Practical Adversarial Attack on WiFi Sensing Through Unnoticeable Communication Packet Perturbation

Changming Li
Rutgers University
cl1361@scarletmail.rutgers.edu

Mingjing Xu
Temple University
mingjing.xu@temple.edu

Yicong Du
UESTC
202112081362@std.uestc.edu.cn

Limin Liu
UESTC
202221081035@std.uestc.edu.cn

Cong Shi
New Jersey Institute of Technology
cong.shi@njit.edu

Yan Wang
Temple University
y.wang@temple.edu

Hongbo Liu
UESTC
hongbo.liu@uestc.edu.cn

Yingying Chen
Rutgers University
yingche@scarletmail.rutgers.edu

ABSTRACT

The pervasive use of WiFi has driven the recent research in WiFi sensing, converting communication tech into sensing for applications such as activity recognition, user authentication, and vital sign monitoring. Despite the integration of deep learning into WiFi sensing systems, potential security vulnerabilities to adversarial attacks remain unexplored. This paper introduces the first physical attack focusing on deep learning-based WiFi sensing systems, demonstrating how adversaries can subtly manipulate WiFi packet preambles to affect channel state information (CSI), a critical feature in such systems, and thereby influence underlying deep learning models without disrupting regular communication. To realize the proposed attack in practical scenarios, we rigorously analyze and derive the intricate relationship between the pilot symbol and CSI. A novel mechanism is proposed to facilitate quantitive control of receiver-side CSI through minimal modifications to the pilot symbols of WiFi packets at the transmitter. We further develop a perturbation optimization method based on the Carlini & Wagner (CW) attack and a penalty-based training process to ensure the attack's universal efficacy across various CSI responses and noise. The physical attack is implemented and evaluated in two representative WiFi sensing systems (i.e., activity recognition and user authentication) with 35 participants over 3 months. Extensive experiments demonstrate the remarkable attack success rates of 90.47% and 83.83% for activity recognition and user authentication, respectively.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACM MobiCom '24, September 30-October 4, 2024, Washington D.C., DC, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0489-5/24/09…$15.00
https://doi.org/10.1145/3636534.3649367

CCS CONCEPTS

• **Security and privacy → Mobile and wireless security**.

KEYWORDS

Adversarial Attack, WiFi Sensing, Unnoticeable Attack, Communication Packet Perturbation

ACM Reference Format:
Changming Li, Mingjing Xu, Yicong Du, Limin Liu, Cong Shi, Yan Wang, Hongbo Liu, and Yingying Chen. 2024. Practical Adversarial Attack on WiFi Sensing Through Unnoticeable Communication Packet Perturbation. In *The 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '24), September 30-October 4, 2024, Washington D.C., DC, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3636534.3649367

1 INTRODUCTION

Being one of the most prevalent wireless communication techniques, WiFi has gained considerable attention in recent years due to its contactless sensing capabilities. Recent research has shown the great potential of WiFi sensing technology when integrated with deep learning methodologies, providing exceptional precision in various applications (e.g., activity recognition [16], user identification/authentication [37, 38], and vital sign monitoring [44]), providing the essential foundation for a more diverse range of applications, including smart cities, healthcare innovation, and security for homes and businesses.

While the advanced WiFi sensing technology empowered by deep learning offers the potential of immense convenience, it also creates opportunities for new attack vectors with severe consequences. For instance, adversaries could exploit these weaknesses to manipulate a smart entrance system

that relies on WiFi user authentication, potentially breaching the security of homes or offices. Additionally, adversaries could tamper with WiFi-based elderly care systems, causing them to wrongly categorize critical events like falls as regular activities such as watching TV. Furthermore, adversaries could control a WiFi-based vital sign monitoring system to generate falsified health data, leading to the omission of vital precautionary alerts and essential medical interventions. In this work, we demonstrate the feasibility of launching this attack in physical environments. We develop the first physical adversarial attack on classification-based WiFi sensing systems and uncover a surprising finding: *By physically tampering with the transmitted WiFi packets, adversaries can covertly manipulate WiFi sensing systems into generating predictions aligned with their adversarial goals.*

Our attack capitalizes on the predominant use of deep learning in WiFi sensing systems for predictions. We find that a few prior studies have conducted initial investigations into adversarial attacks on WiFi and mmWave sensing systems [4, 48, 50]. These studies focus on digital attacks that assume the received WiFi data can be directly modified. However, many receivers are users' devices (e.g., smartphones, wearables) that are hard to access. Instead of modifying the received WiFi data, it is more feasible if the adversary can physically alter the WiFi signals on the fly without accessing the receiver. A few more recent studies [22, 25, 57] show the potential of disrupting the wireless communication channel to interfere with WiFi sensing. Such disruptions degrade the communication quality (e.g., packet loss and signal-to-noise ratio), making the attacks noticeable to users. In this paper, we propose a more practical approach, an unnoticeable physical adversarial attack achievable through the transmitters in WiFi sensing systems (e.g., access points). Our work reveals that adversaries can covertly change the WiFi signals at the receiver by altering the preamble, a pre-defined sequence within the transmitted WiFi packets, which can potentially threaten any WiFi sensing systems.

Toward this end, we develop an imperceptible adversarial attack by optimizing the preamble, specifically the *pilot symbols* (or the long training sequence in IEEE 802.11 [35]), as illustrated in Figure 1. Our attack embodies two fundamental characteristics: *(i) Unnoticeability*: Our attack is unnoticeable to users as it has minimal impacts on the communication quality. Executed via WiFi transmitters, our attack enables remote control over WiFi sensing systems across all connected devices. *(ii) Untargeted & Targeted Attack:* Based on our theoretical exploration, we derive a quantitative relationship between pilot symbol modifications and the corresponding changes in received CSI, a measurement widely used in WiFi sensing. Building upon this understanding, we design an untargeted attack capable of disrupting WiFi sensing systems by inducing incorrect predictions. We further achieve the targeted attack, which enables the manipulation of WiFi sensing systems to provide adversary-specific predictions. The targeted attack allows adversaries to perform more complicated tasks, such as unauthorized access or triggering specific functions (e.g., turning on the oven). Both attacks raise serious security problems, particularly as WiFi sensing gains prominence within current WiFi infrastructures.

To realize such an attack, we need to address several critical challenges. Achieving effective attacks requires precise control over CSI. The relationship between the CSI responses and the pilot symbols can be modeled as a multiplicative relationship [26]. Our attack utilizes this relationship to generate adversarial perturbations, which can be precisely converted into pilot symbols and replayed with WiFi transmitters. We distinctively propose to generate optimal adversarial perturbations as multiplicative factors upon CSI data by adapting the Carlini & Wagner (C&W) attack [7]. In addition, enabling a universally effective attack in different scenarios with diverse channel variations presents a significant challenge. To overcome this problem, we optimize the pilot symbols across a small set of channel variations via penalty-based universal training [30]. This technique makes the generated pilot symbols highly effective across diverse new channel variations.

During the experiments of launching physical attacks with the adversarial pilot symbols, we notice that environment variations (e.g., interference of neighboring WiFi devices, furniture placement differences) could degrade the attack's effectiveness. To address this challenge, we design a channel augmentation technique that incorporates channel fading effects across different environments, improving the robustness of the attack. Furthermore, the inherent hardware imperfections induce unpredictable noises upon the received CSI. For example, voltage fluctuations in the circuit board may deviate WiFi signals passing the radio frequency front end or the local oscillator, potentially distorting the CSI patterns associated with the pilot symbols. To make the attack realistic on commercial WiFi devices, we fine-tune the pilot symbols by integrating synthesized hardware noises. We summarize our main contributions as follows:

- We demonstrate the first physical, unnoticeable, and universal adversarial attacks targeting deep learning-empowered WiFi sensing systems. By manipulating pilot symbols within WiFi packets, our attack enables control over the underlying models without disrupting regular communication.
- We model the multiplicative relationship between the pilot symbols in WiFi packets and received CSI to quantify how altering pilot symbols influences the CSI. Extensive experimental studies are conducted to confirm the effectiveness of perturbation generated based on this modeling.
- We developed pilot symbol optimization methods to enable both untargeted and targeted attacks by adapting the C&W

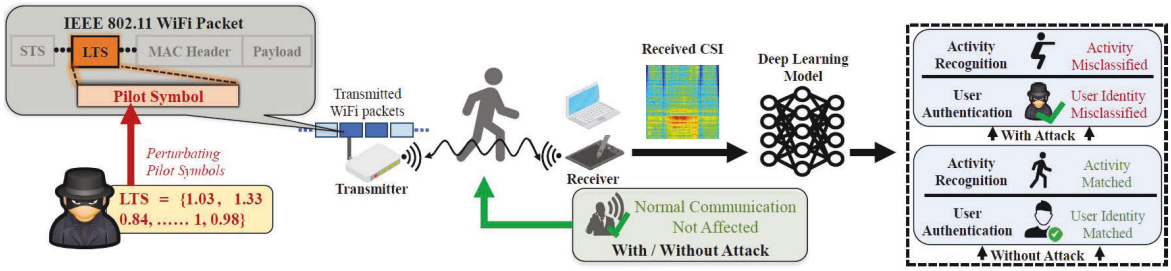


Figure 1: Illustration of the proposed physical, unnoticeable, and universal adversarial attack against deep-learning-based WiFi sensing system.

attack algorithm. We further design training techniques to generate pilot symbols robust across diverse channel variations and environments.

- We evaluate our attack by conducting extensive experiments with two representative WiFi sensing systems (i.e. activity recognition and user identification), involving 35 volunteers across three universities in a period of three months. The results show that our attack can achieve over 85% untargeted attack success rates and 85% targeted attack success rates.

2 BACKGROUND AND PRELIMINARIES

2.1 Wireless Sensing via Learning on Channel State Information

CSI-enabled WiFi Sensing. WiFi communication relies on orthogonal frequency division multiplexing (OFDM) techniques, where multiple data streams are simultaneously transmitted through a group of closely spaced narrow-band subchannels (subcarriers). For example, 802.11n 2.4GHz/5GHz WiFi normally operates on a 20MHz bandwidth divided into 52 subcarriers. Along with OFDM, CSI is used to represent how WiFi signals propagate at each subcarrier, which can be denoted as:

$$H = \begin{bmatrix} h_{1,1} & \cdots & h_{1,p} & \cdots & h_{1,P} \\ \vdots & & \vdots & & \vdots \\ h_{k,1} & \cdots & h_{k,p} & \cdots & h_{k,P} \\ \vdots & & \vdots & & \vdots \\ h_{52,1} & \cdots & h_{52,p} & \cdots & h_{52,P}, \end{bmatrix} \quad (1)$$

where k and p respectively represent the indices of the subcarrier and the WiFi packet, and $h_{k,p}$ is the complex coefficient representing the corresponding CSI. The CSI carries rich and detailed information that reflects dynamic and static characteristics of wireless channels, such as multipath fading, transmission loss, and Doppler effects. Therefore, it can be leveraged to sense the changes in the surrounding environment induced by the activities and behaviors of users. Compared to sensing via cameras and motion sensors, sensing with CSI reduces the risks of visual privacy as well as removing the requirement of wearing on-body sensors.

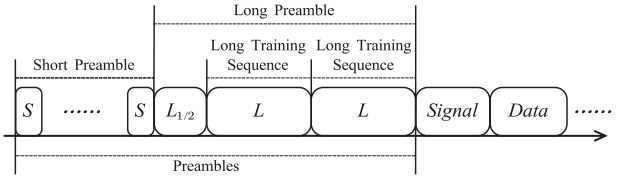


Figure 2: Illustration of the pilot symbol (also known as the long training sequence (LTS)) in the IEEE 802.11 b/g/n/ac frame structure [6].

Existing studies have shown that the CSI signals extracted from regular communication of WiFi can be utilized to achieve various sensing capabilities, including activity recognition [45], gesture recognition [33, 34] and user authentication [20]. These studies show that sensing capabilities can be seamlessly integrated into current WiFi communication systems, and they can work simultaneously.

Deep Learning-empowered Applications. To facilitate practical implementation, more and more research in WiFi sensing has started to adopt deep learning models to enhance accuracy and robustness. Unlike conventional machine learning models, deep learning models can establish intricate linear and nonlinear connections between input CSI data and resulting labels, guaranteeing excellent WiFi sensing performance. For instance, convolutional neural networks (CNNs) are frequently employed in activity recognition for feature extraction and classification [16, 37]. Beyond activity recognition, deep learning techniques have demonstrated considerable success in domains such as gesture recognition [33, 34], human presence detection [14], vital sign monitoring [44], and user authentication [37, 38]. In this study, we investigate security vulnerabilities within deep learning-empowered WiFi sensing systems in two representative domains: activity recognition and user identification. These WiFi sensing technologies are essential elements across a wide range of applications, such as smart entry systems, access controls for smart homes/offices, and systems for elderly care.

2.2 CSI Computation on WiFi Systems

WiFi systems conform to IEEE 802.11 b/g/n/ac computes the CSI based on the pilot symbols, also referred to as the long

training sequence (LTS), of each received WiFi packet. As shown in Figure 2, a WiFi packet contains four basic elements: a short preamble for onset detection, a long preamble (containing the pilot symbols), a signal field with basic packet information, and the data symbols containing the data. To estimate the CSI, the WiFi system examines the differences between the transmitted pilot symbols in the long preamble and the proportion of received signals corresponding to the pilot symbols. By denoting the transmitted signals of pilot symbols as x_L, we formulate the received signal as:

$$y_L = Hx_L + n, \tag{2}$$

where H is the physical channel and n denotes the noise. To estimate H representing the channel properties, channel estimation techniques could be employed, such as least square (LS), minimum mean square error, and linear minimum mean square error. Among them, LS is the most representative algorithm, which is widely employed in commercial WiFi systems. LS is designed to find a $\hat{H}$ that minimizes the sum of squares of errors between the y_L and x_L. It can be realized by solving the following estimation problem [35]:

$$\hat{H} = [(\overline{x}_L)^* \overline{x}_L]^{-1} (\overline{x}_L)^* y_L, \tag{3}$$

where $\hat{H}$ is the estimation of channel response. Note that the signal template $\overline{x}_L$ for CSI computation is pre-known to the receiver, and it will not be affected by the signal x_L.

3 THREAT MODEL

The adversary aims to transmit contaminated WiFi packets with the adversarial pilot symbols to control users' WiFi sensing systems. To achieve this goal, the adversary embeds the pilot contamination mechanism into the firmware of the WiFi transmitter (e.g., access points, routers) through firmware modifications. For example, The adversary can compromise WiFi routers deployed in public places (e.g., office rooms, university buildings, and hotel lobbies) via physical access and install the malicious firmware. In addition, the adversary can utilize online firmware update methods [11] to install the malicious firmware: (i) An adversary could send a malicious update with the pilot contamination mechanism through the Internet to compromise a WiFi router [27]. Firmware updates of many WiFi routers are not sufficiently protected by proper authentication; (ii) The adversary may also spread malicious firmware updates to compromise some routers even if they are protected by authentication (e.g., using the devices' factory default login information); (iii) Furthermore, as open-source WiFi platforms (e.g., OpenWRT [31], DD-WRT [12], Tomato Firmware [40]) are gaining popularity, the malicious firmware can be disguised as a customized online firmware update of these platforms with new functionalities. Users may install the firmware and unintentionally give control of their WiFi devices to the adversary.

To generate the adversarial pilot symbols, an adversary should have clean CSI data from public WiFi datasets or collect the data using his/her own WiFi receiver in the target environment. A similar strategy is adopted in the prior work RAFA [25]. In addition, an adversary can apply gradient-based adversarial machine learning algorithms to generate adversarial examples for two types: *(i) Untargeted attack* that aims to disable the WiFi sensing system by making the CSI at the receiver-side misclassified as an incorrect activity or identity label; *(ii) Targeted attack* that is designed to change the classification result to an adversary-desired activity/identity label. To optimize the pilot symbols of untargeted and targeted attacks, we consider the following two attack scenarios, each associated with a practical strategy for acquiring the model for optimization:

White-box Setting. The adversary utilizes a pre-trained deep learning model to mirror the user's model in architecture and weights. This assumption holds in real-world scenarios since numerous WiFi sensing systems are constructed upon pre-trained deep learning models, which are often accessible online (e.g., WiDar [33]). To generate the pilot symbols, the adversary employs gradient-descent-based optimization techniques on the pre-trained model.

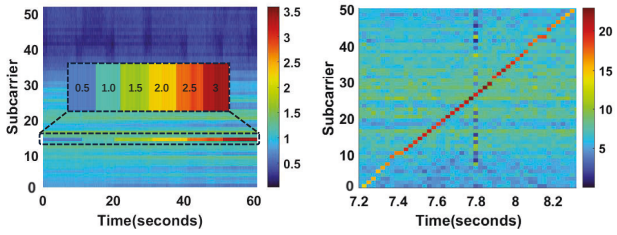
Black-box Setting. In cases where the user's pre-trained model is inaccessible, the adversary can harness the transferability property of deep learning models [24] to craft the adversarial pilot symbols. By exploiting the transferability, the adversary can train a surrogate model with similar classification objectives (e.g., activity or user identification) using their own model architectures and CSI datasets. Deep learning models exhibit similar inference capabilities despite differences in architectures or training datasets. So, the adversary can use the surrogate model to train the pilot symbols.

4 ATTACK OVERVIEW

4.1 Problem Formulation

We aim to craft the adversarial pilot symbols that force the deep learning model for WiFi sensing to produce adversary-specified labels. We represent the deep learning model as a function that takes a CSI sample as input and yields probabilities over a set of predefined labels: $z = f(\hat{H})$, where z denotes the label with the highest probability. The attack's goal is to generate contaminated pilot symbols resulting in perturbed CSI $\hat{H}'$ at the user's devices, compelling the model to produce a target label: $z_t = f(\hat{H}')$. The adversary must address the following practical constraints:

Maintaining Regular Communication. Contamination of the pilot symbols could induce undesired CSI perturbations H', which might degrade the WiFi communication quality on user devices. In this work, we quantify the communication quality with packet loss rate. The high loss rate will cause



(a) CSI responses (subcarrier 14) under different pilot symbol values (b) Manipulating responses of CSI across all 52 subcarriers

Figure 3: Illustration of CSI responses through manipulation of pilot symbols in the transmitted WiFi packet.

slow network speeds and frequent disconnects, which could alert users about potential attacks. Therefore, our attack needs to minimize its negative effects on packet loss rate to remain unnoticeable under regular communication.

Universal Applicability to Users' Interactions. In most real-world situations, the user's activity and the altered pilot symbols will influence the perturbed CSI H'. Given the unpredictability of the user's activity type in advance, the generated adversarial pilot symbols must be effective across diverse user activities.

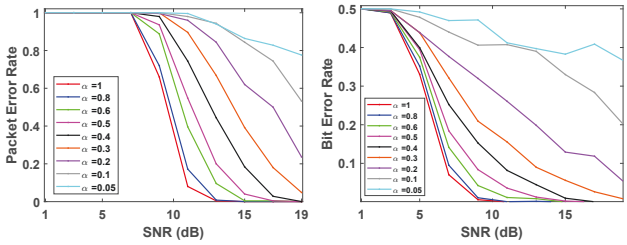
Robustness against Physical Distortions. To realize physical attacks with WiFi devices, the attack must be robust against environmental distortions due to neighboring signal interference and hardware internal noises (e.g., RF front ents, local oscillator).

4.2 Theoretical Analysis on Pilot Contamination and Verification

Theoretical Quantification of Pilot Symbol Modification on CSI. Our adversarial attacks rely on adding precise perturbations to the CSI. Consequently, it becomes imperative to devise a novel approach that guarantees concealed modifications at the subcarrier level of the CSI. Specifically, our attack controls the transmitted signal x_L by scaling the pilot symbol values at the transmitter. The pilot sequence in IEEE 802.11 WiFi packet contains K symbol values $L = [L_1, ..., L_k, ..., L_K]$, each corresponding to a short signal segment of the transmitted signal x_L. We denote the transmitted signals of the k^{th} subcarrier as $x_L^{(k)}$. By scaling the symbol value with a coefficient α, the strength of the transmitted signal segment will also be scaled with the same coefficient $\alpha x_L^{(k)}$. Considering the physical channel h is consistent, the signal of the pilot symbol received at the user's device can be modeled as being multiplied with the same coefficient, equation 2 can be revised as follows:

$$\alpha \cdot y_L^{(k)} \approx \alpha \cdot (h x_L^{(k)}) + n, \tag{4}$$

where the noise term n is usually ignored during CSI computation. Given the scaled received signal, the CSI estimated



(a) Packet Error Rate (b) Bit Error Rate

Figure 4: Simulation of impact pilot symbol with different values of coefficient α. By controlling α, we can reduce the negative impacts on the communication caused by pilot symbol contamination.

at the user' device can be reshaped based on LS method as:

$$\hat{h_k}' = \alpha \cdot \hat{h_k} = \alpha \cdot [(\overline{x}_L^{(k)})^* \overline{x}_L^{(k)}]^{-1} (\overline{x}_L^{(k)})^* y_L^{(k)}, \tag{5}$$

where $\hat{h_k}'$ and $\hat{h_k}$ are the estimated CSI before and after the pilot contamination. Note that the signal template $\overline{x}_L^{(k)}$ for CSI computation is pre-known to the receiver and will not be impacted by the signal scaling at the transmitter. The equation shows a *multiplicative* relationship between the modification to the pilot symbol L_k and the estimated CSI $\hat{h_k}'$, which will be used as the foundation of our attack.

During the attack, the generated adversarial perturbation is converted into an array of coefficient α as a group of contaminated pilot symbols. Once a malicious firmware is installed to compromise the WiFi transmitter, the adversary can inject contaminated pilot symbols into the transmitted WiFi packets. The contaminated pilot symbols are an integral part of the WiFi packet itself. Therefore, the synchronization between contaminated pilot symbols and the packet is not necessary. When these packets are received, the CSI computed from these affected packets is inherently influenced by contaminated pilot symbols.

Experimental Validation. We conduct experiments using two USRP devices to validate the multiplicative relationship between the pilot symbols and CSI changes. Specifically, we use a B210 as the transmitter and a B205min as the receiver. Both devices run GNU Radio and gr-ieee802-11 modules [5]. The CSI is computed at the receiver using the *WiFi_Decode_MAC* module at GNU Radio.

To examine the multiplicative relationship, we try to change the received CSI of subcarrier 14 (h_{14}) by scaling its symbol value L_{14} with α at the transmitter. Specifically, we vary α from 0.5 to 3 in 6 consecutive 10-second time slots, with a step of 0.5. The amplitude of the received CSI is shown in Figure 3(a). We can find that the CSI amplitude increases linearly with the changing α. To further examine the CSI amplitude changes induced by α, we compute the average CSI amplitude at each time slot. With an unaltered CSI amplitude of 1.1929, we observe that the CSI amplitude changes

to 0.65, 1.71, 2.35, 2.96, 3.59 when the α is set to 0.5, 1.5, 2, 2.5, 3, respectively. The relationship of CSI before and after the modification approximates the coefficient α, which is aligned with our theoretical derivation in Equation 5. We further study the feasibility of modifying CSI across different individual subcarriers. We program to take turns to switch a subcarrier for modification, where coefficient α is configured as 3 for each designated subcarrier to highlight the significance of channel manipulation. In Figure 3(b), the channel response extracted from the actual signal shows the change in CSI amplitude corresponding to each subcarrier over time. These results validate that the CSI amplitude can be multiplicatively manipulated based on the designated coefficient α, which can be leveraged to construct complex perturbations.

We further investigate how the constraint on modification coefficient α limits its negative impacts on regular communication. Particularly, we quantify the communication quality with packet error and bit error rates and study how they change under different α values. As shown in Figure 4, under the same signal-to-noise ratio (SNR), both packet error rate and bit error rate increase as α decreases. These observations suggest the value of α, which controls the magnitude of the modified pilot symbols, should be controlled to reduce the impact on communication. We study this problem and design an innovative method to limit the magnitude of the adversarial pilot symbols in our attack, which is explained in Section 5.3.

4.3 Attack Flow

We design a suite of techniques to realize the proposed physical, unnoticeable, and universal adversarial attack against two most representative applications in WiFi sensing, human activity recognition and user authentication. The attack flow is illustrated in Figure 5.

Activity-agnostic Universal Training. Based on the quantified multiplicative relationship between pilot symbols and CSI (validated in Section 4.2), we design an attack scheme that optimizes the perturbation as a multiplicative factor upon CSI ($x' = x \cdot \delta_m$). We adopt the Carlini & Wagner (C&W) attack scheme to craft the perturbation, which minimizes the magnitude of CSI modification as well as the untargeted/target adversarial loss. To ensure the perturbation is effective on various users' activities, we employ penalty-based universal training on a diverse set of CSI samples corresponding to different activities. Such a universal training strategy makes the perturbation agnostic to users' activities and identities.

Perturbation Robustness Enhancement. To ensure the effectiveness of physical attacks across varied conditions, we introduce a channel augmentation technique that integrates channel fading effects from diverse environments into the perturbation training process. Furthermore, to enhance the robustness of our perturbations against noise distortions, we incorporate hardware imperfections by simulating them with Gaussian noises during training.

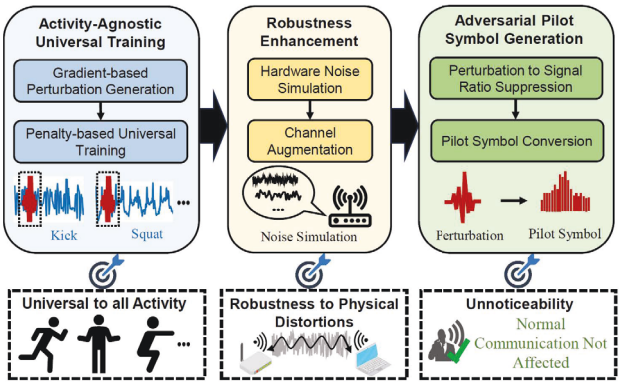


Figure 5: Flow of our physical adversarial attack targeting deep learning-empowered WiFi sensing systems.

Adversarial Pilot Symbol Generation. In our investigation in Section 4.2, we find that a low α may increase the packet loss, potentially alerting users. To avoid such disruptions, our attack suppresses such difference based on perturbation-to-signal ratio (PSR), which sets boundaries of the symbol value to limit the symbol distortions of the attack. Finally, our attack embeds the perturbations in terms of pilot symbol values of WiFi packets.

5 ATTACK DESIGN

5.1 Universal Perturbation Training

To ensure the viability of the proposed attack in real-world scenarios, it is essential to develop a universal perturbation capable of significantly influencing the received CSI. This perturbation should lead the user's model $f(\cdot)$ to incorrectly classify the received CSI, regardless of the initial CSI patterns (e.g., any CSI linked to the preserved activities within the model). In this study, we explore untargeted and targeted adversarial attacks. Given $(x, y) \in \mathcal{D}$, where $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ represents the training data comprising CSI samples x and corresponding ground-truth labels y, we aim to induce misclassification in the untargeted adversarial example such that $f(x + \delta) \neq y$. Conversely, the targeted adversarial example is designed to manipulate the model's output to a predetermined target label, with the objective of achieving $f(x + \delta) = y_t$, where y_t signifies the target label and δ stands for the universal perturbation. To achieve this, we frame the generation of perturbations for the untargeted and targeted attacks as:

$$\text{Untargeted Attack: } \arg\min_{\delta} \|\delta\|_p \text{ s.t. } f(x + \delta) \neq y, \quad (6)$$

$$\text{Targeted Attack: } \arg\min_{\delta} \|\delta\|_p \text{ s.t. } f(x + \delta) = y_t, \quad (7)$$

where $\|\cdot\|_p$ denotes the L_p norm. The optimization stops when it finds the optimal perturb adversarial sample x' with the minimal $\delta = x' - x$.

In this study, we employ a similar approach to the C&W attack [7] in order to solve the optimization problems outlined above. The C&W attack stands as a potent adversarial technique widely employed for untargeted and targeted misclassification. Compared to other adversarial attack methods (e.g., FGSM [15] and DeepFool [28]), the C&W attack demonstrates enhanced effectiveness and robustness against established defense mechanisms, including defensive distillation [32]. Based on the C&W attack, the optimization problem for the untargeted attack in Equation 6 is transformed into the following equation:

$$\text{minimize } -(c \cdot \mathcal{L}_u(x')) + \|\delta\|_p, \tag{8}$$

More specifically, for the untargeted attack, the optimization contains two terms: the first term encourages the model to misclassify the perturbed input x' into any class other than the original ground-truth label y, and the second term encourages the model to minimize the magnitude of the perturbation (i.e., less detectable). The untargeted attack loss function is shown in the following equation:

$$\mathcal{L}_u(x') = \max(Z(x')_y - \max_{i \neq y}\{Z(x')_i\}, -\xi), \tag{9}$$

where $Z(\cdot)$ is the logits (the output of the model $f(\cdot)$ before the softmax function), and ξ is a constant that controls the untargeted misclassification confidence.

The optimization problem for the targeted attack in Equation 7 is transformed into the following equation:

$$\text{minimize } c \cdot \mathcal{L}_t(x') + \|\delta\|_p. \tag{10}$$

Similarly, for the targeted attack, the optimization also contains two terms: the first term encourages the model to classify the perturbed input as the target label, and the second term encourages the model to minimize the magnitude of the perturbation. The targeted attack loss function is shown in the following equation:

$$\mathcal{L}_t(x') = \max(\max_{i \neq y_t}\{Z(x')_i\} - Z(x')_{y_t}, -\xi). \tag{11}$$

Note that we adopt L_2 norm and the best loss function mentioned in the original C&W attack scheme [7] to realize the $\mathcal{L}_u(\cdot)$ and $\mathcal{L}_t(\cdot)$.

Unlike the original C&W attack using $0 \leq x + \delta \leq 1$ as a "box constraint" [7], we generate perturbations within a positive and reasonable CSI amplitude range that can be applied to OFDM transmission using the following equation:

$$x' = \sigma \cdot \frac{1}{2}(\tanh(w) + 1), \tag{12}$$

where w is the variable to be optimized in $\tanh(\cdot)$, σ is the maximum value that we used to constrain the value of the perturbation in our attack model. In our algorithm. we initially generate x' using a given $w = \text{arctanh}(\frac{2x}{\sigma} - 1)$. Then, we employ Equation 8 or Equation 10 to update w and x' at each training epoch to iteratively find the optimal perturbation.

To further preserve the concealed nature of the generated universal adversarial perturbation, it is crucial to impose a constraint on its magnitude, ensuring it remains smaller than that of the WiFi signal. Specifically, we introduce the perturbation-to-signal ratio (*PSR*) as the following equation, which quantifies the relative magnitude of the perturbation with respect to the signal magnitude.

$$PSR = \frac{\|\delta\|_2}{\|\bar{x}\|_2}, \tag{13}$$

where $\bar{x}$ represents the average magnitude of WiFi frames from the legitimate transmitter. As the *PSR* should be set as low as possible to remain undetectable (i.e., $PSR \ll 1$), we adopt a threshold-based method to constrain the magnitude of the perturbation when searching for the optimal perturbation. In this work, we empirically determine a threshold $\tau_{PSR} = 0.05$, which constrains the magnitude of the perturbation to the minimum that can support high attack successful rates based on our experimental results.

5.2 Robustness Enhancement

CSI is often susceptible to diverse environmental changes and hardware-induced signal distortions [56]. Even slight noise interference can potentially alter our perturbation and disrupt its optimized pattern in practical environments. To address this challenge, we collect clean CSI samples (i.e., without any people in the same environment) at different times of a day, denoted as $\hat{n} = [n_1, n_2, ..., n_i]$. By adding these collected noises into the perturbation optimization process, the loss function will augment the robustness of the perturbation against various environment and hardware noises. Building upon this design, we showcase enhancing the robustness of targeted attack by revising Equation 10 as follows:

$$\text{minimize } c \cdot \mathcal{L}_t(\hat{x} + \delta) + \|\delta\|_p, \tag{14}$$

where $\hat{x}$ is equal to $x + rand(\hat{n})$. At each perturbation training epoch, the sample x will be affected by not only the perturbation but also the noise randomly selected from $\hat{n}$. Such a process will enhance the robustness of our attack against random signal variations in practical scenarios.

5.3 Adversarial Pilot Symbol Generation

The process of generating universal adversarial perturbations described earlier assumes the incorporation of perturbation δ into the benign sample through addition operations. However, our findings in Section 4.2 indicate that the alteration in CSI follows a linearly multiplicative pattern with the

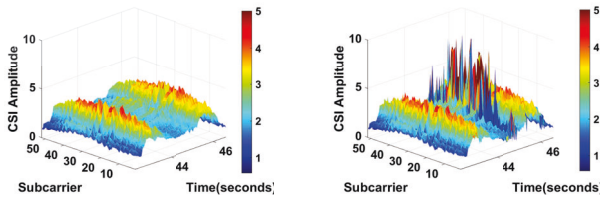
Algorithm 1: Targeted multiplicative universal adversarial perturbation generation.

Input: Input dataset $\mathcal{D}$, target class y_t, target model f, hyperparameters η.
Output: Multiplicative Universal Perturbation δ_m.

Initialize: Additive and Multiplicative Universal Perturbation δ, δ_m.
while ASR($\mathcal{D}$) $< \tau_{ASR}$ **do**
  for $(x, y) \in \mathcal{D}$ **do**
    Initialize $x \leftarrow \hat{x} \cdot \delta_m$, c_{upper}, c_{lower}, w and $c = c_{lower}$;
    for *numbers of epochs* **do**
      $x' = \sigma \cdot \frac{1}{2}(\tanh(w) + 1)$;
      $\delta = x' - x$;
      $\mathcal{L}_{total} \leftarrow \|\delta\|_2 + c \cdot \mathcal{L}_t(x')$;
      $w = w - \eta \cdot \nabla_w \mathcal{L}_{total}$;
      if $\mathcal{L}_t(x') > 0$ **then** $c_{lower} = c$;
      else $c_{upper} = c$;
      $c = (c_{upper} + c_{lower})/2$;
    end
    if $PSR > \tau_{PSR}$ **then** $\delta = \frac{\delta \cdot \tau_{PSR}}{PSR}$;
    Update $\delta_m \leftarrow (x + \delta)/x$;
    $\delta_m = max(min(\delta_m, \delta_m^{max}), \delta_m^{min})$;
    ASR($\mathcal{D}$) $= \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}(f(\hat{x}_i \cdot \delta_m) = y_t)$;
  end
end

coefficient α when applied to adversarial pilot symbols. As a result, it becomes essential to transform the universal adversarial perturbation generation represented in Equation 14 into a multiplicative format to effectively execute the attack on the physical channel. Specifically, we need to transform the generated perturbation δ to a multiplicative perturbation δ_m through $\delta_m \leftarrow (x + \delta)/x$.

In addition, it is essential to ensure that the values of δ_m maintain a reasonable magnitude, aligning with the amplitude distribution of real CSI samples. For the lower bound constraint of δ_m, we aim to retain the orthogonality of the pilot symbols defined by OFDM (i.e., keep the positive and negative signs of the pilot symbols unchanged). Given that the perturbation is applied to the pilot symbol through multiplication, we must ensure the universal adversarial perturbation is non-negative. For the upper bound constraint of δ_m, we aim to avoid excessive amplitude that may cause failure in decoding and a substantial number of bit errors. Such distortion in data communication can be easily detected. Therefore, we empirically set the lower and upper bounds of δ_m as $\delta_m^{min} = 0.5$ and $\delta_m^{max} = 3$, respectively. The value of δ_m will be adjusted to $max(min(\delta_m, \delta_m^{max}), \delta_m^{min})$ after determining the optimal δ_m. Figure 6 presents a comparison of the spectrograms of two CSI time series. One remains unaffected by the adversarial pilot symbol injection (Figure 6(a)), while the other reflects the impact of the adversarial pilot symbol introduced around 44s (Figure 6(b)). The results indicate that our method can successfully generate the adversarial perturbations transferable to pilot symbols of WiFi packets to control received CSI. The whole process of generating targeted multiplicative universal adversarial perturbations is depicted in Algorithm 1. Note that Algorithm 1 terminates when the attack success rate (ASR) on the perturbed dataset exceeds the preconfigured threshold τ_{ASR}.



(a) unaffected CSI (b) Affected CSI

Figure 6: A comparison of the spectrograms of two CSI time series unaffected and affected by the adversarial perturbation generated by our attack.

6 EVALUATION

6.1 Experiment Setup

Hardware Setup. We deploy two USRP devices—B210 serving as the transmitter and B205 as the receiver—utilizing the GNUradio platform [13] to broadcast an OFDM signal. Both devices are configured to operate on WiFi Channel 100, which spans from $5.49 \sim 5.51GHz$, using a central frequency of $5.5GHz$. The transmission power is set at -28.25dBm, equivalent to a normalized gain of 0.75 in the GNUradio setup. The packet transmission rate is set at 50Hz. Both transmitter and receiver are operated via laptops running Ubuntu 18.04. To filter out extraneous packets, we employ a MAC address filtering technique, ensuring only data from our transmitter is considered. The receiver then gathers CSI measurements across 52 subcarriers, which are used for dataset construction.

Training CSI Data Collection. We evaluate the proposed attack in three different environments: an apartment, a university office, and a balcony with dimensions of $6.2m \times 4.5m$, $5.3m \times 5.0m$, and $4.1m \times 6.0m$, respectively. Figure 7 illustrates the specific floorplans for these environments. Both the transmitter and receiver are positioned 2m apart horizontally and raised 1m above the ground. We enlist 35 volunteers who perform a total of six activities, chosen based on typical indoor actions: kicking, pushing, raising arms, sitting, squatting, and walking. Each of these activities takes place

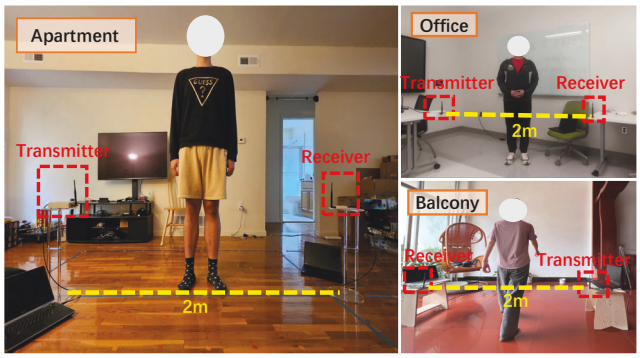


Figure 7: Experimental setup for physical attack in three different room environments.

in all three environments. Raw CSI data from the receiver is processed using a sliding window and short-time energy (STE) method [38] to segment activity-based CSI samples. In total, our dataset comprises 8,280 activity samples (performed by 15 volunteers) from the apartment, 8,715 samples (from another set of 15 volunteers) in the office, and 2,625 samples (conducted by 5 volunteers) from the balcony.

Training Deep Learning Models. We utilize the clean CSI data from volunteers to train deep learning models for two distinct CSI-based applications. For human activity recognition, we employ a CNN + GRU architecture consisting of three convolutional layers followed by an one-layer Gated Recurrent Unit (GRU) [52]. Each activity is linked with its corresponding CSI measurements, which are labeled as activity profiles. For user authentication, we adopt a CNN model as presented in existing literature [38]. Activities performed by specific users are labeled as user profiles. In both models, time-domain features are extracted from the CSI measurements, and they are trained using a supervised learning approach. The models demonstrate robust performance, with a classification accuracy of 96.57% for human activity recognition and 94.85% for user authentication.

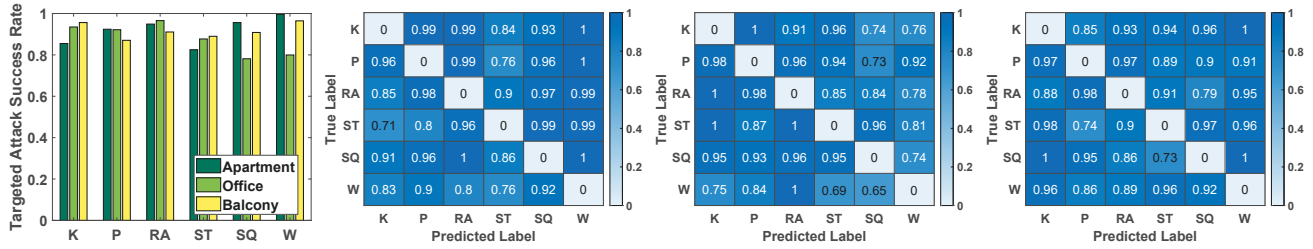
In this work, we define the baseline as the model's misclassification rate under the regular usage without attacks. The collected CSI data already contains random background and hardware noises, which distort the CSI pattern and cause misclassifications. For both untargeted and targeted attacks, we compare the attack success rates with the baseline without launching the attacks. For the untargeted attack, the attack is considered successful if the WiFi samples are misclassified (i.e., different from their ground-truth label). The percentages of misclassifications for activity recognition and user identification are 3% and 5% in our baseline, respectively. For the targeted attack, the success rate is defined as the percentage of WiFi samples classified as an adversary-specified target label. We take turns to set each activity/user label as the target label to evaluate our attack. The average success rates of baseline are 0.58% and 1.02% for activity recognition and user identification, respectively.

Attack Algorithm Evaluation. Our adversarial perturbation generation is implemented using the Tensorflow framework [1]. The digital perturbations were generated as detailed in Section 5. For both target models, we execute our attack algorithm to craft specific adversarial universal perturbations for each target label. To assess the real-world efficacy of our attack, we introduce the perturbation to the transmitter, broadcasting the perturbed OFDM signal. This signal was subsequently captured by the receiver while volunteers performed the six activities. The perturbed CSI samples were then extracted from the received packets to compile the adversarial testing datasets. In total, we collected 1600 perturbed CSI samples for testing human activity recognition and 1250 samples for user authentication. The data collection procedures were approved by our university's IRB.

Evaluation metrics. *1) Attack Success Rate:* For attacking model classification result, we define the two metrics to quantify the attack effectiveness: untargeted attack success rate (UASR) and targeted attack success rate (TASR). Untargeted attack success rate is the probability that a contaminated pilot symbol causes the model to produce false activity or user label. It can be calculated as $UASR = ACC - \frac{N_{correct}}{N_{total}}$, where ACC is the recognition accuracy of the target model; $N_{correct}$ is the number of perturbed CSI sample be correctly classified; N_{total} is the total number of perturbed CSI samples. Targeted Attack success rate is the probability that a contaminated pilot symbol causes the model to classify perturbed CSI sample to the any label desired by the adversary. It can be calculated as, $TASR = \frac{N_{target}}{N_{total}}$, where N_{target} is the number of perturbed CSI samples that are classified as the target activity or user. *2) Packet Loss Rate:* In addition, we use packet loss rate (PLR) to quantify the quality of data communication of the WiFi link. It can be calculated as $PLR = 1 - \frac{N_{received}}{N_{transmitted}}$, where $N_{received}$ is the number of packet received, $N_{transmitted}$ is the number of packet transmitted. The smaller the packet loss rate, the better the data communication quality.

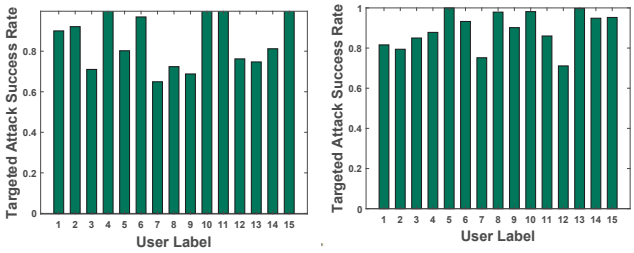
6.2 Human Activity Recognition

6.2.1 Physical Attack Performance. We first evaluate the effectiveness of our physical attack on human activity recognition. As depicted in Figure 8(a), the targeted attack success rate for human activity recognition is tested under three distinct environments: apartment, office, and balcony. Within each environment, we assess the attack's potency against 6 distinct activity classes. Our formulated attack registers an average success rate of 90.47% across all activity categories and throughout the three settings. This result significantly outperforms the performance of the targeted attack baseline, 0.58%. Notably, the office setting, though recording the



(a) Overall attack success rates (b) Confusion matrix for apartment (c) Confusion matrix for office (d) Confusion matrix for balcony

Figure 8: Targeted attack success rate for activity recognition: (K)ick, (P)ush, (R)aise (A)rm, (S)i(T), (SQ)uat, (W)alk.



(a) Apartment (b) Office

Figure 9: Targeted attack success rate for user authentication in different environments.

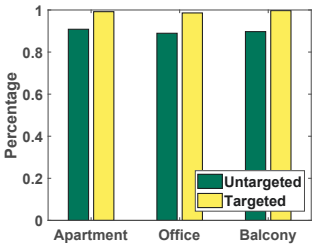


Figure 10: Digital attack success rate at human activity recognition.

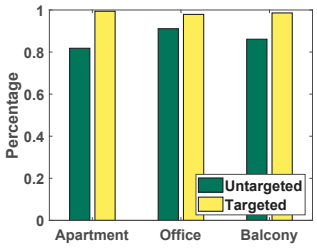


Figure 11: Digital attack results for user authentication.

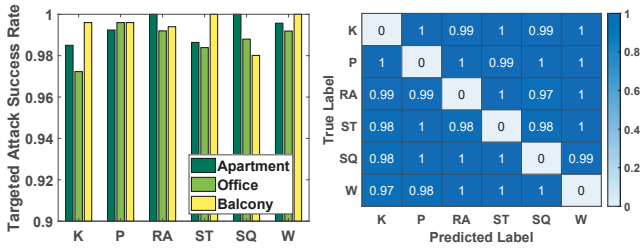
lowest average success rate, still achieves over 88%, underscoring the resilience of our attack across diverse environments. Detailed insights into the attack outcomes for each environment are visualized in Figures 8(b), 8(c), and 8(d). Intriguingly, in certain environments, like the apartment and balcony, success rates for some specific true-predicted combinations (such as squat-sit) dip marginally. This is attributed to the pronounced resemblance between the native and targeted CSI samples. Nevertheless, in the larger scheme, our attack consistently posts commendable success averages of 91.75%, 88.01%, and 91.67% for the three of environments.

6.2.2 Digital Attack Result. Figure 10 depicts the performance for human activity recognition on the digital side. In the scenario of an untargeted attack, the results show the effectiveness of our perturbation generation algorithm, achieving an overall 89.83% UASR across all three environments. This result shows that our attack is 83% better than the baseline when against WiFi-based activity recognition. With the perturbation's influence, a mere average of 6.4% of all perturbed samples get classified accurately, suggesting that a significant proportion of the samples have their classes altered from their original labels.

In the targeted attack setting, our approach consistently show high performance, nearing a 100% success rate in all three settings. Both Figure 12(a) and Figure 12(b) show the effectiveness of our formulated perturbation across the spectrum of the 6 activity categories. It's noteworthy to highlight that, percentage-wise, TASR outperforms UASR. Given our UASR definition—which hinges on the differential between the pristine model's recognition accuracy and the proportion of correct classifications. It is plausible for the UASR to generally lag behind TASR, especially when the recognition accuracy of the model reaches around 96%.

6.3 User Authentication

6.3.1 Physical Attack Result. We subsequently evaluate our physical attack within the context of the user authentication. Figure 9(a) and Figure 9(b) show the targeted attack success rate on user authentication across different user identities. Our attack achieves an average TASR of 83.83% throughout all three settings. This result is 82% better than the targeted attack baseline. In particular, for the apartment and office settings, our approach yields TASRs of 84.56% and 89.03% across the respective 15-user groups. Meanwhile, in the balcony environment, a TASR of 77.9% is accomplished across the 5 user labels. Note that some user labels exhibit relatively low TASRs. For instance, in the apartment environment, the targeted attack outcome for user label 7 barely achieves a TASR of slightly over 60%. This discrepancy can be attributed mainly to the fact that certain user labels inherently share similar features, displaying limited variability in attributes like body shape, height, and weight as manifested in CSI. This overlap renders the attack less effective for specific individuals. Nonetheless, the attacks targeting the majority of user labels consistently register TASRs equal to or exceeding 80% across all tested environments.



(a) Attack success rate (b) Confusion matrix

Figure 12: Digital targeted attack for human activity recognition involving (K)ick, (P)ush, (R)aise (A)rm, (S)i(T), (SQ)uat, (W)alk.

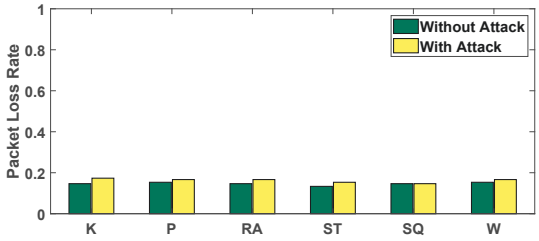


Figure 13: Comparison of packet loss rate before and after launching our physical attack.

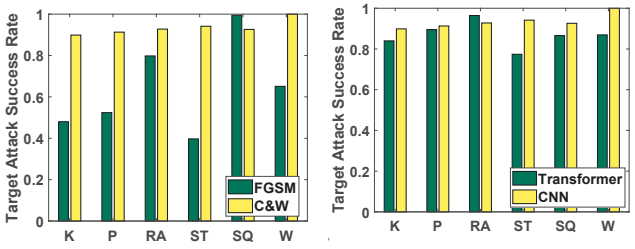
6.3.2 Digital Attack Result. As shown in Figure 11, both the untargeted and targeted attack success rates attain averages of 86.33% UASR and 98.63% TASR, respectively. Compared with the baseline,our attack can achieve 81% higher UASR when against WiFi-based user authentication. The generated perturbation effectively causes the expected misclassifications.

6.4 Impacts on Communication Quality

To assess the unnoticeability of our proposed attack, we conduct a case study comparing the average packet loss rate in scenarios both with and without the attack over a fixed signal transmission duration. Given the packet transmission rate of 50Hz as described in our hardware setup, we would theoretically expect to receive 5000 packets over 100 seconds. We determine the PLR for various activities based on the count of packets lost during this transmission. As illustrated in Figure 13, the average PLR stands at 14.24% in the absence of an attack and slightly rises to 16.23% during an attack as CSI is perturbed while users are engaged in their activities. This minimal increase in PLR indicates the subtlety of our attack. Such consistent communication quality underscores that our approach not only ensures effective adversarial outcomes but also remains discreet in the context of data communication, reflected by the low packet loss rate.

6.5 Impacts of Attack Algorithms

To investigate the effects of varying attack algorithms, we conducted experiments on human activity recognition with



(a) FGSM and C&W algorithm (b) Transformer and CNN models

Figure 14: Comparisons of physical attack performance with different attack algorithms and deep learning models.

a group of five participants in an apartment setting. Both FGSM [15] and C&W algorithms were adapted to our perturbation generation workflow. Figure 14(a) illustrates that while the C&W algorithm yields a TASR of 93.45%, the FGSM yields a mean TASR of 64.08% across all six activities.

It's noteworthy that, with the exception of the SQ (squat) activity, TASRs derived from the FGSM algorithm consistently lag behind those procured via the C&W algorithm. Although the FGSM-generated perturbation for the squat activity boasts a TASR of 99.5%, such high-efficacy perturbations are rarely produced by FGSM. This can be attributed to the fact that the FGSM algorithm wasn't originally crafted for targeted adversarial attacks, whereas the C&W method is more nuanced and aims at specific labels. Furthermore, FGSM's susceptibility to environmental nuances and its vulnerability to minor channel fluctuations become evident even when channel augmentation strategies are in play.

Nevertheless, FGSM's ability to surpass a TASR of 60% accentuates the adaptability of our attack to a broader spectrum of extant adversarial attack techniques, further emphasizing its efficacy in perturbation generation.

6.6 Impacts of Deep Learning Models

To assess the viability of our proposed attack against various existing deep learning architectures, we compare its effectiveness on two target models: the Transformer and our CNN. The evaluation focuses on the human activity recognition task, involving a consistent group of 5 volunteers from the apartment environment.

Figure 14(b) reveals that both the CNN and Transformer models consistently achieve TASRs greater than 85% for the majority of activities. More specifically, the attacks yield average TASRs of 93.45% for the CNN model and 86.81% for the Transformer model. It's also worth noting that, in the absence of any adversarial attacks, both clean models maintain classification accuracies for human activity recognition beyond 95%. These results collectively underscore the versatility of our attack, highlighting its potential for targeting an array of deep learning architectures.

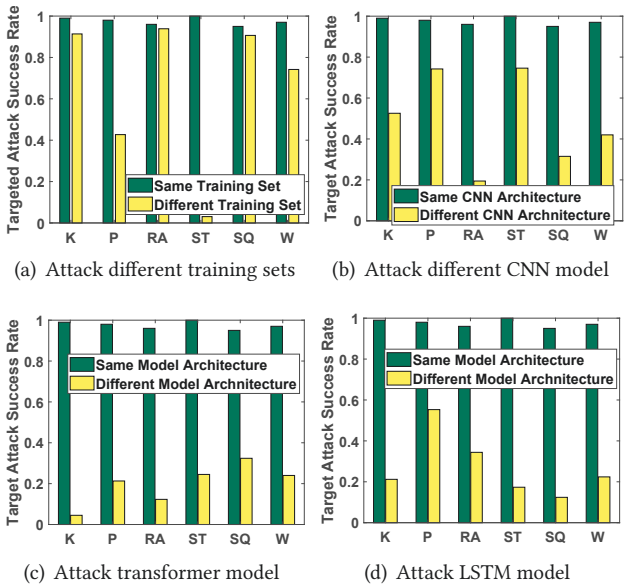


Figure 15: Black-box attacks on human activity recognition under different settings.

6.7 Black-box Attack

To further investigate the practicability of our attack, we perform transfer-based black-box attack experiments under two cases. In the first case, we use the same CNN model to assess the impact of varying training sets on TASR during black-box attacks. Specifically, we generate a perturbation using a training set from a group of 8 individuals and direct the attack towards a model trained with a dataset from 2 different individuals. In the second case, we examine the impact of disparate model architectures on TASR under black-box conditions. This case includes scenarios where the adversary, unfamiliar with the exact architecture of the target model, uses the initial model for his training set. We consider three types of different model architectures: CNN, transformer and LSTM. CNN, as a fundamental building block in diverse WiFi sensing tasks, such as activity recognition, user authentication, and gesture recognition, is necessary to be considered in black-box attack. Even through transformer and LSTM are rarely used for WiFi sensing applications, they are investigated in prior studies of WiFi sensing [9, 54]. Therefore, we also need to take them into the consideration. Firstly, we employ a CNN with 3 convolutional layers to train the dataset and then initiate an attack on another CNN model consisting of 2 convolutional layers. Then, we utilize the same adversarial perturbation, which is trained by the CNN with 3 convolutional layers, to attack a transformer model and an LSTM model separately. The results of these experiments are presented in Figures 15(a), 15(b), 15(c) and 15(d). The results reveal that the average TASR for black-box attacks employing different training sets stands at 65.98%. Meanwhile, black-box attacks using the CNN with 3 convolutional layers against the CNN with 2 convolutional layers achieve an average TASR of 49.05%. In the case of attacking different model architectures, the average TASR achieves 20.01% and 27.17% when against the transformer and LSTM model, respectively. While these figures represent a drop from the 98% TASR observed in prior white-box attacks, these cases still surpass random guesses, which are statistically expected to yield a mere 18%.

7 RELATED WORK

WiFi Sensing Systems. Existing CSI-based sensing systems are mainly focused on activity and gesture recognition, user authentication, and localization. Regarding activity recognition, Nakamura et al. [29] designed a CSI-based fall detection system using CNN with high accuracy across environments. Shalaby et al. [36] developed deep learning models on CSI amplitude data for recognizing different types of activities. Wang et al. [42] enabled accurate human authentication through a few-shot learning without requiring a large number of CSI data. Liu et al. [23] proposed tracking human vital signs via the detection of CSI variations induced by breathing and heartbeats. WiFi sensing has also been utilized for user authentication tasks. For example, AR-Alarm [20] performs real-time intrusion detection using CSI on commodity WiFi devices. Furthermore, CSI-based sensing can be extended to realize indoor locations [18].

Adversarial Attacks. Adversarial attacks aim to add imperceptible perturbations to original inputs, causing targeted deep learning models to yield incorrect decisions. Particularly, deep learning models are inherently susceptible to such attacks in various applications. For example, in the image and audio domains, most existing studies launched digital adversarial attacks on video [7, 46, 47] and speech/speaker recognition models [3, 10, 19, 30, 39, 41, 49], which feed adversarial examples directly into the target model without considering effects in real physical environments. Some studies proposed over-the-air audio attacks [2, 8, 21, 51, 53], focusing on modeling sound distortions during audio propagation. With the proliferation of deep learning in wireless sensing, wireless sensing systems also have security vulnerability to adversarial attacks [4, 17, 43, 48, 50]. For instance, Wang et al. [43] showed the threat of digital adversarial attacks through the manipulation of CSI data in a DNN-based indoor localization system. WiCAM. [50] investigates a digital adversarial perturbation generation method to attack a WiFi-based activity recognition system. However, these studies manipulate digital WiFi data directly to compromise the deep learning model at the receiver side without considering the practicability of establishing these attacks in

real-world scenarios, in which receiver devices (e.g., laptops, smartphones) are hard to access.

Recent studies have designed adversarial attacks using RF signals to alter the decisions of deep learning models. For example, WiAdv.[57] investigated adversarial signals against the gesture recognition system by modifying the Doppler shift of CSI with adversarial perturbations. Similarly, Liu et al. [22] exploited a physical adversarial example as a jamming signal to disrupt the behavior recognition system. Furthermore, RAFA [25], which launches a physical adversarial attack on a WiFi-based localization system, requires the attack signal to be sustained for a duration exceeding that of the preamble to be effective. However, these signal-based attacks disrupt the wireless communication channel to inject the adversarial perturbations. They compromise the packet's payload, leading to noticeable packet loss rates or data decoding errors for users. Such errors or delays can impact the communication quality. In this paper, we show a practical attack compatible with commercial WiFi devices while having a negligible impact on WiFi communication quality.

8 DISCUSSION

Enhancement of Black-box Attack. For more practical black-box attack scenarios, it is hard for the adversary to acquire specific labels, model architecture, and parameters for the target classification system. To enhance the transferability of black-box attacks, we could train adversarial examples across diverse datasets and model architectures. Besides, the adversary usually can only access part of the training data for the target model. To address the lack of information about the deep learning model, generative adversarial networks (GANs) could be harnessed to synthesize sufficient adversarial samples with the original training data. Overall, these explorations have the potential to enhance the practicality of black-box attacks in real environments.

Extending the Applicability and Robustness of Attacks. Several potential areas can be explored in future work. First, since our research did not fully utilize the highly sensitive CSI phase data for human activity recognition, we could expand the utilization of CSI to incorporate phase information within CSI, making it adaptable to a broader spectrum of WiFi-based sensing systems. In addition, existing studies in the image domain show that adversarial attacks have the capability to attack image classification classifiers (i.e., ImageNet) for over 1000 classes [55]. Our attack techniques can be further generalized to models with more labels. To enhance the robustness of perturbations in real physical environments, our proposed channel augmentation technique could be further incorporated with simulated physical distortions from ambient noise using specialized channel models. These research avenues could provide valuable insights into improving the applicability and robustness of attack methods across various wireless sensing systems.

Countermeasures. Our proposed attack allows the adversary to generate perturbations that are correctly decoded and hardly perceived during wireless transmission through the establishment of specific symbol value thresholds. Therefore, a potential countermeasure is to detect adversarial examples by analyzing differences between clean and perturbed CSI data. Particularly, the alternations in pilot symbols in our attack would introduce some fluctuations in the CSI phase data. This could be due to the changes of I and Q components, a complex sample that is defined by each half of the long training symbols. A value changes to a specific subcarrier also emits magnitude modification to either I or Q component and further changes the complex number which induces the phase shift. Such a phase shift may vary based on different modulation modes in wireless transmission and serve as a potential indicator for adversarial example detection. This detection can be conducted by examining the statistical characteristics or building a machine learning classifier (e.g., SVM). For instance, statistical measures such as mean, variance, and skewness could help discern perturbed CSI data associated with specific activities or users.

9 CONCLUSION

In this work, we investigate the security issues of WiFi sensing systems through designing a physical, unnoticeable, and universal adversarial attack. Our investigations revealed that by tampering with the pilot symbols in transmitted WiFi packets, adversaries can covertly manipulate the predictions of deep learning models without disrupting regular communication. Through a rigorous theoretical analysis, we quantified the multiplicative relationship between the influence of pilot modifications at the transmitter and the receiver-side CSI. Based on this quantification, we designed an adapted Carlini & Wagner attack scheme that optimizes adversarial pilot symbols as multiplicative factors upon CSI, realizing both untargeted and targeted attacks. In addition, we developed a penalty-based universal training approach that optimize robust pilot symbols independent to users' activities and identities. Furthermore, various channel interference and hardware noises are considered to improve the robustness of the pilot symbols under physical attacks. Extensive experiments against both activity recognition and user identification models under various realistic scenarios demonstrated the attack's effectiveness.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation Grants CNS2120396, CCF2211163, IIS2311596, CNS2120276, CNS2145389, IIS2311597.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 265–283.

[2] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin RB Butler, and Joseph Wilson. 2019. Practical hidden voice attacks against speech and speaker recognition systems. *arXiv preprint arXiv:1904.05734* (2019).

[3] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. 2018. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554* (2018).

[4] Harshit Ambalkar, Xuyu Wang, and Shiwen Mao. 2021. Adversarial human activity recognition using Wi-Fi CSI. In *2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 1–5.

[5] Bastian Bloessl. 2019. *A physical layer experimentation framework for automotive WLAN*. Gesellschaft für Informatik eV.

[6] Bastian Bloessl, Michele Segata, Christoph Sommer, and Falko Dressler. 2017. Performance assessment of IEEE 802.11 p with an open source SDR-based prototype. *IEEE Transactions on Mobile Computing* 17, 5 (2017), 1162–1175.

[7] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. Ieee, 39–57.

[8] Guangke Chen, Sen Chenb, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. 2021. Who is real bob? adversarial attacks on speaker recognition systems. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 694–711.

[9] Zhenghua Chen, Le Zhang, Chaoyang Jiang, Zhiguang Cao, and Wei Cui. 2018. WiFi CSI based passive human activity recognition using attention based BLSTM. *IEEE Transactions on Mobile Computing* 18, 11 (2018), 2714–2724.

[10] Moustapha M Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. 2017. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. *Advances in neural information processing systems* 30 (2017).

[11] Ang Cui, Michael Costello, and Salvatore Stolfo. 2013. When firmware modifications attack: A case study of embedded exploitation. (2013).

[12] DD-WRT. accessed February 2012. https://dd-wrt.com/

[13] GNU Radio Website. accessed February 2012. http://www.gnuradio.org

[14] Liangyi Gong, Wu Yang, Zimu Zhou, Dapeng Man, Haibin Cai, Xiancun Zhou, and Zheng Yang. 2016. An adaptive wireless passive human detection via fine-grained physical layer information. *Ad Hoc Networks* 38 (2016), 38–50.

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

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

[17] Silvija Kokalj-Filipovic, Rob Miller, and Joshua Morman. 2019. Targeted adversarial examples against RF deep classifiers. In *Proceedings of the ACM Workshop on Wireless Security and Machine Learning*. 6–11.

[18] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. 2015. Spotfi: Decimeter level localization using wifi. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. 269–282.

[19] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. 2018. Fooling end-to-end speaker verification with adversarial examples. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1962–1966.

[20] Shengjie Li, Xiang Li, Kai Niu, Hao Wang, Yue Zhang, and Daqing Zhang. 2017. Ar-alarm: An adaptive and robust intrusion detection system leveraging csi from commodity wi-fi. In *Enhanced Quality of Life and Smart Living: 15th International Conference, ICOST 2017, Paris, France, August 29-31, 2017, Proceedings 15*. Springer, 211–223.

[21] Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. 2020. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*. 1121–1134.

[22] Jianwei Liu, Yinghui He, Chaowei Xiao, Jinsong Han, Le Cheng, and Kui Ren. 2022. Physical-World Attack towards WiFi-based Behavior Recognition. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 400–409.

[23] Jian Liu, Yan Wang, Yingying Chen, Jie Yang, Xu Chen, and Jerry Cheng. 2015. Tracking vital signs during sleep leveraging off-the-shelf wifi. In *Proceedings of the 16th ACM international symposium on mobile ad hoc networking and computing*. 267–276.

[24] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770* (2016).

[25] Zikun Liu, Changming Xu, Emerson Sie, Gagandeep Singh, and Deepak Vasisht. 2023. Exploring Practical Vulnerabilities of Machine Learning-based Wireless Systems. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 1801–1817.

[26] Yongsen Ma, Gang Zhou, and Shuangquan Wang. 2019. WiFi sensing with channel state information: A survey. *ACM Computing Surveys (CSUR)* 52, 3 (2019), 1–36.

[27] mandylionlabs. 2013. *Reverse Engineering a D-Link Backdoor – /dev/ttyS0*. https://devttys0.com/2013/10/reverse-engineering-a-d-link-backdoor/

[28] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2574–2582.

[29] Takashi Nakamura, Mondher Bouazizi, Kohei Yamamoto, and Tomoaki Ohtsuki. 2020. Wi-fi-CSI-based fall detection by spectrogram analysis with CNN. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 1–6.

[30] Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. 2019. Universal adversarial perturbations for speech recognition systems. *arXiv preprint arXiv:1905.03828* (2019).

[31] OpenWRT. accessed February 2012. https://openwrt.org/

[32] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*. IEEE, 582–597.

[33] Kun Qian, Chenshu Wu, Zheng Yang, Yunhao Liu, and Kyle Jamieson. 2017. Widar: Decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 1–10.

[34] Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. 2018. Widar2. 0: Passive human tracking with a single Wi-Fi link. In *Proceedings of the 16th annual international conference on mobile systems, applications, and services*. 350–361.

[35] Anand Kumar Sah and Arun Kumar Timalsina. 2015. Improvement of Complexity and Performance of Least Square Based Channel Estimation in MIMO System. *Journal of Advanced College of Engineering and*

Management 1 (2015), 11–24.

[36] Eman Shalaby, Nada ElShennawy, and Amany Sarhan. 2022. Utilizing deep learning models in CSI-based human activity recognition. *Neural Computing and Applications* (2022), 1–18.

[37] Cong Shi, Jian Liu, Nick Borodinov, Bruno Leao, and Yingying Chen. 2020. Towards environment-independent behavior-based user authentication using wifi. In *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 666–674.

[38] Cong Shi, Jian Liu, Hongbo Liu, and Yingying Chen. 2017. Smart user authentication through actuation of daily activities leveraging WiFi-enabled IoT. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 1–10.

[39] Cong Shi, Tianfang Zhang, Zhuohang Li, Huy Phan, Tianming Zhao, Yan Wang, Jian Liu, Bo Yuan, and Yingying Chen. 2022. Audio-domain position-independent backdoor attack via unnoticeable triggers. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 583–595.

[40] Tomato Firmware. accessed February 2012. https://www.polarcloud.com/tomato

[41] Jon Vadillo and Roberto Santana. 2019. Universal adversarial examples in speech command classification. *arXiv preprint arXiv:1911.10182* (2019).

[42] Dazhuo Wang, Jianfei Yang, Wei Cui, Lihua Xie, and Sumei Sun. 2022. CAUTION: A Robust WiFi-based human authentication system via few-shot open-set recognition. *IEEE Internet of Things Journal* 9, 18 (2022), 17323–17333.

[43] Xiangyu Wang, Xuyu Wang, Shiwen Mao, Jian Zhang, Senthilkumar CG Periaswamy, and Justin Patton. 2022. Adversarial deep learning for indoor localization with channel state information tensors. *IEEE internet of things journal* 9, 19 (2022), 18182–18194.

[44] Xuyu Wang, Chao Yang, and Shiwen Mao. 2017. PhaseBeat: Exploiting CSI phase data for vital sign monitoring with commodity WiFi devices. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1230–1239.

[45] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. 2014. E-eyes: Device-free location-oriented activity identification using fine-grained WiFi signatures. In *Annual International Conference on Mobile Computing and Networking (MobiCom)*. 617–628.

[46] Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. 2022. Cross-Modal Transferable Adversarial Attacks from Images to Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15064–15073.

[47] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. 2018. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610* (2018).

[48] Yucheng Xie, Ruizhe Jiang, Xiaonan Guo, Yan Wang, Jerry Cheng, and Yingying Chen. 2022. Universal targeted attacks against mmWave-based human activity recognition system. In *IEEE Conference on Computer Communications (INFOCOM)*. 541–542.

[49] Yi Xie, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan. 2021. Real-time, robust and adaptive universal adversarial attacks against speaker recognition systems. *Journal of Signal Processing Systems* (2021), 1–14.

[50] Leiyang Xu, Xiaolong Zheng, Xiangyuan Li, Yucheng Zhang, Liang Liu, and Huadong Ma. 2022. WiCAM: Imperceptible Adversarial Attack on Deep Learning based WiFi Sensing. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 10–18.

[51] Hiromu Yakura and Jun Sakuma. 2018. Robust audio adversarial example for a physical attack. *arXiv preprint arXiv:1810.11793* (2018).

[52] Jianfei Yang, Xinyan Chen, Han Zou, Chris Xiaoxuan Lu, Dazhuo Wang, Sumei Sun, and Lihua Xie. 2023. SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing. *Patterns* 4, 3 (2023).

[53] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A Gunter. 2018. {CommanderSong}: A systematic approach for practical adversarial voice recognition. In *USENIX security symposium (USENIX security)*. 49–64.

[54] Zhongfeng Zhang, Hongxin Du, Seungwon Choi, and Sung Ho Cho. 2022. TIPS: Transformer based indoor positioning system using both CSI and DoA of WiFi signal. *IEEE Access* 10 (2022), 111363–111376.

[55] Zekun Zhang and Tianfu Wu. 2020. Learning ordered top-k adversarial attacks via adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 776–777.

[56] Shengli Zhou and Georgios B Giannakis. 2004. Adaptive modulation for multiantenna transmissions with channel mean feedback. *IEEE Transactions on Wireless Communications* 3, 5 (2004), 1626–1636.

[57] Yuxuan Zhou, Huangxun Chen, Chenyu Huang, and Qian Zhang. 2022. WiAdv: Practical and Robust Adversarial Attack against WiFi-based Gesture Recognition System. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–25.