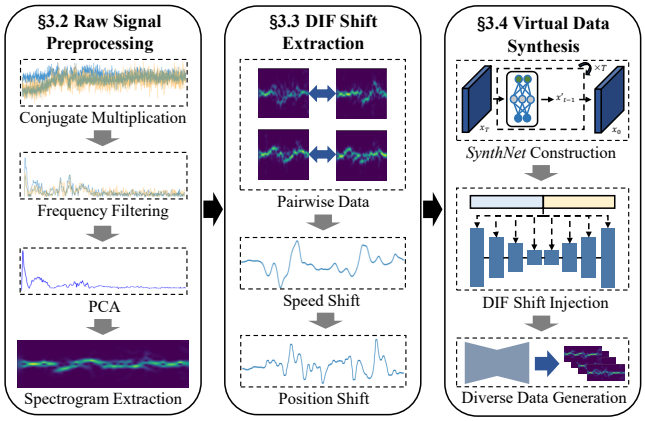


Figure 5: Overview of *SynthFi*.

4 *SynthFi* Design

In this section, we first outline the overview of *SynthFi*, followed by a detailed presentation of its technical components.

4.1 Overview

As illustrated in Fig. 5, the *SynthFi* framework consists of three primary modules: raw signal preprocessing, DIF shift extraction, and virtual data synthesis. To enable accurate and efficient gesture recognition, these modules operate across two phases, namely bootstrapping and customization.

In the bootstrapping phase, the developer or user defines a set of base gesture classes and collects corresponding WiFi data for each class. The raw signal preprocessing module converts the collected CSI into clean spectrograms. For any two spectrograms within the same class, *SynthFi* computes both speed-shift and position-shift embeddings through the DIF shift extraction module, capturing physically meaningful behavioral variations. These embeddings, together with the spectrograms, are then used to train the virtual data generation model, referred to as *SynthNet*.

During the customization phase, the user only needs to collect a few real samples for each customized-yet-unseen gesture class. After identical preprocessing, the virtual data synthesis module employs the pretrained *SynthNet*, conditioned on the DIF shift embeddings extracted from base-gesture instances, to generate a large, diverse set of synthetic spectrograms for the unseen gestures.

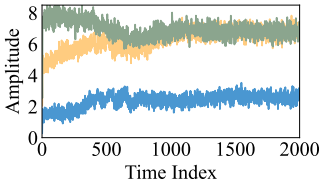


Figure 6: Three subcarriers of raw signals.

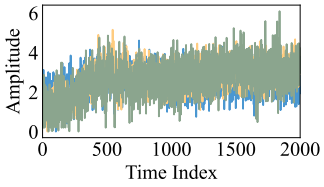


Figure 7: Result of conjugate multiplication.

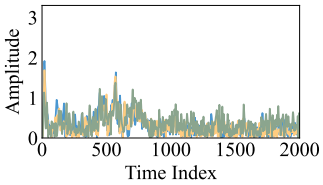


Figure 8: Result of signal filtering.

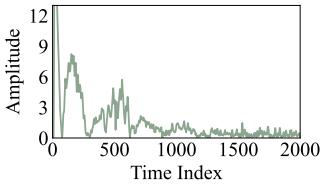


Figure 9: Result of PCA.

The user can then use both real and synthesized data to train a gesture recognition model with high generalization performance.

4.2 Signal Preprocessing

As shown in Fig. 6, raw CSI measurements typically contain substantial noise that is unrelated to behavioral information. To mitigate this issue, we apply a series of preprocessing techniques, including conjugate multiplication, filtering, and Principal Component Analysis (PCA), to obtain cleaner CSI signals. Moreover, WiFi signals are reflected not only by human bodies but also by static objects in the environment. These static reflections can significantly degrade recognition performance when the environment changes, thereby limiting cross-environment performance. To address this issue, we extract Doppler frequency–shift features from the CSI, which capture only human-induced dynamics and are subsequently represented as spectrograms.

Conjugate multiplication. A slight synchronization mismatch between the Tx and Rx may result in a time-dependent random phase perturbation, which can blur the underlying gesture-related patterns in CSI and degrade recognition performance. Fortunately, because all antennas on a given receiver rely on a common Radio Frequency (RF) oscillator, their phase deviations remain essentially identical. Leveraging this property, the undesired phase offset can be effectively canceled by applying conjugate multiplication to the CSI from two antennas of the same receiver, as depicted in Fig.7.

Signal filtering. Beyond the phase distortion, the captured CSI is further contaminated by stationary signal components and additive white Gaussian noise. To mitigate these effects, we begin by applying a high-pass filter with a 2 Hz cutoff, which suppresses the low-frequency contributions introduced by static propagation paths. Subsequently, a low-pass filter with a 60 Hz cutoff is employed to attenuate the high-frequency noise. As depicted in Fig. 8, the resulting CSI waveforms exhibit significantly smoother profiles.

PCA. Following the filtering stage, PCA is applied to the refined CSI, with the dominant principal component retained for further processing. This operation not only accentuates gesture-relevant characteristics but also suppresses residual noise. As depicted in Fig. 9, the CSI series after PCA appear considerably denoised.

Spectrogram extraction. In a gesture recognition scenario, the received signals are both from static objects and dynamic human body.

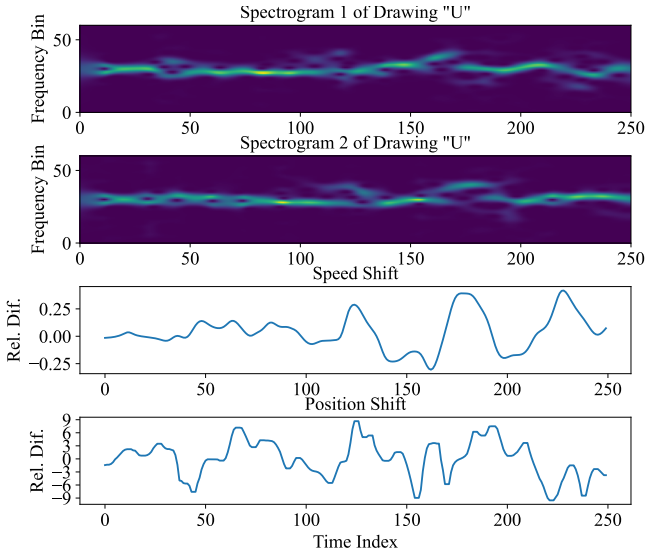


Figure 10: Two WiFi spectrograms and their DIF shift embeddings. "Rel. Dif." means relative difference.

The static-reflection could impair the cross-environment robustness of the system. To achieve stable performance across environments, we aim at extracting information only related to human motions. Doppler Frequency Shift (DFS) can effectively reflect such human dynamics. According to [60], we can represent DFS information by performing Short-Time Fourier Transform (STFT). After that, we can get spectrograms that are only related to gesture information, as shown in Fig. 10.

4.3 DIF Shift Extraction

As aforementioned, speed and position are two critical DIFs that should be incorporated into virtual data generation. However, directly and precisely measuring these factors in real-world sensing is extremely challenging. To address this, *SynthFi* instead extracts the DIF differences between pairs of instances, which offers two key advantages: (1) The extracted differences preserve the dynamic physical variations of the DIFs between instances, effectively capturing how speed and position change in practice. (2) This approach substantially enhances the utility of limited training samples. For instance, if the developer collects 20 samples for a base gesture class, only 20 distinct speed/position patterns can be obtained. In contrast, by computing pairwise DIF shifts, we can derive $2 \times C_{20}^2 = 380$ distinct difference patterns, expanding the set of available behavioral cues by twentyfold. With this enriched DIF information, the subsequent generative model can more effectively learn how to guide the virtual data synthesis process under varying behavioral conditions. In the following, we detail how to extract DIF shifts from one spectrogram instance to another.

4.3.1 Speed Shift Quantification. Recall that we extract Doppler spectrograms from the CSI matrix as cross-environment gesture features. Although the spectrogram intrinsically encodes velocity information, directly subtracting one spectrogram from another would merely capture instantaneous amplitude discrepancies, rather than reflecting the temporal evolution of the underlying velocity dynamics embedded in the Doppler spectrum. To deal with this issue, we

propose a centroid–correlation fusion method that estimates the time-resolved speed shift between two spectrograms. This method effectively preserves the gesture dynamics while maintaining consistency with Doppler physics.

Specifically, given two magnitude spectrograms $\mathbf{S}_1(t, f_k)$ and $\mathbf{S}_2(t, f_k)$ with time index t and discrete frequency bins f_k, we first normalize each time frame to suppress amplitude biases caused by signal attenuation or antenna gain variations. For each normalized frame, we compute the spectral centroid [43] as:

$$f_c^{(i)}(t) = \frac{\sum_k f_k \, \tilde{\mathbf{S}}_i(t, f_k)}{\sum_k \tilde{\mathbf{S}}_i(t, f_k)}, \tag{7}$$

which represents the first-order spectral moment of Doppler energy.

The centroid difference between the two spectrograms is calculated as $\Delta f(t) = f_c^{(2)}(t) - f_c^{(1)}(t)$, capturing the shift of Doppler energy over time. Because Doppler frequency is linearly proportional to radial velocity, this frequency shift has a direct velocity-domain interpretation:

$$\Delta v(t) = \frac{\lambda}{2} \Delta f(t), \tag{8}$$

where λ denotes the carrier wavelength. Thus, $\Delta v(t)$ can be directly regarded as the speed shift between the two motion instances, reflecting how their movement speeds differ over time, as shown in the third part of Fig. 10. Since the spectral centroid captures the global distribution of Doppler energy rather than individual frequency components, it is also inherently robust to multipath-induced perturbations.

4.3.2 Position Shift Quantification. Beyond motion speed, arm position constitutes another dominant DIF that induces substantial variability across gesture instances. Compared with spectrogram-derived features, time-series amplitudes provide a more direct and higher-resolution reflection of range-dependent fluctuations. Building on this, we introduce a time-resolved method for estimating position shift using temporal amplitude traces.

Specifically, we first approximate the instantaneous received power through short-time frame energy. Let each time-series signal be divided into overlapping frames indexed by t, with the corresponding sample set $\mathcal{W}_t$. For the i-th signal instance, the frame energy is calculated as:

$$E_i(t) = \sum_{n \in \mathcal{W}_t} x_i[n]^2, \tag{9}$$

which serves as a surrogate for short-term received power. Normalizing by the frame length L yields $\hat{P}_i(t) = \frac{E_i(t)}{L}$, where a small constant $\varepsilon > 0$ is added in subsequent steps to avoid numerical degeneracy. As received power decreases monotonically with propagation distance, $\hat{P}_i(t)$ naturally encodes range-dependent variations induced by limb movement.

To further convert short-time power into a position estimate, we employ the classical power-law path-loss model [26]. Inverting the model provides the per-frame range approximation $\hat{d}_i(t) = d_0 \left(\frac{P_0}{\hat{P}_i(t) + \varepsilon} \right)^{1/\gamma}$, where γ denotes the path-loss exponent, and P_0 and d_0 represent reference power and distance, respectively. Since our objective is to quantify *relative* positional differences rather than to recover absolute metric distances, the exact values of (γ, P_0, d_0) are not required. We therefore set them to constant nominal values

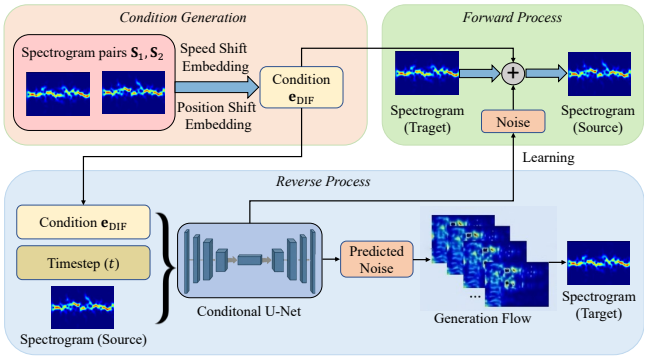


Figure 11: Architecture of our generative model *SynthNet*.

(i.e., $\gamma = 2$, $P_0 = 1$, $d_0 = 1$). which preserves the monotonic mapping from power to distance.

Given the inferred per-frame distances, the instantaneous position shift between two instances is computed as

$$\Delta p(t) = \hat{d}_2(t) - \hat{d}_1(t). \tag{10}$$

This produces a temporally aligned trajectory that describes how the relative limb position evolves across frames. Because amplitude-based distance estimates can fluctuate under multipath and noise, we finally apply a temporal smoothing. Let $h(\tau)$ denote a smoothing kernel; the refined trajectory can be expressed as:

$$\bar{\Delta p}(t) = \sum_{\tau} h(\tau) \, \Delta p(t - \tau). \tag{11}$$

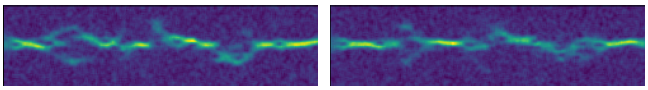
In practice, we adopt median filtering [30], which effectively suppresses impulsive outliers while enforcing the temporal continuity characteristic of human dynamics.

Through this procedure, we obtain a time-resolved estimate of position shift from amplitude traces. The resulting $\Delta p(t)$ (as shown in the last part of Fig. 10) captures how the relative limb configurations differ between gesture instances and thus forms a second key DIF complementary to the speed-induced variability.

4.4 Behaviorally Diverse Data Generation

Data synthesis is the core component of DA. Its effectiveness hinges on the realism of the generated samples. To achieve this, we prefer to use deep learning–based generative models to construct a principled synthesis mechanism. RF-Diffusion [5], a recently proposed wireless data generation framework, has demonstrated the superiority of diffusion models in producing high-fidelity WiFi signals. However, existing diffusion-based approaches face two key limitations that restrict their scalability: (1) Scalability to unseen classes. They rely on class labels as generation conditions, which inherently confines the model to synthesizing samples only for classes observed during training, preventing generalization to unseen gestures. (2) Scalability to unseen distributions. Since the model learns only the empirical distribution of the training set, the diversity of synthesized data is fundamentally limited, diminishing its augmentation effectiveness.

In this paper, we propose a novel diffusion-based generative paradigm that overcomes these limitations, as shown in Fig. 11. Unlike prior works [5], we discard class labels as generation conditions. Instead, we use DIF shifts to guide the synthesis process.



(a) Instance 1. (b) Instance 2.

Figure 12: Synthesized instances of "swift left & right".

Under this formulation, data generation is no longer performed by denoising a random noise matrix. Rather, the model learns to transform one real instance into another conditioned on their DIF differences. This design not only removes label constraints and enables synthesis for unseen gesture classes, but also substantially enhances the behavioral diversity of generated data through the rich space of DIF difference patterns. In the following, we present the architecture of the proposed generative model and describe its training and DIF-conditioned synthesis pipeline in detail.

4.4.1 Generative Architecture. Our generative model is based on a U-Net denoiser $\epsilon_\theta(\cdot)$ equipped with a DIF shift embedding module, which conditions all U-Net layers through feature-wise affine modulation. The full process consists of a forward (noising) path and a reverse (denoising) path.

Forward process. Instead of diffusing a clean spectrogram into pure Gaussian noise as in RF-Diffusion [5], we construct a controlled interpolation between two spectrograms S_1 and S_2. At diffusion step t, the forward variable is defined as:

$$\mathbf{x}_t = (1-\gamma_t)\mathrm{S}_1 + \gamma_t \mathrm{S}_2 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{12}$$

where γ_t monotonically increases with $\gamma_0 = 0$ and $\gamma_T = 1$, and σ_t determines the noise level. This formulation yields a smooth trajectory that remains structurally aligned with the transition from S_1 to a "noisy" version of S_2.

Reverse process. The reverse process aims to invert the above perturbation and progressively remove noise under the guidance of the DIF embedding $\mathbf{e}_{\mathrm{DIF}} = f_{\mathrm{DIF}}(\mathrm{S}_1, \mathrm{S}_2)$. At timestep t, the denoiser predicts the noise component:

$$\hat{\boldsymbol{\epsilon}}_t = \epsilon_\theta(\mathbf{x}_t,\ t,\ \mathbf{e}_{\mathrm{DIF}}), \tag{13}$$

and the reverse update is computed by removing the estimated noise and undoing the γ_t-dependent attenuation:

$$\mathbf{x}_{t-1} = \frac{\mathbf{x}_t - \sigma_t \hat{\boldsymbol{\epsilon}}_t}{1-\gamma_t} + \eta_t \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{14}$$

where η_t introduces controlled stochasticity analogous to the posterior variance. Iterating this update yields the final clean reconstruction $\mathrm{S}_1^\star = \mathbf{x}_0$.

4.4.2 Training and Generation. The training of *SynthNet* encourages the denoiser to correctly predict the injected Gaussian noise. For a training pair $(\mathrm{S}_1, \mathrm{S}_2)$ and its DIF embedding $\mathbf{e}_{\mathrm{DIF}}$, the denoiser is trained using an L_2 noise-prediction objective:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t,\ t,\ \mathbf{e}_{\mathrm{DIF}})\|_2^2\right]. \tag{15}$$

This loss function allows the model to learn consistent, DIF-controlled structural transformations between spectrograms.

During generation, given an unseen-class instance $\mathrm{S}^{(u)}$, we synthesize realistic intra-class variations by conditioning on DIF embeddings sampled from the base dataset. Two randomly selected base instances $(\mathrm{S}_a, \mathrm{S}_b)$ yield $\mathbf{e}_{\mathrm{DIF}}^{(a,b)} = f_{\mathrm{DIF}}(\mathrm{S}_a, \mathrm{S}_b)$, which serves as a plausible variation direction. The generation follows the reverse diffusion process, where $\mathrm{S}^{(u)}$ is updated for $t = T, \ldots, 1$ as:

$$\mathbf{x}_{t-1} = \frac{\mathbf{x}_t - \sigma_t\, \epsilon_\theta(\mathbf{x}_t,\ t,\ \mathbf{e}_{\mathrm{DIF}}^{(a,b)})}{1-\gamma_t} + \eta_t \boldsymbol{\xi}. \tag{16}$$

After the final step, the synthesized sample is obtained as $\tilde{\mathrm{S}}^{(u)} = \mathbf{x}_0$, preserving the semantic identity of the unseen class while incorporating realistic variation patterns induced by the DIF shifts.

Fig. 12(a) and (b) show two synthesized samples of "swift left & right" generated by *SynthNet*. It can be observed that they exhibit both high fidelity and clear diversity induced by DIF shifts.

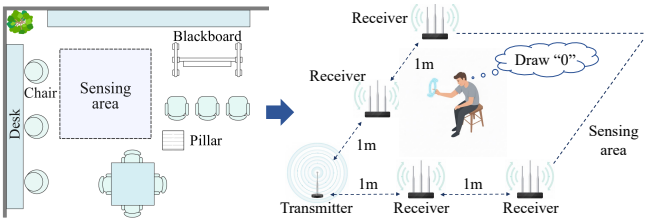


Figure 13: Experiment setup.

5 Evaluation

This section first introduces the implementation of *SynthFi*, and then details its quantitative performance.

Experiment setup. We implement a prototype of *SynthFi* using COTS WiFi devices and conduct basic experiments in a laboratory environment, as illustrated in Fig. 13. We follow a common setup of WiFi HGR [60, 64, 65, 68]: the system is composed of one transmitter and four receivers, each equipped with an Intel 5300 Network Interface Card (NIC). CSI is extracted using the Linux CSI Tool [15] from WiFi packets transmitted in the 5.54 GHz band. The transmission rate is configured to 1000 packets per second. *SynthNet* is trained on a server equipped with an H20-NVLink (96 GB) GPU using the Adam optimizer with a learning rate of 5e-4.

Data collection. We recruit ten persons to participate in the data collection process. Each participant performs 20 gesture classes (as shown in Fig. 2) within a 2m × 2m sensing area, consistent with typical deployment settings for interactive gesture recognition [60, 68]. In total, we collect more than 2900 gesture instances for evaluation. Under the default configuration, 15 gesture categories are designated as base classes, while the remaining 5 categories are treated as unseen classes for testing the generalization performance of *SynthFi*. A Convolutional Neural Network (CNN) is used as the HGR model. It includes three convolutional blocks, each comprising a convolution layer, ReLU activation, and max-pooling. These are followed by a fully connected head that contains a flattening stage, one hidden layer, and a final classification layer.

Metrics. To evaluate the performance of *SynthFi*, we adopt three quantitative metrics: Fréchet Inception Distance (FID) [5], accuracy [65], and accuracy gain (AG). FID measures generative fidelity by computing the Fréchet distance between the feature representations of real and synthesized samples. In our evaluation, a CNN based on LeNet [33] pretrained on real data is used as the feature encoder. A lower FID indicates stronger fidelity. Accuracy quantifies the likelihood that *SynthFi* correctly predicts the gesture label of an input sample, reflecting both the fidelity and diversity of the generated data. AG measures the increase in accuracy resulting

Table 2: Details of the gesture classes considered in the evaluation. "CCW" means counterclockwise.

Base	Draw '1'	Draw '2'	Draw '3'	Draw '4'	Draw '5'
	Draw '6'	Draw '7'	Draw '8'	Draw '9'	Draw '0'
	Draw '0' (CCW)	Draw 'U'	Draw 'U' reverse	Draw 'N'	Draw 'N' reverse
Unseen	Swing left and right	Swing right and left	Draw Rectangle	Draw Rectangle (CCW)	Dig

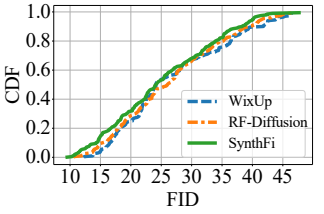


Figure 14: FID distributions of three frameworks.

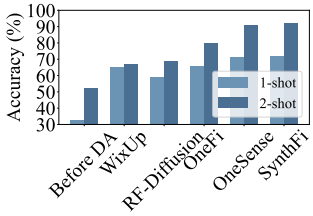


Figure 15: Overall accuracies before and after DA.

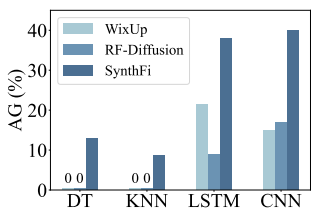


Figure 16: AGs on different classifiers.

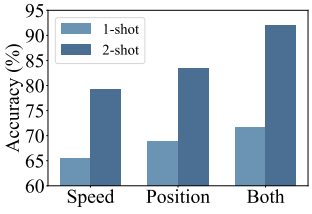


Figure 17: Impact of DIF embeddings.

from data augmentation. It is calculated as the difference between post-augmentation and pre-augmentation accuracies if this difference is positive; otherwise, AG is set to zero. Higher accuracy and AG signify better fidelity, diversity, and generalization performance.

5.1 Overall Performance

We first assess the overall performance of *SynthFi*. Under the default setting, we consider challenging scenarios where only one or two real samples are available for unseen classes (i.e., 1-shot or 2-shot settings). The number of generated samples is six times that of the real ones. The DIF embeddings are randomly selected from base-class samples during generation. We compare *SynthFi* with two state-of-the-art (SOTA) DA methods, WixUp [39] and RF-Diffusion [5], as well as two FSL methods, OneFi [60] and OneSense [65]. WixUp employs a Gaussian mixture model combined with probability-based transformations to generate new samples. RF-Diffusion develops a time-frequency diffusion model to synthesize additional data. Since RF-Diffusion is unable to generate data for unseen classes, we train it on all 20 classes and supply each unseen class with the same number of real samples as used in *SynthFi*. This setup ensures a fair comparison among these methods.

FID distribution. To assess whether the synthesized data is sufficiently realistic to capture the underlying signal propagation physics, we compute the FIDs for all three DA frameworks and present their Cumulative Distribution Functions (CDFs) in Fig. 14. The mean FIDs for WixUp, RF-Diffusion, and *SynthFi* are 26.9, 26.4, and 25.1, respectively. The FID distribution of *SynthFi* closely matches those of the two baselines. This alignment is important because previous studies have shown that both WixUp and RF-Diffusion are capable of generating high-fidelity wireless data. Achieving comparable FID distributions indicates that *SynthFi* provides a similar level of realism as these baseline approaches. As a result, *SynthFi* can reliably produce high-fidelity samples that are consistent with real signal propagation patterns and thus support downstream tasks such as DA.

Accuracy. To comprehensively evaluate the effectiveness of generated samples, we measure the classification accuracies achieved by models trained with each DA framework and FSL method. The results are presented in Fig. 15. Under the 1-shot and 2-shot settings, *SynthFi* improves the accuracies from 32.4% to 71.7% and from 52.1% to 92.1%, respectively. When WixUp is used for augmentation, the accuracies increase to 65.0% and 67.0% under the 1-shot and 2-shot settings, respectively. RF-Diffusion improves the 1-shot and 2-shot accuracies to 59.3% and 68.9%, respectively. OneFi and OneSense achieve 1/2-shot accuracies of 66.0%/80.0% and 71.5%/90.8%, respectively. These results demonstrate that *SynthFi* achieves the best performance among all five methods. Its superior performance indicates that the generated samples are sufficiently realistic to facilitate effective model learning and sufficiently diverse to capture a broader range of intra-class variations. The larger accuracy improvements achieved by *SynthFi* further confirm its ability to improve generalization to unseen samples with diverse behavioral patterns that are overlooked by existing DA frameworks. Furthermore, the 1-shot and 2-shot variances across five classes are 1.4 and 1.2, respectively. This suggests that *SynthFi* maintains stable recognition performance across different gestures.

AG on different classifiers. To further evaluate the DA effectiveness of *SynthFi*, we measure its 2-shot AG across four classifiers: two traditional models (Decision Tree, DT [8], and K-nearest Neighbor, KNN [7]) and two deep learning models (CNN and Long Short-Term Memory, LSTM [63]). The results are shown in Fig. 16. *SynthFi* achieves AG values of 37%+ on both CNN and LSTM, while WixUp remains below 22% and RF-Diffusion remains below 17%. For DT and KNN, *SynthFi* reaches AGs of 12.9% and 8.6%, whereas the other two baselines provide no performance improvement. These observations suggest that the two baselines are unable to extract sufficient diversity from limited reference samples. They are constrained by the narrow distribution of the few available examples and do not make effective use of the rich behavioral variations present in base-class data. As a result, the potentially useful diverse features become buried in the samples generated by WixUp and RF-Diffusion, making them difficult for traditional classifiers to exploit. In contrast, the behavioral diversity introduced by *SynthFi* is more easily recognized by both traditional and deep sensing models, which leads to larger gains across classifier types. Furthermore, the absolute 2-shot accuracies achieved by *SynthFi* are as high as 92.1% (CNN), 81.4% (LSTM), 83.7% (DT), and 87.9% (KNN). These results indicate that *SynthFi* provides high flexibility in sensing model selection. For example, a DT can be used in low-power or resource-constrained devices, while a CNN can be deployed on

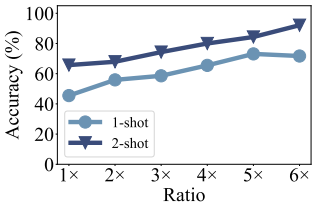


Figure 18: Effect of synthesized data's ratio.

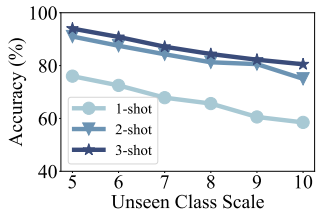


Figure 19: Effect of unseen class scale on accuracies.

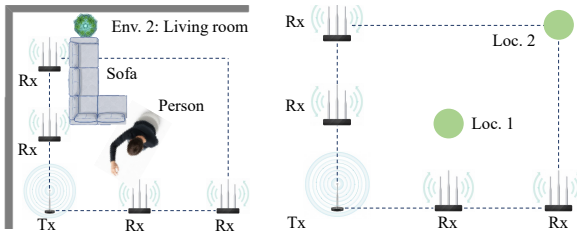


Figure 20: Experiment setup for cross-domain evaluation.

more capable platforms. This flexibility offers a clear advantage over many prior FSL paradigms that often depend on computationally heavy backbone models such as CNNs [65] or Transformers [60].

Effectiveness of DIF shift embedding. *SynthFi* enhances DA performance by introducing behavioral diversity through two explicit DIF shift embeddings, namely speed shift embedding and position shift embedding. To examine the individual contribution of each embedding, we enable only one of them during synthesis and re-evaluate the accuracy. As shown in Fig. 17, both embeddings improve recognition accuracy in the 1-shot and 2-shot settings. In both cases, the accuracy achieved with the position shift embedding is consistently higher than that obtained with the speed shift embedding. This suggests that position shift provides richer behavioral diversity than speed shift. When both embeddings are used together, the accuracy increases further. These results indicate that both speed and position shifts effectively enrich the behavioral diversity of synthesized data and substantially enhance the generalization performance of the HGR model.

Effect of γ. Note that γ, the path-loss exponent, directly determines the position DIF. To evaluate its impact on *SynthFi*, we vary γ from 1.5 to 3 in step of 0.5. The resulting 1-shot accuracies are 69.9%, 71.7%, 69.6%, and 70.6%, while the 2-shot accuracies are 89.1%, 91.4%, 89.8%, and 90.0%, respectively. These results demonstrate that *SynthFi* remains robust under different path-loss settings.

5.2 Effect of Synthesized Data's Ratio

To investigate how the ratio of synthesized samples influences DA performance, we reassess the 1-shot and 2-shot accuracies while increasing the ratio of virtual samples to real samples from 1× to 6×. The results are shown in Fig. 18. The accuracies in both settings rise as the ratio increases. When the ratio becomes larger, the accuracy curves gradually become flatter, indicating that the performance improvement approaches convergence. These findings suggest that increasing the synthesized data ratio enhances DA effectiveness, although the improvement eventually saturates when the ratio continues to grow.

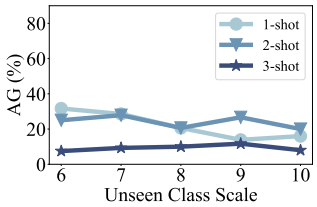


Figure 21: Effect of unseen class scale on AGs.

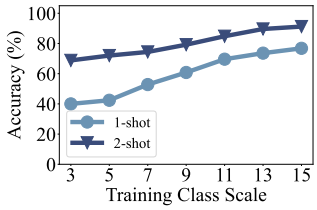


Figure 22: Effect of training class scale.

5.3 Effect of Unseen Class Scale

To investigate how performance changes as the number of unseen classes grows, we reassess the 1-shot and 2-shot accuracies while increasing the number of unseen classes from 5 to 10. In this experiment, we designate the first n classes out of all 20 as the base classes, and treat the remaining $20 - n$ classes as the customized classes. As shown in Fig. 19, the accuracy decreases in both settings as the unseen class scale becomes larger. This trend is expected because constructing precise decision boundaries becomes more challenging when the classifier must distinguish among more target classes. Even so, the recognition performance remains acceptable when the number of unseen classes reaches 10. Moreover, as illustrated in Fig. 19, the accuracy loss can be compensated by collecting just one additional real instance: the 3-shot accuracy for the 10-class setting exceeds 80%. Fig. 21 shows the AGs as the scale of the unseen classes increases. It can be observed that *SynthFi* consistently provides relatively stable gains across all settings. Furthermore, the AGs appear to increase when fewer real samples are available, which may be because the room for improvement contributed by synthesized samples becomes smaller as more real data are provided. The above observations demonstrate that *SynthFi* maintains commendable scalability when applied to unseen sensing tasks.

5.4 Effect of Training Class Scale

Recall that *SynthFi* leverages base-class samples collected by the developer or user to synthesize data for unseen classes, for which only a small number of real samples are required. This design significantly reduces the data collection burden of customized-yet-unseen gesture classes. In this experiment, we vary the number of base classes from 3 to 15 in increments of 2 and re-evaluate the recognition performance. The resulting 1-shot and 2-shot accuracies are presented in Fig. 22. The accuracies increase noticeably as the base-class scale becomes larger, which is reasonable because a richer set of base classes provides more abundant behavioral diversity for synthesis. When the number of base classes reaches 15, *SynthFi* achieves the highest performance. Encouragingly, based on the trends of the accuracy curves, the accuracy appears to continue improving as the base-class scale increases further.

5.5 Cross-domain DA Effectiveness

It is important to examine whether *SynthFi* generalizes under cross-domain conditions. In this experiment, we train *SynthNet* using data collected from a single participant in a laboratory environment under Location 1, and then evaluate the AGs across Environment 2 (living room), an unseen user, and Location 2, as shown in Fig. 20. Note that the cross-environment/user/location experiments are

three independent evaluations, where only one domain factor is varied at a time. The results show that *SynthFi* achieves AGs of 24.3%, 26.2%, and 23.0% under cross-environment, cross-user, and cross-location settings, respectively. Meanwhile, the AGs of WixUp and RF-Diffusion are 8.5%/14.1%/12.2% and 12.4%/13.8%/15.6% under these settings. These results demonstrate that *SynthFi* can effectively generalize across domain shifts by capturing class-agnostic and physically grounded DIF variations, outperforming existing works. It can be observed that the cross-domain AGs are lower than those achieved in in-domain settings. This could be because domain shifts introduce additional variations beyond DIF factors, such as environment-dependent multipath characteristics and user-specific body reflections, which reduce the distributional alignment between synthesized and target-domain data.

5.6 Extension to Other Application

We design *SynthFi* to incorporate behavioral diversity into the data synthesis process, enhancing the effectiveness of DA in HCI scenarios. To verify that our approach remains effective beyond gesture recognition, we further evaluate *SynthFi* on two open-source datasets, WiAR [13] and Widar3.0 [68]. For WiAR, we focus on large-scale activity recognition. Ten activity classes are treated as base classes, and DA is evaluated on the remaining six classes. The results show that the 1-shot and 2-shot AGs reach 13.3% and 15.1%, respectively. For Widar3.0, we consider a user identification task, where data from five users serve as base classes and the other five are regarded as unseen classes. The experiment results demonstrate that the 1-shot and 2-shot AGs reach 13.5% and 27.8%. These findings indicate that *SynthFi* can be readily and effectively extended to broader wireless sensing tasks to generate diverse data.

5.7 Overhead

In the *SynthFi* framework, the diffusion model constitutes the main computational overhead. Our analysis shows that *SynthNet* contains 12.6M parameters, which is moderate compared to lightweight architectures such as MobileNetV2 [44] (3.47M parameters) that can run smoothly on edge devices (e.g., iPhone 6s). Although the per-sample generation cost is 9.2T floating-point operations, data synthesis is conducted offline during the training stage and incurs no overhead at inference time. In practice, users can upload a small set of real samples to a server or developer for offline synthesis of virtual data, eliminating the need for on-device, real-time generation. Therefore, the proposed framework introduces negligible runtime overhead and maintains acceptable storage and computational costs for practical deployment.

5.8 User Study

To investigate users' perceptions toward WiFi-based gesture interfaces and the practical benefits enabled by *SynthFi*, we conduct a questionnaire-based user study with 42 participants. Participants are asked to rate three statements on a 5-point Likert scale (1: strongly disagree, 5: strongly agree). The questions focused on three aspects: (1) preference for privacy-preserving WiFi-based gesture interaction over camera-based approaches in privacy-sensitive scenarios; (2) the perceived capability of *SynthFi*-like techniques in facilitating the development of customized gesture interfaces for different users and applications; and (3) users' experience regarding the reduction of time and effort required for gesture data collection enabled by *SynthFi*. According to the participants' feedback in Fig. 23, we can conclude that: (1) For privacy-sensitive scenarios, participants showed a strong preference for WiFi-based gesture interfaces over camera-based approaches, with an average score of 4.50/5. This indicates that privacy preservation is an important factor influencing gesture interaction adoption. (2) Participants positively evaluate the capability of *SynthFi*-like techniques in facilitating customized gesture interfaces, achieving an average score of 4.43/5. This demonstrates the perceived value of efficient data synthesis for expanding HCI applications. (3) Participants report the most positive experience regarding *SynthFi*'s reduction of gesture data collection effort and time, with an average score of 4.64/5. This suggests that *SynthFi* can substantially improve the practicality of developing data-efficient gesture interfaces.

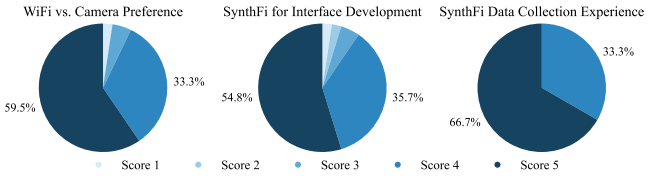


Figure 23: Score distributions of the user study questionnaire.

6 Discussion

In this work, we introduce behavior-induced factors like speed and position into the data synthesis process, which are major sources of sample variability in HCI scenarios but have been overlooked in prior studies. Our theoretical and empirical results show that these factors play an important role in enhancing intra-class diversity, although additional attributes such as arm rotation may further increase variability in practice. A possible direction for capturing a more complete set of behavioral variations is to use advanced deep learning techniques, e.g., disentangled representation learning [57], to automatically extract these factors. We leave the exploration of these broader behavioral cues for future work.

7 Conclusion

We present *SynthFi*, a novel DA framework designed to overcome the scarcity and limited diversity of real-world WiFi sensing data. *SynthFi* first quantifies key diversity-inducing physical factors to extract rich behavioral variation cues. Building upon these quantified factors, we further develop a diffusion-based generative model capable of shifting real instances along diverse behavioral dimensions, thereby producing high-fidelity synthetic data even for previously unseen classes. Extensive real-world experiments demonstrate substantial improvements in sensing performance.

Acknowledgments

We sincerely thank the anonymous ACs and reviewers for their valuable suggestions. This paper is supported in part by the National Natural Science Foundation of China under grant No. 62372400, the "Pioneer" and "Leading Goose" R&D Program of Zhejiang under grant No. 2026LDC01029(JRB), and the 2026–2028 Young Talent Support Program of the Zhejiang Association for Science and Technology under grant No. ZJSKXQT2026031.

References

[1] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM COMPUTING SURVEYS* 46, 3 (2014), 33:1–33:33.

[2] Xi Chen, Chen Ma, Michel Allegue, and Xue Liu. 2017. Taming the inconsistency of Wi-Fi fingerprints for device-free passive indoor localization. In *Proceedings of the IEEE Conference on Computer Communications*. 1–9.

[3] Xingyu Chen and Xinyu Zhang. 2023. RF Genesis: Zero-Shot Generalization of mmWave Sensing through Simulation-Based Data Synthesis and Generative Diffusion Models. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems*. 28–42.

[4] Weng Cho Chew. 1999. *Waves and fields in inhomogenous media*. John Wiley & Sons.

[5] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. 2024. RF-diffusion: Radio signal generation via time-frequency diffusion. In *Proceedings of the ACM Annual International Conference on Mobile Computing and Networking*. 77–92.

[6] Guoxuan Chi, Guidong Zhang, Xuan Ding, Qiang Ma, Zheng Yang, Zhenguo Du, Houfei Xiao, and Zhuang Liu. 2024. XFall: Domain Adaptive Wi-Fi-Based Fall Detection With Cross-Modal Supervision. *IEEE Journal on Selected Areas in Communication* 42, 9 (2024), 2457–2471.

[7] Pádraig Cunningham and Sarah Jane Delany. 2021. K-nearest neighbour classifiers-a tutorial. *ACM computing surveys* 54, 6 (2021), 1–25.

[8] Barry de Ville. 2013. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics* 5, 3 (2013), 448–455.

[9] Long Fan, Yinghui He, Lei Xie, Serene Zhang, and Jun Luo. 2026. Sense with Polyface Mirror: Enhancing Wi-Fi Sensing Diversity via Programmable Metasurfaces. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems*.

[10] Georgia Gkioxari, Ross B. Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and Recognizing Human-Object Interactions. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*. 8359–8367.

[11] Chen Gong, Bo Liang, Wei Gao, and Chenren Xu. 2025. Data Can Speak for Itself: Quality-guided Utilization of Wireless Synthetic Data. In *Proceedings of the ACM Annual International Conference on Mobile Systems, Applications and Services*. 209–222.

[12] Jun Gong, Yang Zhang, Xia Zhou, and Xing-Dong Yang. 2017. Pyro: Thumb-Tip Gesture Recognition Using Pyroelectric Infrared Sensing. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*. 553–563.

[13] Linlin Guo, Lei Wang, Chuang Lin, Jialin Liu, Bingxian Lu, Jian Fang, Zhonghao Liu, Zeyang Shan, Jingwen Yang, and Silu Guo. 2019. Wiar: A Public Dataset for Wifi-Based Activity Recognition. *IEEE Access* 7 (2019), 154935–154945.

[14] Unsoo Ha, Junshan Leng, Alaa Khaddaj, and Fadel Adib. 2020. Food and Liquid Sensing in Practical Environments using RFIDs. In *Prpceedings of the USENIX Symposium on Networked Systems Design and Implementation*. 1083–1100.

[15] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 2011. Tool release: gathering 802.11n traces with channel state information. *Computer Communication Review* 41, 1 (2011), 53.

[16] Xueqiang Han, Tianyue Zheng, Tony Xiao Han, and Jun Luo. 2025. RayLoc: Wireless Indoor Localization via Fully Differentiable Ray-tracing. https://arxiv.org/abs/2501.17881

[17] Yinghui He, Xin Li, and Jun Luo. 2026. Task-Oriented Integrated Sensing and Semantic Communications for Multi-Device Video Analytics. *IEEE Transactions on Mobile Computing* 25, 5 (2026), 7323–7337.

[18] Yinghui He, Jianwei Liu, Mo Li, Guanding Yu, Jinsong Han, and Kui Ren. 2023. SenCom: Integrated Sensing and Communication with Practical WiFi. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*. 60:1–60:16.

[19] Yuan He, Yi-Miao Sun, and Xiuzhen Guo. 2025. RF Computing: A New Realm of IoT Research. *Journal of Computer Science and Technology* 40, 4 (2025), 941–956.

[20] Yinghui He, Mingming Xu, Zhe Chen, Fu Xiao, and Jun Luo. 2025. Beam-Fi: Integrated Sensing and Communication via MU-MIMO upon Commodity Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 84:1–84:22.

[21] Yinghui He, Mingming Xu, Fu Xiao, and Jun Luo. 2025. VersaBeam: Versatile Beamforming for Integrated Sensing and Communication over Commodity Wi-Fi. In *Proceedings of the IEEE Conference on Computer Communications*. 1–10.

[22] Yinghui He, Guanding Yu, Yunlong Cai, and Haiyan Luo. 2024. Integrated Sensing, Computation, and Communication: System Framework and Performance Optimization. *IEEE Transactions on Wireless Communications* 23, 2 (2024), 1114–1128.

[23] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.

[24] Jingzhi Hu, Xin Li, Jin Gan, and Jun Luo. 2025. Poison to Cure: Privacy-preserving Wi-Fi Multi-User Sensing via Data Poisoning. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*. 47–62.

[25] Jingzhi Hu, Tianyue Zheng, Zhe Chen, Hongbo Wang, and Jun Luo. 2023. MUSE-Fi: Contactless MUti-person SEnsing Exploiting Near-field Wi-Fi Channel Variation. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*. 75:1–75:15.

[26] Hazer Inaltekin, Mung Chiang, H Vincent Poor, and Stephen B Wicker. 2009. On unbounded path-loss models: effects of singularity on wireless network performance. *IEEE Journal on Selected Areas in Communications* 27, 7 (2009), 1078–1092.

[27] Sijie Ji, Lixiang Lian, Yuanqing Zheng, and Chenshu Wu. 2024. MuSAC: Mutualistic Sensing and Communication for Mobile Crowdsensing. In *Proceedings of the IEEE International Conference on Distributed Computing Systems*. 243–254.

[28] Sijie Ji, Yaxiong Xie, and Mo Li. 2022. SiFall: Practical online fall detection with RF sensing. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems*. 563–577.

[29] Sijie Ji, Xuanye Zhang, Yuanqing Zheng, and Mo Li. 2023. Construct 3d hand skeleton with commercial wifi. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems*. 322–334.

[30] BI Justusson. 2006. Median filtering: Statistical properties. In *Two-dimensional digital signal prcessing II: transforms and median filters*. Springer, 161–196.

[31] Kaustubh Kalgaonkar and Bhiksha Raj. 2009. One-handed gesture recognition using ultrasonic Doppler sonar. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1889–1892.

[32] Belal Korany, Chitra R. Karanam, Hong Cai, and Yasamin Mostofi. 2019. XModal-ID: Using WiFi for Through-Wall Person Identification from Candidate Video Footage. In *TProceedings of the ACM Annual International Conference on Mobile Computing and Networking*. 36:1–36:15.

[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[34] Qiyue Li, Heng Qu, Zhi Liu, Nana Zhou, Wei Sun, Stephan Sigg, and Jie Li. 2021. AF-DCGAN: Amplitude Feature Deep Convolutional GAN for Fingerprint Construction in Indoor Localization Systems. *IEEE Transactions on Emerging Topics in Computational Intelligence* 5, 3 (2021), 468–480.

[35] Tianxing Li, Qiang Liu, and Xia Zhou. 2016. Practical Human Sensing in the Light. In *Proceedings of the ACM Annual International Conference on Mobile Systems, Applications and Services*. 71–84.

[36] Xin Li, Yinghui He, and Jun Luo. 2025. μCeiver-Fi: Exploiting Spectrum Resources of Multi-Link Receiver for Fine-Granularity Wi-Fi Sensing. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*. 1045–1059.

[37] Xin Li, Hongbo Wang, Zhe Chen, Zhiping Jiang, and Jun Luo. 2024. UWB-Fi: Pushing Wi-Fi towards Ultra-wideband for Fine-Granularity Sensing. In *Proceedings of the Annual International Conference on Mobile Systems, Applications and Services*. 42–55.

[38] Yichen Li, Tianxing Li, Ruchir A. Patel, Xing-Dong Yang, and Xia Zhou. 2018. Self-Powered Gesture Recognition with Ambient Light. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*. 595–608.

[39] Yin Li and Rajalakshmi Nandakumar. 2025. WixUp: A Generic Data Augmentation Framework for Wireless Human Tracking. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems*. 449–462.

[40] Jianwei Liu, Jiantao Yuan, Guanding Yu, and Jinsong Han. 2025. Efficient One-Shot Gesture Recognition for WiFi ISAC via Aug-Meta Learning. *IEEE Journal on Selected Areas in Communication* 43, 11 (2025), 3766–3781.

[41] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (2018), 861.

[42] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. CovertBand: Activity Information Leakage using Music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–24.

[43] T. Nguyen-Nhat, L. Vuong-Cong, V. Le-Ngoc, and T. Pham-Bao. 2024. Inspecting spectral centroid and relative power of allocated spectra using artificial neural network for damage diagnosis in beam structures under moving loads. *Journal of Vibration Engineering & Technologies* 12, 3 (2024), 4617–4635.

[44] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4510–4520.

[45] Jonathan Savin, Clarisse Gaudez, MA Gilles, Vincent Padois, and Philippe Bidaud. 2021. Evidence of movement variability patterns during a repetitive pointing task until exhaustion. *Applied Ergonomics* 96 (2021), 103464.

[46] Souvik Sen, Jeongkeun Lee, Kyu-Han Kim, and Paul Congdon. 2013. Avoiding multipath to revive inbuilding WiFi localization. In *Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services*. 249–262.

[47] SIGRobotics. 2025. Github for LeKiwi Robot. https://github.com/SIGRobotics-UIUC/LeKiwi/tree/main.

[48] Wenfan Song, Weiye Xu, Jianwei Liu, Yuanqing Zheng, Xinhuai Wang, Xian Xu, and Jinsong Han. 2025. Anti-Spoofing and Mask-Supported Face Authentication using mmWave without On-site Registration. *IEEE Transactions on Dependable*

and Secure Computing (2025), 1–18.

[49] Shunta Suzuki, Takashi Amesaka, Hiroki Watanabe, Buntarou Shizuki, and Yuta Sugiura. 2024. EarHover: Mid-Air Gesture Recognition for Hearables Using Sound Leakage Signals. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*. 129:1–129:13.

[50] Fei Wang, Jinsong Han, Shiyuan Zhang, Xu He, and Dong Huang. 2018. CSI-Net: Unified Human Body Characterization and Action Recognition. *CoRR* abs/1810.03064 (2018).

[51] Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. 2024. XRF55: A Radio Frequency Dataset for Human Indoor Action Analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 21:1–21:34.

[52] Fei Wang, Tingting Zhang, Wei Xi, Han Ding, Ge Wang, Di Zhang, Yuanhao Cui, Fan Liu, Jinsong Han, Jie Xu, and Tony Xiao Han. 2025. A Survey on Wi-Fi Sensing Generalizability: Taxonomy, Techniques, Datasets, and Future Research Prospects. https://arxiv.org/abs/2503.08008

[53] Fei Wang, Tingting Zhang, Wei Xi, Han Ding, Ge Wang, Di Zhang, Yuanhao Cui, Fan Liu, Jinsong Han, Jie Xu, and Tony Xiao Han. 2026. A Survey on Wi-Fi Sensing Generalizability: Taxonomy, Techniques, Datasets, and Future Research Prospects. *IEEE Communications Surveys & Tutorials* (2026).

[54] Jie Wang, Qinhua Gao, Xiaorui Ma, Yunong Zhao, and Yuguang Fang. 2020. Learning to Sense: Deep Learning for Wireless Sensing with Less Training Efforts. *IEEE Wireless Communications* 27, 3 (2020), 156–162.

[55] Jie Wang, Yunong Zhao, Xiaorui Ma, Qinghua Gao, Miao Pan, and Hongyu Wang. 2020. Cross-Scenario Device-Free Activity Recognition Based on Deep Adversarial Networks. *IEEE Transactions on Vehicular Technology* 69, 5 (2020), 5416–5425.

[56] Minsi Wang, Bingbing Ni, and Xiaokang Yang. 2017. Recurrent Modeling of Interaction Context for Collective Activity Recognition. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*. 7408–7416.

[57] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. 2024. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 9677–9696.

[58] Anusha I. Withana, Roshan Lalintha Peiris, Nipuna Samarasekara, and Suranga Nanayakkara. 2015. zSense: Enabling Shallow Depth Gesture Recognition for Greater Input Expressivity on Smart Wearables. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 3661–3670.

[59] Chunjing Xiao, Daojun Han, Yongsen Ma, and Zhiguang Qin. 2019. CsiGAN: Robust Channel State Information-Based Activity Recognition With GANs. *IEEE Internet of Things Journal* 6, 6 (2019), 10191–10204.

[60] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. 2021. OneFi: One-Shot Recognition for Unseen Gesture via COTS WiFi. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems*. 206–219.

[61] Weiye Xu, Wenfan Song, Jianwei Liu, Yajie Liu, Xin Cui, Yuanqing Zheng, Jinsong Han, Xinhuai Wang, and Kui Ren. 2022. Mask does not matter: anti-spoofing face authentication using mmWave without on-site registration. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*. 310–323.

[62] Koji Yatani and Khai N. Truong. 2012. BodyScope: a wearable acoustic sensor for activity recognition. In *Proceedings of the ACM International Conference on Ubiquitous Computing*. 341–350.

[63] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation* 31, 7 (2019), 1235–1270.

[64] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2021. Widar3. 0: Zero-effort cross-domain gesture recognition with Wi-Fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8671–8688.

[65] Leqi Zhao, Rui Xiao, Jianwei Liu, and Jinsong Han. 2024. One is Enough: Enabling One-shot Device-free Gesture Recognition with COTS WiFi. In *Proceedings of the IEEE Conference on Computer Communications*. 1231–1240.

[66] Tianming Zhao, Jian Liu, Yan Wang, Hongbo Liu, and Yingying Chen. 2018. PPG-based Finger-level Gesture Recognition Leveraging Wearables. In *Proceedings of the IEEE Conference on Computer Communications*. 1457–1465.

[67] Lili Zheng, Bin-Jie Hu, Jinguang Qiu, and Manman Cui. 2020. A Deep-Learning-Based Self-Calibration Time-Reversal Fingerprinting Localization Approach on Wi-Fi Platform. *IEEE Internet of Things Journal* 7, 8 (2020), 7072–7083.

[68] Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi. In *Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services*. 313–325.