# Towards Metasurface-Assisted Sample Synthesis for Wi-Fi Human Sensing

Jiaming Gu, Shaonan Chen, Yimiao Sun, Yadong Xie, Rui Xi,
Yuan He *Senior Member, IEEE*, Qiang Cheng, *Senior Member, IEEE*

**Abstract**—Wi-Fi human sensing has attracted numerous research studies over the past decade. The rapid advancement of machine learning technology further boosts the development of Wi-Fi human sensing. However, current Wi-Fi human sensing suffers from the "data scarcity" problem: all the existing proposals require collecting a large amount of human-based datasets to train the sensing models, which is labor-intensive and may raise ethical concerns in certain scenarios. This obstacle seriously restricts the size, quality, and diversity of available datasets, thereby affecting the sensing performance in terms of accuracy and cross-domain applicability. In order to solve this problem, we in this paper propose Metasurface-Assisted Sample Synthesis (MASS), a novel approach to synthesize high-fidelity Wi-Fi sensing samples that effectively capture both the essential features of human motion and environment-specific multipath characteristics without requiring human involvement. The evaluation results show that MASS is effective in boosting machine learning performance, improving classification accuracy by 18%, and enhancing the cross-domain sensing accuracy by 22%. We further analyze inherent synthesis distortions stemming from hardware limitations and introduce a mitigation technique, which significantly enhances data fidelity, achieving 91.8% accuracy even when training exclusively on synthesized samples. These findings underscore the potential of MASS to facilitate the creation of high-quality, diverse datasets with minimal human involvement and associated labor costs.

**Index Terms**—Wi-Fi Sensing, Data Augmentation, Metasurface, Machine Learning, Cross-domain Sensing.

## 1 Introduction

Benefiting from the ubiquitous Wi-Fi infrastructures, Wi-Fi sensing technologies exhibit versatile abilities to enable diverse applications, including location estimation [1], [2], health monitoring [3], activity recognition [4], [5], *etc.* Recent advancements further release the potential of Wi-Fi sensing by integrating deep learning (DL) models into sensing systems, with which the sensing ability and accuracy can be significantly enhanced [2], [6], [7].

Comprehensive and diverse labeled datasets are necessary for training a robust DL model. When it comes to Wi-Fi human sensing, a sample of sensing data is the coupling result of human motions and the complex multipath propagation of the indoor environment (commonly referred to as the sensing domain). This inherent coupling results in data scarcity through two primary avenues. First, the requirement for human participation makes the process labor-intensive and may raise ethical concerns, especially when collecting health-related samples from patients or individ-

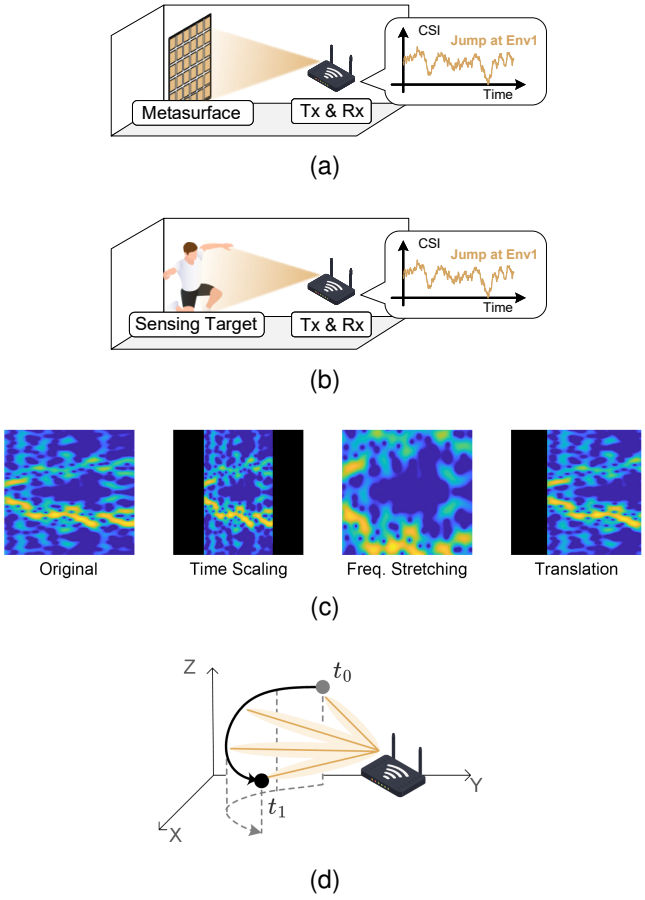



Fig. 1. Comparison of different approaches for obtaining Wi-Fi sensing data, contrasting MASS with traditional human-based data collection and existing data augmentation methods. (a) Data collection using MASS. (b) Traditional human-based data collection. (c) Transformation-based data augmentation. (d) Illustration of model-driven data augmentation.

- *(Co-primary authors: Jiaming Gu and Shaonan Chen.)*
- *(Corresponding authors: Yuan He and Qiang Cheng.)*
- *J. Gu, Y. Sun, Y. Xie, Y. He are with the School of Software and BNRist, Tsinghua University, Beijing, China. E-mail: {gujm24, sym21}@mails.tsinghua.edu.cn, {ydxie, heyuan}@tsinghua.edu.cn.*
- *R. Xi is with the University of Electronic Science and Technology of China, China. E-mail: ruix.ryan@gmail.com.*
- *S. Chen and Q. Cheng are with the School of Information Science and Engineering, Southeast University, Nanjing, China. E-mail: {220230750, qiangcheng}@seu.edu.cn.*

uals with disabilities. Second, the multipath characteristics are a crucial factor in sensing. The usability of data collected in one domain may degrade when it is used to train a DL model for sensing in other domains [2], [8], [9].

Data augmentation is deemed a feasible solution to tackle the data scarcity problem. Some existing works acquire more data by simply applying different transformations on the original signals [10] (*e.g.,* scaling, translating the axis, or adding noise), but are likely to alter the physical properties of signals. Model-driven approaches take into account the signal propagation process [11], [12], but overlook the complex multipath effects, which lead to low robustness of the DL models. Recent works propose to employ style transfer or generative AI for sample synthesis [9], [13]–[16], but still require a large amount of data for training and fall into the "chicken-and-egg" dilemma.

Inspired by the advancements in metasurface research [17]–[20], we in this paper propose **M**etasurface-**A**ssisted **S**ample **S**ynthesis (**MASS**) for synthesizing Wi-Fi sensing samples. Unlike the aforementioned data-driven generative models, MASS operates based on physical principles and requires no prior model training. This gives it a "plug-and-play" nature, allowing it to be deployed in new environments to generate data immediately. Fig. 1 compares MASS with human-based data collection and other data augmentation approaches. Leveraging its inherent capacity for waveform manipulation, a metasurface can emulate the effect of human motion on Wi-Fi signals. By replacing the human with the metasurface in the task of collecting sensing samples in real environments, MASS presents a new and much more efficient method to synthesize sensing samples without human participation, significantly reducing labor costs and avoiding ethical concerns.

Our comprehensive evaluation demonstrates that MASS effectively synthesizes high-quality Wi-Fi sensing samples, significantly boosting machine learning performance in various scenarios. The thorough fidelity analysis confirms the high overall similarity but also characterizes subtle distortions that primarily stem from the inherent physical limitations of the metasurface hardware.

Crucially, we propose an optinal Distortion Filtering Module (DFM) to mitigate these distortions. The key observation is that these synthesis distortions manifest predictably within the micro-Doppler spectrum (MDS) as low-energy components, and thus the DFM is designed to suppress these identified distortions while preserving the primary motion signatures, thereby further enhancing the quality and practical utility of the synthesized data.

Our contributions can be summarized as follows:

- We present the theoretical framework underpinning MASS and develop a complete workflow for its realization, elucidating how metasurfaces can be utilized to emulate the effect of human motion on Wi-Fi signals.
- We present the approach of MASS and elaborate on the procedure from collecting an activity template to acquiring metasurface-synthesized samples. This approach significantly reduces labor costs and alleviates ethical concerns while efficiently capturing the intrinsic domain characteristics of the sensing environment.
- We implement and evaluate the core MASS approach, demonstrating its significant potential. Without DFM, MASS improves classification accuracy by 18% (outperforming noise augmentation by 9%) and enhances cross-domain sensing accuracy by up to 22% (outperforming noise augmentation by 13%).
- We conduct an in-depth fidelity analysis, mathematically characterize the synthesis distortions, and propose DFM to mitigate these side-effects. Evaluations show that integrating the DFM substantially elevates MASS's performance, boosts cross-domain accuracy by an additional 11.4% (reaching 94.1% in drastic domain change scenarios), and achieves the accuracy of 91.8% even when training exclusively on synthesized samples.

The structure of this paper is as follows: Sec. 2 provides a comparison between MASS and related studies. Sec. 3 elaborates on the design of MASS. The evaluation is presented in Sec. 4. Finally, Sec. 5 concludes the study.

## 2 Related Work

This section provides an overview of related studies and contrasts them with our research to highlight the novelty and significance of our work.

### 2.1 Data Augmentation for Wi-Fi Sensing

**Transformation-Based Augmentation**: In computer vision, data augmentation is widely used to mitigate data scarcity. New samples are generated by applying transformations to the original images. As shown in Fig. 1(c), Dense-LSTM [10] extends this concept to Wi-Fi sensing by adding noise, scaling the time axis, or stretching the frequency axis in spectrograms. However, these methods may alter the inherent physical properties of the Wi-Fi signal [11]. For example, stretching the Doppler frequency dimension might incorrectly transform "walking" into "running."

**Model-Driven Augmentation**: To preserve the intrinsic physics of Wi-Fi signals, model-driven data augmentation methods have been proposed [11], [21]–[23]. For example, SimHumalator [22] models human limbs as scatter points shown in Fig. 1(d) and computes their impact on Wi-Fi signals to synthesize samples. However, these studies ignore the modeling of complex environmental multipath effects, such as wall reflections, despite being critical for data diversity [8], [9].

**Data-Driven Augmentation**: With advancements in DL, data-driven augmentation methods have emerged. Some studies [15], [16] employ style transfer for cross-domain adaptation, while others [13], [14], [24] utilize generative AI to create new samples. However, they encounter issues with interpretability, as AI models may not adhere to physical laws [25], [26]. Furthermore, data-driven approaches face the "chicken-and-egg" dilemma, as they require extensive data for training before they can generate new data.

In contrast, MASS is theoretically robust and makes no simplifications to the environment. It synthesizes new Wi-Fi sensing samples that capture human motions and the environment multipath simultaneously without the need for an extensive Wi-Fi sensing dataset.

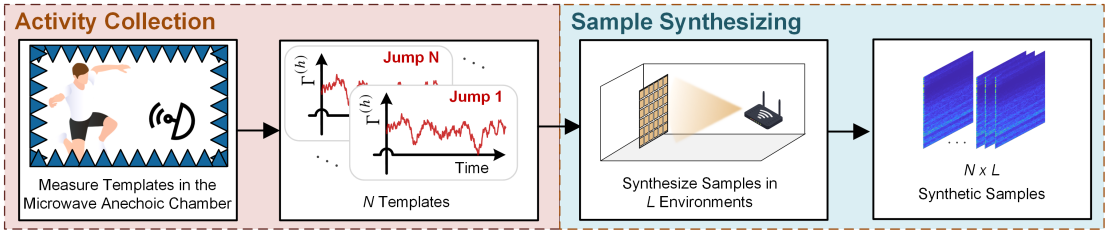



Fig. 2. Overview of the MASS. The two-stage process involves collecting $N$ activity templates and then synthesizing $N \times L$ sensing samples by using a metasurface to emulate these templates in $L$ different environments.

### 2.2 Domain-Independent Sensing

Besides data augmentation, some studies aim to extract domain-independent features, thus performing well in cross-domain sensing scenarios [27]–[29]. Some other works use transfer learning to adapt models to unseen domains with minimal labeled data [30], [31]. However, DPSense [32] suggests that domain-specific features cannot be entirely removed. Although these works have improved cross-domain sensing performance, they still cannot avoid the labor-intensive sample collection process. For example, the CrossSense [9] dataset required a three-month-long effort to collect approximately 1.2 million samples. With MASS, a one-time template collection of only a few hours enables the autonomous synthesis of 1520 high-fidelity samples samples across four different environments. The number of 1520 is not an upper limit. Our method can generate an unlimited number of samples in any new environment with marginal additional human labor. This makes our work orthogonal and complementary to the aforementioned methods.

## 3 Design

This section first presents an overview of MASS, then elaborates on its sample synthesis capability using the MASS theory, outlines the procedure of sample synthesis, characterizes the distortions introduced during the synthesis process, and finally proposes the DFM to mitigate these distortions.

### 3.1 Overview

As shown in Fig. 2, MASS operates in two stages: the activity collection stage and the sample synthesizing stage. In the first stage, activity templates are gathered to capture the impact of human activities on Wi-Fi signals. In the second stage, the metasurface is deployed alongside Wi-Fi sensing equipment across various environments. At this point, the metasurface simulates human movements based on the recorded activity templates. Taking into account the real-world setting in which the process occurs, the samples collected by the Wi-Fi sensing equipment represent a combination of human activity and environmental characteristics. By replacing humans with the metasurface, MASS avoids high labor costs while still producing substantial sensing samples. For example, a total of $N \times L$ samples can be synthesized from $N$ activity templates and $L$ environments.

### 3.2 MASS Theory

But how does a metasurface emulate the human activity? This section presents the MASS theory, which first models the environment when either a human or metasurface is present and then translates the problem of emulating human activities to finding the appropriate control voltage sequence, or coding sequence, for the metasurface.

When a human is present in a multipath-rich indoor environment, as illustrated in Fig. 3(a), the CSI of the $n$-th Wi-Fi package is modeled as [33]:

$$\begin{aligned}\boldsymbol{H}_n(f) &= \boldsymbol{H}_n^{(s)}(f) + \boldsymbol{H}_n^{(d)}(f) \\ &= \boldsymbol{H}_n^{(s)}(f) + \sum_{p=1}^{P_n} \boldsymbol{H}_n^{(p_r)}(f)\Gamma_n^{(h)}(\boldsymbol{q})\boldsymbol{H}_n^{(p_i)}(f),\end{aligned} \tag{1}$$

where $\boldsymbol{H}_n^{(s)}(f)$ is the static component that arises from static objects such as walls and chairs. $\boldsymbol{H}_n^{(d)}(f)$ is the dynamic component from the interaction between the human and the environment, further decomposable into $P_n$ multipath components. For the $p$-th component, the signal propagates along the path $p_i$, and is reflected by a part of the human body simplified as a scatter point at $\boldsymbol{q}$ with reflection coefficient $\Gamma_n^{(h)}(\boldsymbol{q})$. The signal then propagates along the path $p_r$ to the receiver. Each path, $p_k$, is modeled as $\boldsymbol{H}_n^{(p_k)}(f) = a_{n,p_k}(f)\exp(-j\,2\pi f\,\tau_{n,p_k})$, where $a_{n,p_k}(f)$ is the attenuation and $\tau_{n,p_k}$ is the path delay.

When the metasurface, rather than the human, is present in the same environment, the $n$-th CSI can be expressed as:

$$\boldsymbol{H'}_n(f) = \boldsymbol{H}_n^{(s)}(f) + \sum_{p=1}^{P_n} \boldsymbol{H}_n^{(p_r)}(f)\Gamma_n^{(m)}(\boldsymbol{q}_A)\boldsymbol{H}_n^{(p_i)}(f), \tag{2}$$

where $\Gamma_n^{(m)}(\boldsymbol{q}_A)$ denotes the reflection coefficient of the meta-atom at $\boldsymbol{q}_A$. $\Gamma_n^{(m)}(\boldsymbol{q}_A)$ is determined by the metasurface design and the voltage $V_n(\boldsymbol{q}_A)$ applied to the meta-atom. As depicted in Fig. 3(b) and Fig. 3(c), our metasurface comprises 16×8 meta-atoms, each capable of controlling phase reflection from 0 to $2\pi$, while keeping the amplitude almost unchanged. The $16 \times 8$ array size is chosen to provide a sufficiently large reflection aperture while maintaining manageable production costs. The full $2\pi$ phase control is achieved through the dual-resonance design within each meta-atom. Each meta-atom integrates a varactor diode, whose capacitance is modulated by the applied control voltage, which than tunes the resonant frequency of the meta-atom and enables precise and continuous control over the reflection phase.

Eq. 2 diverges from Eq. 1 solely by substituting $\Gamma_n^{(h)}(\boldsymbol{q})$ with $\Gamma_n^{(m)}(\boldsymbol{q}_A)$. Thus, the human activity is replicated by equating $\Gamma_n^{(m)}(\boldsymbol{q}_A)$ to $\Gamma_n^{(h)}(\boldsymbol{q})$. To emulate an **activity template** $\{\Gamma_n^{(h)}(\boldsymbol{q})\}$, an accurate **coding sequence** $\{V_n(\boldsymbol{q}_A)\}$ is

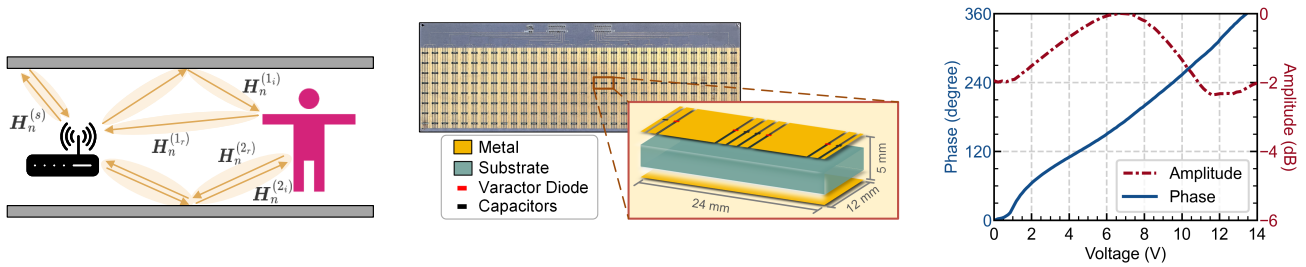



Fig. 3. Core components and models underlying MASS. (a) Wi-Fi sensing model in a multipath-rich indoor environment. (b) The metasurface used in MASS. (c) Reflection coefficient of the meta-atom as a function of control voltage.

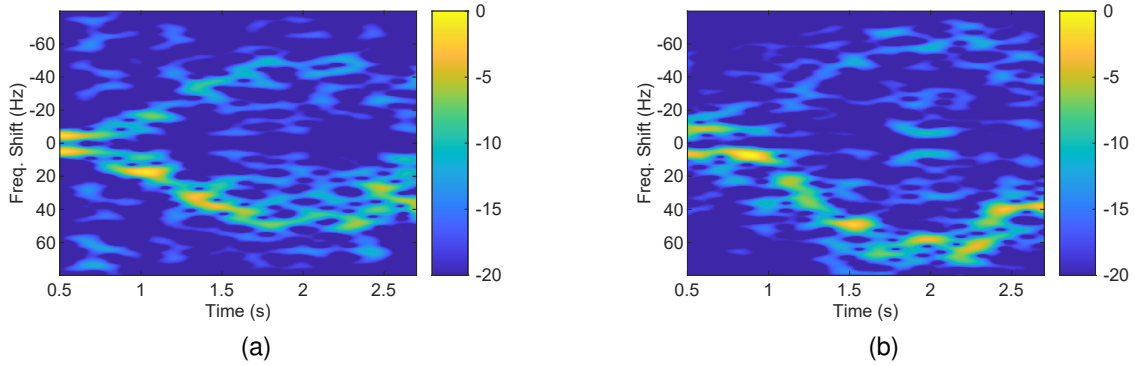



Fig. 4. Comparison of MDS illustrating the fidelity of synthesized samples in capturing human motion dynamics. (a) MDS from a human-based sensing sample. (b) MDS from a synthesized sample.

required so that the impacts of metasurface, $\Gamma_n^{(m)}(\boldsymbol{q}_A)$, are aligned with the impacts of the human body, $\Gamma_n^{(h)}(\boldsymbol{q})$.

Note that the MASS theory is fundamentally general. The dynamic scatter point $\boldsymbol{q}$ described in Eq. (1) can theoretically represent any moving object, not just a human. In this work, we focus on human motion as it is the core of synthesizing high-fidelity Wi-Fi sensing samples.

### 3.3 Sample Synthesis Procedure

Building on the MASS theory, the sample synthesis process unfolds as follows:

**Activity Template Collection**: Activity templates are collected in a microwave anechoic chamber. In the chamber, multipath effects are minimized and the impacts of human activities on Wi-Fi signals are isolated. A vector signal analyzer [34] is used to transmit a single-frequency sinusoidal wave, a volunteer performs the activity in the chamber, and the echo signal is collected as the activity template. In practice, since the human sensing necessitates a space where the volunteer can move freely, we modified an office space and surrounded the sensing area with absorbing materials to reduce reflections (Domain A, Fig. 6(a)). During the measurement of activity templates, the antennas are placed on a 1.5-meter-high stand, directly facing the human torso.

**Coding Sequence Derivation**: The coding sequence is then derived from the activity template. Each activity template is filtered by a bandpass filter to remove the direct current component and focuses on the human motion. The control voltage series is then determined by referring to the phase extracted from the filtered template according to the relationship depicted in Fig. 3(c). This resulting voltage sequence directly determines the metasurface's time-varying reflection coefficient, $\Gamma_n^{(m)}$, which, as modeled in Eq. (2), allows the method to replicate the impact of human motions into the final synthesized CSI, $\boldsymbol{H'}_n(f)$.

The derivation process is computationally efficient. Generating the coding sequence from a 3-second template takes less than 50 ms on a consumer-grade laptop.

**Sample Synthesis**: With the coding sequences, the metasurface is prepared to emulate human activities. The metasurface and Wi-Fi devices are deployed in a real-world environment, with coding sequences applied to the metasurface. And Wi-Fi devices capture the CSI as synthesized samples.

Critically, the metasurface does not need to occupy the exact same spatial location as the original human. As described by our model in Eq. (2), the Wi-Fi receiver captures a signal that holistically combines the emulated motion (from the coding sequence) with the authentic multipath characteristics of the environment, which are determined by the metasurface's physical location. This allows us to generate diverse samples from a single template by simply placing the metasurface at different locations.

The synthesis process itself occurs in real-time, (*e.g.*, a 3-second sample takes 3 seconds to generate), underscoring the efficiency of MASS for creating large-scale datasets.

As illustrated in Fig. 4, a comparison of the MDS for the "running approach" activity reveals strong similarities between the human-based sensing sample and the syn-

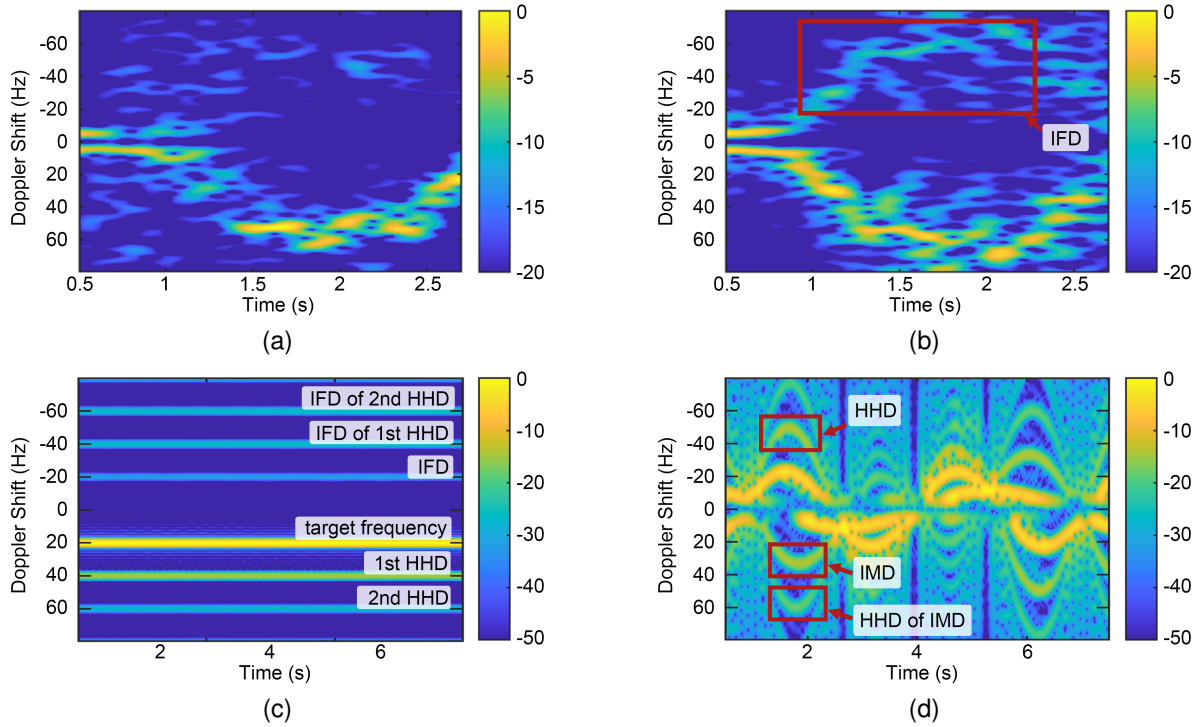



Fig. 5. Characterization of synthesis distortions in synthesized samples. (a) MDS from a real human activity sample in the anechoic chamber, serving as a baseline. (b) MDS from a corresponding synthesized sample, illustrating the presence of IFD. (c) MDS from a synthesized sample with a specially crafted template, showing the presence of IFD and HHD. (d) MDS from another specially synthesized sample, demonstrating the presence of IMD.

thesized sample. Both samples are collected in Domain B (Fig. 6(a)) and both spectra demonstrate a clear sequence: initiation of running around 1 second, speed maintenance until approximately 2 seconds, and deceleration around 2.6 seconds. This demonstrates MASS's ability to capture the dynamic characteristics of human activities. The discrepancies between the two spectra are primarily due to the inherent, non-deterministic nature of human movement.

It is important to note that our method focuses on simulating the entire activity as a whole, rather than simulating specific, isolated time slots. Fig. 4 serves as an intuitive demonstration of the overall similarity for all activities. For a more rigorous analysis, readers are referred to Sec. 4.

### 3.4 Characterizing Synthesis Distortions

To further investigate the fidelity of MASS, we conducted a preliminary analysis in the anechoic chamber, where multipath effects are minimized. As shown in Fig. 5, distortions are acknowledged. There are two primary distortion types: Harmonic Distortion (HMD) and Intermodulation Distortion (IMD). HMD is further divided into Image Frequency Distortion (IFD) and Higher Harmonic Distortion (HHD).

(1) IFD: Comparing the MDS of real (Fig. 5(a)) and synthesized (Fig. 5(b)) samples for the same activity collected in the anechoic chamber reveals this type of distortion. While the real sample primarily shows expected positive Doppler shifts, the synthetic sample exhibits additional prominent spectral lines in the negative frequency range, mirroring the true positive frequencies. This physically unrealistic "mirror ghost" effect is termed IFD.

(2) HHD: Further investigation involves constructing a template simulating a single object approaching the Wi-Fi device at a constant velocity (*e.g.*, 20 Hz Doppler). The MDS for the synthesized sample is shown in Fig. 5(c). Ideally, it should yield only the 20 Hz line. However, there are also distinct spectral lines at integer multiples of 20 Hz such as 40 Hz and 60 Hz, and their corresponding negative images at -40 Hz and -60 Hz. These spurious components at harmonic frequencies are termed HHD.

(3) IMD: When the synthesis involves multiple simultaneous motions (*e.g.*, two objects moving sinusoidally, Fig. 5(d)), another type of distortion appears, highlighted by the red box. Unlike IFD and HHD related to a single motion's frequency, IMD manifests as spectral components at linear combinations of the fundamental frequencies associated with the objects (*e.g.*, $f_1 \pm f_2$, $2f_1 \pm f_2$, etc.). This interaction between multiple frequency components is termed IMD.

These distortions stem from two primary sources: (1) the inherent limitation of the metasurface hardware, specifically the coupling between phase and amplitude control (Fig. 3(c)), and (2) non-linearity introduced during the phase-prioritized synthesis process. The phase-amplitude coupling forces a non-ideal amplitude profile when achieving the target phase, which is the cause of IFD and HHD. The synthesis process itself, particularly the step of extracting phase from a template, introduces non-linearity that manifests as IMD.

To elucidate these mechanisms mathematically, ignoring static reflections and multipath propagation, consider a single meta-atom's contribution to CSI at time $n$: $H_n = \Gamma(V_n) = A(V_n)e^{jP(V_n)}$. $V_n$ is the control voltage, while $A(V)$ and $P(V)$ are the amplitude and phase responses

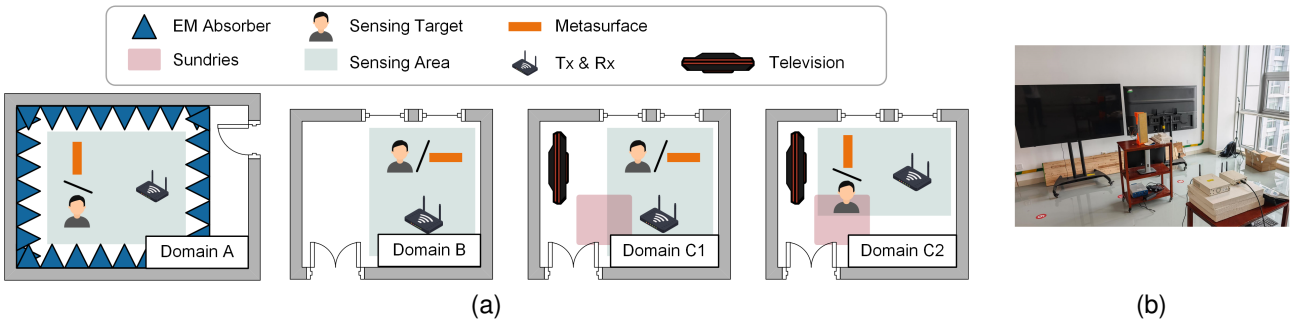



Fig. 6. Sensing domain layouts and the scenario photograph. (a) Layouts of the Domain A, B, C1, and C2. (b) The scenario of domain C2.

(Fig. 3(c)). To achieve a target phase $\omega_n$, the required voltage is $V_n = P^{-1}(\omega_n)$, leading to the synthesized signal $H_n = A(P^{-1}(\omega_n))e^{j\omega_n}$. Let $\bar{A}(\omega) = A(P^{-1}(\omega))$. Due to the $2\pi$ periodicity of phase, $\bar{A}(\omega)$ is periodic and can be expanded in a Fourier series: $\bar{A}(\omega) = \sum_{k=-\infty}^{\infty} c_k e^{jk\omega}$. Substituting this gives:

$$H_n = \bar{A}(\omega_n)e^{j\omega_n} = \sum_{k\in\mathbb{Z}} c_k e^{j(k+1)\omega_n}. \tag{3}$$

The origin of IFD and HHD can be understood by using the example from Fig. 5(c), where the desired phase evolves linearly, $\omega_n = \omega_0 n$ (corresponding to a constant Doppler frequency $f_0 = \omega_0/(2\pi T_s)$), Eq. (3) becomes:

$$H_n = c_0 e^{j\omega_0 n} + \sum_{k\in\mathbb{Z}, k\neq 0} c_k e^{j(k+1)\omega_0 n}. \tag{4}$$

In Eq. (4), the first term ($c_0 e^{j\omega_0 n}$) represents the desired signal at the frequency $f_0$. The summation term encapsulates harmonic distortions during synthesis. Terms with $k = 1, 2, \ldots$ ($c_1 e^{j2\omega_0 n}, c_2 e^{j3\omega_0 n}, \ldots$) correspond to HHD components at frequencies $2f_0, 3f_0, \ldots$. And terms with $k = -3, -4, \ldots$ correspond to the IFD of these higher harmonics (*e.g.*, $-2f_0, -3f_0$).

The origin of IMD can be understood from the synthesis non-linearity, even assuming ideal amplitude control ($A(V) = 1$). Consider a template with two motion components $T_n = s_1(n) + s_2(n)$ (with frequencies $f_1, f_2$, respectively). Synthesis targets the phase $\omega_n = \angle T_n = \angle(s_1(n) + s_2(n))$. The resulting signal is $H_n = e^{j\omega_n} = e^{j\angle(s_1(n)+s_2(n))}$. The operation $T_n \to \angle T_n \to e^{j\angle T_n}$ is inherently non-linear. Viewing $H_n$ as a function $H_n(s_1, s_2)$, its multivariate Taylor series expansion around zero is:

$$H_n = \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty} c'_{k_1 k_2} s_1^{k_1}(n) s_2^{k_2}(n). \tag{5}$$

The higher-order terms ($k_1 + k_2 > 1$) involve products $s_1^{k_1}(n)s_2^{k_2}(n)$. Since $s_1(n)$ contains $f_1$ and $s_2(n)$ contains $f_2$, these products generate frequencies of the form $k_1 f_1 + k_2 f_2$, which are the IMD components.

### 3.5 Distortion Filtering Module

To further refine the synthesized samples and enhance its utility for downstream machine learning tasks, we in this section propose DFM.

Our key observation is that while the primary signatures corresponding to the intended emulated motion dominate the MDS, potential distortions often manifest as distinct spectral components with significantly lower energy levels. Leveraging this characteristic, the DFM employs a threshold-based filtering strategy directly on the MDS to suppress these low-energy components. Specifically, DFM normalizes the MDS so the peak energy is 0 dB, and applys a threshold $t$, so that spectral bins with energy below $t$ are set to $t$. The selection of this threshold, $t$, is guided by a careful empirical analysis. We observe a distinct trade-off: overly permissive thresholds are insufficient to remove noticeable distortions, while overly aggressive thresholds begin to erode subtle but meaningful components of the human motion signature. The value of $t = -18$ dB is ultimately chosen as it provides an effective balance. The DFM thus yields cleaner spectral representations, improving downstream model performance.

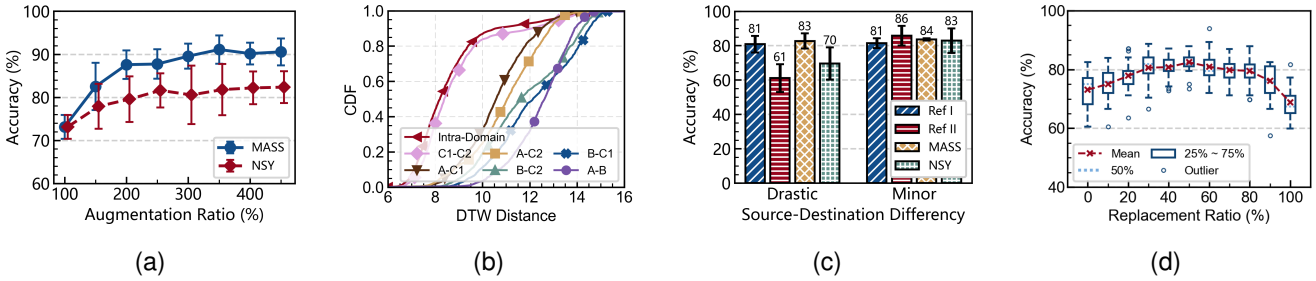
## 4 Evaluation

### 4.1 Experiment Setup

**Overview**: Our evaluation is structured to comprehensively assess the MASS framework, proceeding in two stages. We employ LeNet [35] as the activity recognition classifier throughout the evaluation. In the first stage, we the accuracy gain achieved by MASS (Sec. 4.2, 4.3, 4.4). In this stage, raw CSI undergoes only minimal preprocessing (phase correction and static removal in [28], [36]) before direct input to the classifier, providing a stringent test of the raw synthesized data. However, due to the distortions (Sec. 3.4), the benefit is limited by using raw data directly. To address this practically, we introduced the DFM in the second stage, where we evaluate the performance enhancements achieved by applying the DFM (Sec. 4.5).

**Experiment Environments**: Sensing samples are collected in four distinct environments, as depicted in Fig. 6(a). Domain A, the microwave anechoic chamber, serves both as the activity template collection site and a sensing sample collection site. Domains B and C emulate typical indoor environments using the same room, with Domain B being more spacious and Domain C containing more miscellaneous items. The key difference between Domain C1 and C2 is the positioning of the sensing target and the Wi-Fi equipment, and Fig. 6(b) shows the scenario of Domain C2.

**Sensing Equipment**: PicoScenes is selected as the Wi-Fi sensing platform, which is compliant with the standard Wi-Fi protocol [37]. For data collection, PicoScenes is configured to transmit packets at 800 Hz over a duration of 3 seconds, resulting in each sample comprising 2400 packets.




(a) (b) (c) (d)

Fig. 7. Evaluation results demonstrating MASS effectiveness and fidelity. (a) Classification accuracy across augmentation ratios. (b) DTW distance between samples from different domains. (c) Classification accuracy for different cross-domain sensing groups and methods. (d) Classification accuracy across replacement ratios.

**Collected Dataset**: Four volunteers of varying heights and weights participate in performing seven common daily actions within the specified domains. These actions include Walk Approach (WA), Walk Away (WW), Run Approach (RA), Run Away (RW), Sit Down (SD), Stand Up (SU), and Jump (JMP). Each action is repeated multiple times to ensure the diversity of the dataset. After filtering out corrupted samples (*e.g.*, instances where a volunteer executed the wrong action or when accidental interference occurred), the dataset comprises 140 activity templates, 659 human-based samples, and 1520 synthesized samples. Our experiments have strictly followed the IRB of our institute and we have confirmed the consent from the volunteers.

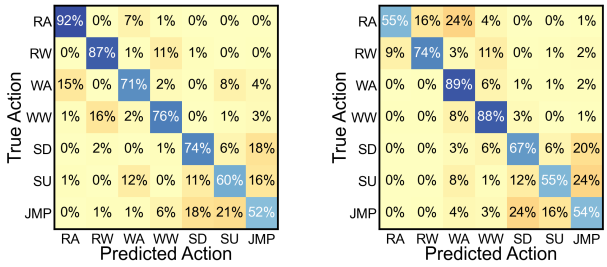
### 4.2 Accuracy Improvement with MASS

This section verifies the effectiveness of MASS in improving activity recognition accuracy by augmenting human-based sensing samples with synthetic samples at different augmentation ratios. The augmentation ratio is defined as the ratio of the training set size to the number of human-based samples in that set. The baseline method, which involves duplicating real samples and adding Gaussian noise, is referred to as NSY. The results are shown in Fig. 7(a).

Given the inherent instability of AI training, we utilize the repeated k-fold cross-evaluation [38]. All human-based samples are partitioned into $k$ folds. For each fold and augmentation ratio, $M$ different subsets of synthetic samples are merged with $k-1$ folds for training, and the remaining fold is used for testing. It ensures that each fold is tested $M$ times with distinct sets of synthetic samples. The mean accuracy and standard deviation provide a more reliable metric.

At 100% augmentation, representing no augmentation, both methods achieve the identical accuracy of 73% with only human-based samples. As the augmentation ratio increases, noticeable accuracy improvements are observed for both methods. At 350% augmentation, MASS achieves an accuracy of 91%, which is 18% higher than no augmentation and 9% higher than NSY at its maximum accuracy of 82% at 450% augmentation. Importantly, MASS consistently outperforms NSY across all augmentation ratios, demonstrating higher accuracy and more stability as indicated by tighter standard deviations.

These outcomes suggest that MASS effectively synthesizes samples that capture essential features of human motion, offering more substantial benefits for training compared to simple noise perturbation.




(a) (b)

Fig. 8. Confusion matrices revealing distinct misclassification patterns of models trained solely on human-based versus synthesized data. (a) Confusion matrix for a model trained exclusively on human-based samples (0% replacement). (b) Confusion matrix for a model trained exclusively on synthesized samples (100% replacement).

### 4.3 Cross-Domain Sensing Performance

This section evaluates the capability of MASS to characterize varying sensing environments and facilitate cross-domain generalization. In cross-domain trials, the model is trained in one domain (source) and tested in another (destination), resulting in 12 distinct combinations derived from our four-domain dataset. For convenience, we denote training on domain $X$ and testing on domain $Y$ as $X \rightarrow Y$.

To illustrate the differences between domains, Fig. 7(b) depicts the cumulative distribution function (CDF) of Dynamic Time Warping (DTW) [39] distances for the same activity across domains. The figure reveals that inter-domain distances substantially exceed intra-domain distances. Specifically, $C1 - C2$ has a mean DTW distance of 8.92, close to the intra-domain mean distance of 8.61, indicating minimal variance due to minor differences in sensing positions and directions. In contrast, other inter-domain distances are considerably larger, signaling significant changes in multipath characteristics. Notably, room size and layout do not solely determine domain differences. $A - C1$ and $A - C2$ exhibit smaller DTW distances than $B - C1$ and $B - C2$.

Based on these DTW distance observations, we categorized the 12 cross-domain trials into two groups reflecting their difficulty. The **Minor Change Group** includes trials between domains with highly similar multipath characteristics, identified by low inter-domain DTW distances (*i.e.*, $C1 - C2$). The **Drastic Change Group** comprises all other

pairings, where domains have significantly different multipath profiles and thus exhibit high inter-domain DTW distances (*e.g.*, transfers between the anechoic chamber A and indoor rooms B or C). These groups reflect different levels of difficulty in cross-domain tasks. The Drastic Change Group is more challenging, whereas the Minor Change Group poses less difficulty.

During each trial, synthetic samples from the destination domain are added to enhance the training set, allowing the model to learn domain-specific features. We compare MASS with multiple reference and baseline approaches. Ref I represents an ideal, albeit unrealistic scenario where the training set includes samples from both the source and destination domains. Ref II trains solely on the source domain, reflecting real-world cross-domain sensing scenarios. NSY serves as the baseline by enhancing the source domain training set with Gaussian noise, a common practice in traditional data augmentation.

The results are summarized in Fig. 7(c). In the Drastic Change Group, MASS gives an average accuracy of 83%, outperforming Ref II by 22% and NSY by 13%. In the Minor Change Group, MASS achieves an accuracy of 84%, which is comparable to Ref I and Ref II. These results suggest that synthetic samples effectively capture domain-specific multipath characteristics, thereby validating the feasibility of MASS as a solution for cross-domain sensing.

### 4.4 Accuracy Comparison: Training on Real vs. Synthesized Samples

Although previous evaluations yield positive results, we remain curious about the extent to which MASS approximates human-based samples. To explore this, we assess whether synthetic samples can be distinguished from human-based ones. We replace different ratios of human-based samples with synthetic samples to train the LeNet. The model is then tested on an independent set of human-based samples. Similar to Section 4.2, the repeated k-fold cross-evaluation is used to ensure robust and reliable outcomes.

Fig. 7(d) illustrates the results. The peak at the replacement ratio 50% indicates the maximum diversity of the training set and therefore produces the maximum precision of 82%. Although the accuracy at 100% replacement is comparable to 0% replacement, the confusion matrices shown in Fig. 8(a) and 8(b) reveal discrepancies. Models frequently misclassify action directions (*i.e.*, approaching vs. departing) when trained with only synthetic samples, whereas models trained solely on human-based samples tend to misclassify action speed (*i.e.*, walking vs. running).

Based on the results, we find that MASS does not completely replicate the direct human-based sensing data, and real samples cannot be entirely replaced by synthetic ones. These differences highlight gaps between synthetic and real samples that warrant further investigation, as discussed in Sec. 3.4. However, we observe that substituting some real samples with synthetic ones enhances the model's performance. This effect can be linked to the greater diversity of data provided by synthetic samples and supports the findings in Sec. 4.2, where real samples were supplemented rather than substituted with synthetic ones.

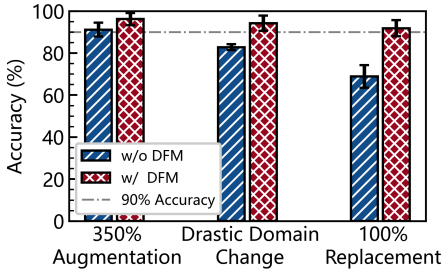



Fig. 9. Impact of the DFM on classification accuracy.

### 4.5 Accuracy Improvements with the DFM

While redesigning the metasurface could potentially eliminate the distortions mentioned, such an approach would incur significant costs and complexity. As a more practical alternative, we introduce the DFM, detailed in Sec. 3.5, which aims to mitigate these distortions through signal processing. To evaluate its efficacy, we reapply the DFM to our key test scenarios: augmentation, cross-domain sensing, and 100% replacement.

Fig. 9 clearly demonstrates the substantial benefits of the DFM. Applying the filter yields significant accuracy improvements across all scenarios compared to using unfiltered synthesized data. Most notably: (1) in the 350% Augmentation setting, accuracy improves by $\sim$ 5.0% (from 91.1% to 96.1%). (2) in the Drastic Domain Change scenario, DFM boosts accuracy by $\sim$ 11.4% (from 82.7% to 94.1%). (3) even in the challenging 100% Replacement scenario, reflecting data fidelity, DFM achieved a remarkable $\sim$ 23.0% gain (from 68.8% to 91.8%).

These results strongly indicate that the DFM effectively mitigates synthesis distortions. This improvement, achieved purely through data processing, substantially bridges the fidelity gap and boosts model performance, particularly in challenging cross-domain and data-scarce scenarios, thereby increasing the practical value of MASS without requiring hardware modifications.

## 5 Conclusion

This study introduces MASS, a novel metasurface-assisted approach for synthesizing Wi-Fi human sensing samples. By leveraging the waveform manipulation capabilities of metasurfaces, MASS effectively emulates human motion's impact on Wi-Fi signals while capturing environment-specific characteristics. Through comprehensive evaluation, including analysis of synthesis fidelity and targeted distortion mitigation, we demonstrate that MASS significantly enhances classification accuracy and cross-domain generalizability, and it is feasible to train a model exclusively on synthesized samples. Our findings validate the capability of metasurfaces to empower Wi-Fi sensing by enabling the generation of substantial, high-fidelity datasets with minimal cost and human involvement, paving the way for more robust and accessible sensing applications.

While practical challenges exist, including template diversity limitations from human-based collection and metasurface hardware constraints, our work shows these can be addressed through methods like DFM and exploring simulation-based templates. Despite these points, MASS

provides a valuable tool for generating high-quality, diverse Wi-Fi sensing datasets. Future work will focus on expanding the activity templates to include a wider range of human activities, validate the methodology in more environments, as well as extending the methodology to emulate the motion of non-human objects, thereby paving the way for more versatile and accessible Wi-Fi sensing applications.

## REFERENCES


[1] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," in *Proceedings of the ACM SIGCOMM*, 2015.

[2] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proceedings of the ACM MobiSys*, 2019.

[3] S. Lee, Y.-D. Park, Y.-J. Suh, and S. Jeon, "Design and implementation of monitoring system for breathing and heart rate pattern using wifi signals," in *Proceedings of the IEEE CCNC*, 2018.

[4] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Communications Surveys & Tutorials*, 2019.

[5] H. Yan, Y. Zhang, Y. Wang, and K. Xu, "Wiact: A passive wifi-based human activity recognition system," *IEEE Sensors Journal*, 2019.

[6] D. Wu, D. Zhang, C. Xu, H. Wang, and X. Li, "Device-free wifi human sensing: From pattern-based to model-based approaches," *IEEE Communications Magazine*, 2017.

[7] J. Yang, X. Chen, H. Zou, C. X. Lu, D. Wang, S. Sun, and L. Xie, "Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing," *Patterns*, 2023.

[8] C. Chen, G. Zhou, and Y. Lin, "Cross-domain wifi sensing with channel state information: A survey," *ACM Computing Surveys*, 2023.

[9] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, and Z. Wang, "Crosssense: Towards cross-site and large-scale wifi sensing," in *Proceedings of the ACM MobiCom*, 2018.

[10] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, "Data augmentation and dense-lstm for human activity recognition using wifi signal," *IEEE Internet of Things Journal*, 2021.

[11] W. Hou and C. Wu, "Rfboost: Understanding and boosting deep wifi sensing via physical data augmentation," *ACM IMWUT*, 2024.

[12] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI CAI*, 2020.

[13] X. Chen and X. Zhang, "Rf genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and generative diffusion models," in *Proceedings of the ACM SenSys*, 2023.

[14] G. Chi, Z. Yang, C. Wu, J. Xu, Y. Gao, Y. Liu, and T. X. Han, "Rf-diffusion: Radio signal generation via time-frequency diffusion," in *Proceedings of the ACM MobiCom*, 2024.

[15] X. Li, L. Chang, F. Song, J. Wang, X. Chen, Z. Tang, and Z. Wang, "Crossgr: Accurate and low-cost cross-target gesture recognition using wi-fi," *ACM IMWUT*, 2021.

[16] C. Xiao, D. Han, Y. Ma, and Z. Qin, "Csigan: Robust channel state information-based activity recognition with gans," *IEEE Internet of Things Journal*, 2019.

[17] C. Feng, X. Li, Y. Zhang, X. Wang, L. Chang, F. Wang, X. Zhang, and X. Chen, "Rflens: Metasurface-enabled beamforming for iot communication and sensing," in *Proceedings of the ACM MobiCom*, 2021.

[18] S. R. Wang, J. Y. Dai, J. C. Ke, Z. Y. Chen, Q. Y. Zhou, Z. J. Qi, Y. J. Lu, Y. Huang, M. K. Sun, Q. Cheng, and T. J. Cui, "Radar micro-doppler signature generation based on time-domain digital coding metasurface," *Advanced Science*, 2024.

[19] G.-B. Wu, J. Y. Dai, K. M. Shum, K. F. Chan, Q. Cheng, T. J. Cui, and C. H. Chan, "A universal metasurface antenna to manipulate all fundamental characteristics of electromagnetic waves," *Nature Communications*, 2023.

[20] S. Ahmad, M. Tariq, M. A. Jan, and H. Song, "Reconfigurable intelligent surfaces assisted 6g communications for internet of everything," *IEEE Internet of Things Journal*, 2023.

[21] A. Virmani and M. Shahzad, "Position and orientation agnostic gesture recognition using wifi," in *Proceedings of the ACM MobiSys*, 2017.

[22] S. Vishwakarma, W. Li, C. Tang, K. Woodbridge, R. Adve, and K. Chetty, "Simhumalator: An open-source end-to-end radar simulator for human activity recognition," *IEEE Aerospace and Electronic Systems Magazine*, 2022.

[23] S. Waqar and M. Pätzold, "A simulation-based framework for the design of human activity recognition systems using radar sensors," *IEEE Internet of Things Journal*, 2023.

[24] C. Xiao, Y. Han, W. Yang, Y. Hou, F. Shi, and K. Chetty, "Diffusion model-based contrastive learning for human activity recognition," *IEEE Internet of Things Journal*, 2024.

[25] B. Kang, Y. Yue, R. Lu, Z. Lin, Y. Zhao, K. Wang, G. Huang, and J. Feng, "How far is video generation from world model: A physical law perspective," *arXiv:2411.02385*, 2024.

[26] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, "Scientific machine learning through physics-informed neural networks: Where we are and what's next," *Journal of Scientific Computing*, 2022.

[27] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, W. Xu, and L. Su, "Towards environment independent device free human activity recognition," *ACM IMWUT*, 2018.

[28] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu, "Widar2.0: Passive human tracking with a single wi-fi link," in *Proceedings of the ACM MobiSys*, 2018.

[29] Y. Zhou, J. Yang, H. Huang, and L. Xie, "Adapose: Towards cross-site device-free human pose estimation with commodity wifi," *IEEE Internet of Things Journal*, 2024.

[30] Q. Bu, G. Yang, X. Ming, T. Zhang, J. Feng, and J. Zhang, "Deep transfer learning for gesture recognition with wifi signals," *Personal and Ubiquitous Computing*, 2022.

[31] S. J. Pan, V. W. Zheng, Q. Yang, and D. H. Hu, "Transfer learning for wifi-based indoor localization," in *Proceedings of the AAAI*, 2008.

[32] 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," *ACM IMWUT*, 2022.

[33] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.

[34] "pxie-5841." [Online]. Available: https://www.ni.com/en-us/shop/model/pxie-5841.html

[35] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Proceedings of the MIT Press NeurIPS*, 1989.

[36] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei, "Indotrack: Device-free indoor human tracking with commodity wi-fi," *ACM IMWUT*, 2017.

[37] "Picoscenes: Enabling modern wi-fi isac research!" [Online]. Available: https://ps.zpj.io/

[38] T. Fushiki, "Estimation of prediction error by using k-fold cross-validation," *Statistics and Computing*, 2011.

[39] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, 1978.