

Universal Targeted Adversarial Attacks Against mmWave-Based Human Activity Recognition



Yucheng Xie, Xiaonan Guo, Yan Wang, Jerry Cheng, and Yingying Chen

1 Introduction

Millimeter Wave (mmWave) technology is one of the promising communication technologies due to its high throughput and wide bandwidth. Recent studies have shown the initial success of using mmWave in the domain of sensing applications, including Human Activity Recognition (HAR). Human activity recognition (HAR) has attracted significant attention since it is an essential technology to enable human-computer interactions in many Internet of Things (IoT) and security applications, including health monitoring and user authentication. Many HAR systems have been designed using various sensing modalities. Traditional camera-based [11, 16] and sensor-based [5, 42] HAR systems capture human activities using video cameras and body sensors, respectively. They usually intrigue privacy concerns or are not convenient. To circumvent these limitations, low-cost, non-intrusive solutions like

Y. Xie

Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA

e-mail: yx11@iupui.edu

X. Guo (✉)

George Mason University, Fairfax, VA, USA

e-mail: xguo8@gmu.edu

Y. Wang (✉)

Temple University, Philadelphia, PA, USA

e-mail: y.wang@temple.edu

J. Cheng

New York Institute of Technology, Old Westbury, NY, USA

e-mail: jcheng18@nyit.edu

Y. Chen (✉)

Rutgers University, New Brunswick, NJ, USA

e-mail: yingche@scarletmail.rutgers.edu

radio frequency-based techniques are being researched. Wireless signals have been commonly used in communication due to their convenience, flexibility, and ability to transmit information over long distances without the need for physical connections. The prevalence of wireless signals in our everyday devices allows for a complex network of reflected rays in indoor environments. Researchers find that the human presence and motion significantly impact these signals, enabling the capture of human body movements involved in daily activities. Recently, wireless signals (e.g., WiFi [19, 73], sound [31, 62], mmWave [39, 63]) have been utilized to track human activities without attaching sensors to the human body. In this direction, mmWave-based HAR systems stand out because they can provide high resolution with their short wavelength and large bandwidths.

Most mmWave-based HAR systems adopt deep learning models for activity identification due to their high accuracy and strong capability of handling interference in the real world. However, recent research has revealed that deep learning models are susceptible to adversarial inputs [68]. Some researchers have designed minor perturbations that cause deep learning networks to make inaccurate predictions in image classification [44] and voice recognition [53]. Nevertheless, few studies have investigated the susceptibility of adversarial targeted attacks in mmWave-based HAR systems. Because mmWave-based HAR systems are usually integrated into many crucial applications such as older patient monitoring and user authentication [6, 91] as shown in Fig. 1, we believe that studying adversarial attacks on these systems is critical and urgent. Most recently, Ozbulak et al. [48] have done an initial investigation with the untargeted adversarial attack on mmWave-based HAR. The designed attack is only applicable to a particular HAR model (i.e., heatmap-based) and cannot trigger the model to generate designated classes. Moreover, many research problems, such as how to design unnoticeable perturbations based on unique patterns of mmWave signals [66], how to launch universal target adversarial attacks [32], or more challenging black-box attacks [34], are still worth further

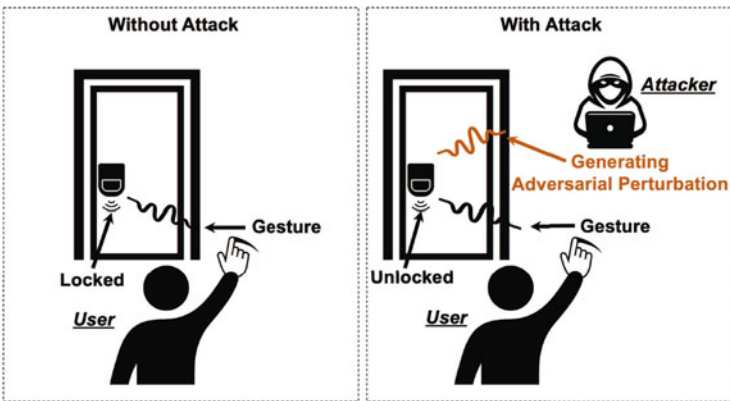


Fig. 1 Illustration of adversarial attacks on gesture-based user authentication system

exploration. Therefore, a more comprehensive study of systematically exploring different types of adversarial attacks on different types of mmWave-based HAR models is highly demanded.

In this chapter, we aim to systematically investigate and reveal the severe security issues of mmWave-based HAR models by developing the following effective adversarial attacks: (1) *White-box Attack*. Unlike existing work that only studied the untargeted attack for a particular mmWave-based HAR model, we successfully design both targeted and untargeted attacks for different mmWave-based HAR models in the white-box scenario. Moreover, because both targeted and untargeted attacks need to train a unique adversarial perturbation for each activity sample [87], which is inefficient and infeasible in time-constrained scenarios, we design a universal adversarial attack that can produce an adversarial perturbation applicable to different activity samples and ready to be used in real-time without additional training; (2) *Black-box Attack*. Besides white-box attacks that assume the attackers have full knowledge of the victim model, we further explore more challenging black-box attacks where attackers may not have sufficient information about the victim system and need to conduct attacks under more realistic conditions (e.g., the victim model is unavailable to the attacker). Black-box attacks are a more probable form of threat in real-world applications. By examining black-box attacks, we hope to better prepare these HAR systems for actual threats. Therefore, we develop an effective method to enable black-box targeted attacks in this chapter. Exploring both white-box and black-box adversarial attacks permits us to comprehensively and practically evaluate and improve the security of mmWave-based HAR.

Designing effective and practical adversarial attacks for different mmWave-based HAR models is nontrivial. Different from traditional replay attacks [51], our attack could fool the HAR system without collecting data samples from the target activity. In particular, we apply gradient-based machine learning algorithms to generate adversarial perturbations for targeted and untargeted attacks while minimizing their size. The adversarial perturbation is generated by solving an optimization problem to concurrently minimize the perturbation loss, which constrains the perturbation size and adversarial loss to ensure the success of the adversarial attacks without being noticed. In addition, mmWave-based HAR systems may use different data representations that require careful attention. Our comprehensive study identifies two representative types of mmWave-based HAR models (i.e., voxel-based and heatmap-based). We design a discretization method to ensure the validity of adversarial samples and further optimize the form of the adversarial samples with two distance metrics. The main challenge in designing the universal adversarial attack is deriving an effective adversarial perturbation for any activity sample without online training. We implement an offline training strategy with an iteration algorithm that crafts universal perturbation across the samples from a small pre-collected activity set. Unlike the existing universal attack that needs inserting padding frames between two successive activities [48], our attack modifies the activity sample directly, which enables the attack on a broader range of mmWave-based HAR applications. Furthermore, to overcome the information deficiency of the victim model in black-box attacks, we utilize a knowledge distillation (KD) approach to generate a robust

replacement model. We further develop a generative adversarial network (GAN) to produce a sufficiently large number of pseudo samples for substitute model construction.

In summary, we explore the security issues of AI-empowered mmWave-based HAR systems as deep learning technologies have been extensively used to help these systems to achieve more accurate and convenient recognition. In particular, we implement a comprehensive assessment of the challenges brought by adversarial attacks on various mmWave-based HAR systems, including both white-box and black-box adversarial attacks. For white-box attacks, we employ adversarial learning to reduce the magnitude of the perturbation, ensuring that the generated perturbation is undetectable by manual examinations while can successfully attack mmWave-based HAR systems. We also develop a discretization method to enable adversarial attacks on different representative models of mmWave-base HAR. To enable universal targeted attacks, we develop an iteration method to construct well-designed universal perturbations that can be applied to various unseen mmWave samples directly without additional training for these samples. We further design a black-box attack that can attack mmWave HAR systems without knowing the model architecture and parameters. In particular, we leverage knowledge distillation to address the information deficiency of the victim model. We also develop a generative adversarial network to address the lack of training data. We assess our implemented attack methods on two representative mmWave-based HAR models and demonstrate the efficacy, efficiency, and practicality of the attacks.

The remainder of this chapter is organized as follows. Section 2 discusses the related work in the field of HAR and adversarial attacks. Section 3 provides background information on sensing using wireless signals, human activity recognition, and adversarial attacks. Section 4 summarizes the victim machine learning models for HAR. Section 5 presents the threat model of our attack. Section 6 describes the developed white-box adversarial attack design and black-box attack design. Section 7 presents the experimental results of the attacks on two different mmWave-based HAR models. Finally, Sect. 8 concludes this chapter and provides directions for future research.

2 Related Work

Because of its wide application, HAR has attracted great attention for the past decade. In general, human activity recognition systems can be classified into three categories: camera-based [28], sensor-based [4], and radio-frequency (RF) signal-based [19, 71]. A couple of camera-based systems have been implemented to recognize human activities [16, 28]. These works use cameras to capture images or videos and apply image-processing algorithms to extract motions. However, camera-based methods may raise privacy concerns. To address this weakness,

sensor-based systems have been developed [4, 20]. These works explore various dedicated sensors such as gyroscopes [20], ECG or FSR sensors [4] to collect different types of signals for further analysis. However, sensor-based approaches require users to wear sensors or other devices, which is inconvenient for senior people or during complex activities. To overcome the above limitations, researchers recently developed RF-based methods (e.g., WiFi and mmWave). WiFi-based approaches [19, 72, 75, 89] use off-the-shelf WiFi devices to infer human activities. However, being easily influenced by surrounding environments remains the main limitation. Compared with WiFi signals, mmWave has been proven to be robust for activity recognition due to the antenna's directionality and stability. Some researchers design HAR systems based on mmWave [6, 25, 39, 50, 55, 57, 58, 71, 77–79, 84].

Most mmWave-based HAR systems adopt deep learning models for activity identification due to their high performance and capability of handling real-life interference. However, machine learning models such as neural networks were susceptible to adversarial perturbations, as pointed out by Szegedy et al. [68]. We discover that the majority of current adversarial attacks are proven in applications related to image recognition and speech authentication [10, 14, 32, 43, 44, 81]. Recently, there has been some work discussing adversarial attacks on radar-based systems. Yang et al. [87] examine the adversarial susceptibility of the Doppler-based HAR system. They analyze the untargeted attack issues for the HAR system and evaluated three white-box attack methods (i.e., FGSM, PGD, and MIM), respectively. Then, Ozbulak et al. [48] examine the vulnerability of radar-based HAR systems to a universal untargeted attack. Nevertheless, none of them explore the feasibility of targeted adversarial attacks to control the HAR system's output, nor do they provide a comprehensive study of adversarial attacks against mmWave-based HAR systems. Moreover, since Ozbulak's method only targets one heatmap-based HAR model, how to launch a universal targeted attack on other types of mmWave-based HAR models is unknown. Besides, based on unique patterns of mmWave activity data, how to develop adversarial activity samples to assure their validity and make them unnoticeable is necessary but seldom explored. In addition, how to enhance attack performance in more challenging black-box scenarios is still an open problem.

In contrast to previous research, we implement a comprehensive study of the threats brought by adversarial attacks, including both untargeted and targeted attacks. We broaden our study on both heatmap-based and voxel-based mmWave-based HAR systems. By optimizing perturbation based on the unique patterns of mmWave activity data, inventing universal attacks to make our attack approach more efficient, and examining the robustness of attacks under black-box scenarios, we intend to give a complete examination of the challenges posed by adversarial attacks on mmWave-based HAR systems.

3 Background

In this section, we present essential background information on three key topics: sensing using wireless signals, human activity recognition, and adversarial attacks.

3.1 Sensing Using Wireless Signals

Wireless signals have become widely adopted for communication due to their convenience, adaptability, and capacity to transmit data over long distances without the need for physical connections. The prevalence of wireless signals in our everyday electronic devices enables complex networks of reflected rays in indoor environments, making it possible to track human motions by examining the received signals. The fundamental premise of sensing using wireless signals is that human movement affects wireless signal propagation. This breakthrough has led to various applications, including human activity recognition [35, 36, 82, 83, 86], human localization [37, 59, 75] and vital sign monitoring [23, 33]. Compared with traditional camera-based methods, which may raise privacy concerns or sensor-based approaches that require users to wear sensors or other devices, wireless-signal-based methods are more convenient, especially for seniors or during complex scenarios. Among the prominent approaches are RFID sensing, WiFi sensing, and mmWave sensing. These techniques offer distinct advantages in terms of operating frequencies, detection capabilities, and potential applications.

The rise of RFID sensing in smart living spaces is due to its hands-free detection abilities with operating frequencies from 125 kHz to 1 GHz. RFID detection considers the changes in waveforms emitted by a transmission antenna. Compared to echo-based detection, RFID transponders can handle higher frequencies, allowing for more extensive range detection. Unique identification is possible through micro-Doppler signatures generated when objects interact with RFID signals. RFID systems, either active or passive, require no line-of-sight and have long lifespans. Thus, they are commonly used in inventory management in warehouses [9, 13, 22, 23, 47, 56].

Owing to the ubiquity of WiFi infrastructure, WiFi detection has become a popular device-free sensing technique. Its Received Signal Strength (RSS) can provide environmental and bodily state information similar to echo-based detection. The RSS, which measures path loss in Decibels (dB), can indicate the presence of people and their activities through its sensitivity to environmental changes. By placing two WiFi devices across a space and observing signal changes, we can effectively track health aspects like breathing and heart rates, and indoor positioning [1, 26, 85]. However, the coarse nature of RSS data, coupled with unstable WiFi signal strength, raises concerns about its reliability in applications. For a more precise use of WiFi signals, Channel State Information (CSI) is developed. CSI measures path loss and also incorporates multipath effects like scattering and

fading. Its higher sensitivity enables better detection of human movements, even unconscious ones like breathing [27, 38, 41, 52, 72, 74]. CSI provides a more detailed set of values, including amplitude and phase information for orthogonal frequency division multiplexing (OFDM) subcarriers, which can capture different frequency ranges and thus provide more fine-grained wireless channel details than RSS. Hence, healthcare applications are increasingly adopting CSI processing using common WiFi devices.

Millimeter wave (mmWave) detection operates at high frequencies, between tens and hundreds of GHz, offering broader bandwidth due to less congestion from commercial technologies such as TV or radio broadcasts. It's utilized in next-generation WiFi protocols like WiGig or 60 GHz WiFi [88]. As mmWave signals can effectively penetrate materials like plastic, drywall, and clothing, they are sensitive to minor changes like vital signs and are useful for high-resolution monitoring, making them ideal for device-free human activity recognition [60]. Studies [30, 63, 76, 90] have demonstrated its potential in recognizing various human activities, even with sparse data. Commercial radar hardware can be used for human activity recognition, dynamic skeleton pose tracking, and even differentiating between a person's motion and static states, including specific activities during static states such as making a call, using an app, or keeping the phone in the pocket.

3.2 *Adversarial Attack*

As Deep Learning (DL) becomes more prevalent, it is anticipated that potential threats will emerge, particularly against DL models. Adversarial attack, also known as evasion attacks, occurs when an adversary makes minute modification to the input of a neural network in order to cause an error in the inference process [17, 43]. These modifications are not noise samples generated at random; rather, they have been purposefully engineered to form a vector in the input feature space that can trick the DL model. Typical examples of these kinds of attacks entail either solving a limited optimization problem in order to construct a deceptive vector in the input feature space or inserting a small value in the model's gradient direction relative to the inputs. Both of these methods involve manipulating the output of the model. These attacks are typically stealthier and more energy-efficient than traditional jamming attacks [45] as they only need to make a small modification over a short period of time to confuse the DL models in their decision-making.

In the disciplines of Computer Vision (CV) and Natural Language Processing (NLP), adversarial attacks have been the subject of much research [10, 14, 32, 43, 44, 81]. One of the most well-known examples is when a machine learning classifier was fooled into thinking that an image of a panda was actually of a gibbon by adding a perturbation to the panda image that was deliberately created to trick the classifier [17]. Similarly, adversarial attacks in the field of wireless sensing can lead to the incorrect classification of human activities. Despite this, we have found that very few studies have looked into the possibility of adversarial targeted attacks occurring

in mmWave-based HAR systems. This motivates us to systematically study and uncover the significant security vulnerabilities of mmWave-based HAR models by building effective adversarial attacks.

3.3 *Human Activity Recognition*

The field of Human Activity Recognition (HAR) is rapidly growing, benefiting from advancements in sensors, cost efficiency, energy optimization, real-time data computation, machine learning, computer vision, AI, and IoT technologies. Activity classifications span diverse domains such as locomotion [18, 29], transportation [12, 46], mobile phone use [15], entertainment [80], health [24, 65], gestures [2, 75], and security [54]. HAR's criticality is emphasized by its application across health monitoring, fitness tracking, home automation, augmented reality, traffic management, targeted advertising, and security. For instance, individual activity logs can inform tailored dietary advice by estimating calorie consumption, and fall detection in seniors can trigger immediate emergency response, mitigating severe accidents.

Conventionally, machine learning methods have been employed to recognize human activity. However, these conventional methods for HAR necessitate the creation and selection of pertinent features, a process requiring significant human effort and expert knowledge. Additionally, the performance of these features might not always be optimal. To mitigate the need for hand-engineering features, deep learning techniques have been introduced in recent years [67]. These techniques offer several advantages for HAR. Firstly, they eliminate the need for manually designing features, which usually require specialist knowledge. Secondly, they have demonstrated higher accuracy in HAR than traditional methods [92]. Thirdly, they can learn from unlabeled data, which is a significant advantage for HAR due to the impracticality of gathering substantial volumes of labeled activity data. Lastly, deep learning techniques can learn useful features from raw data and can handle activity-related data from a range of sources, including different individuals, device models, and device orientations.

A typical deep learning-based HAR system has four main components as shown in Fig. 2. The first component involves data collection from various sensors that could capture images, WiFi CSI, accelerations, gyroscope readings, barometer readings, sound, biosensor readings, etc. The second component involves preprocessing the collected data using techniques like scaling, Principal Component Analysis (PCA) whitening, Zero-phase Component Analysis (ZCA) whitening, or denoising. The third component is the model-building stage, where diverse deep models (e.g., RBM, autoencoder, RNN) can be utilized to learn relevant features. This is followed by the application of a classifier (like softmax classifier, SVM) at the top layer. Once the model is established, it can be trained using the input data. During this training phase, network parameters such as weights are optimized. Finally, the trained model can be employed to predict the activity based on incoming data.

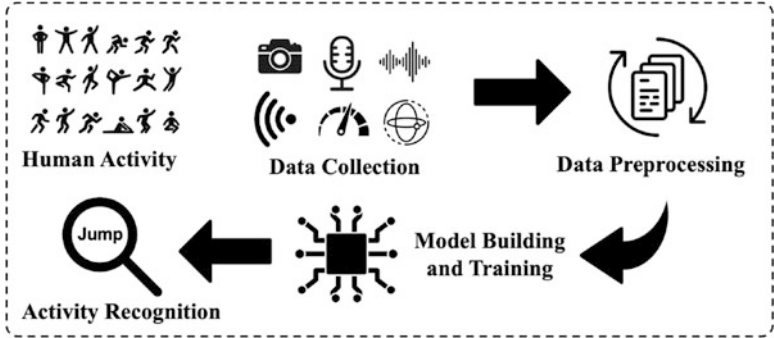


Fig. 2 A typical flow of human activity recognition system

4 Victim Machine Learning Models

The main goal of the mmWave-based human activity recognition system is to identify actions or gestures by examining the dynamics of mmWave signals [6, 39, 57]. As shown in Fig. 2, a typical mmWave-based HAR system captures mmWave signals reflected from the human body via the mmWave sensor. It performs signal processing to determine activity characteristics (e.g., velocity or posture) of users and then estimates the activity class using deep learning models. Deep learning model-empowered HAR systems are vulnerable to adversarial attacks. Existing mmWave-based HAR systems can be categorized into two classes based on the representations of the received mmWave signals. One of the representations is the point cloud derived from the received mmWave signals via a series of FFT operations (i.e., Range-FFT, Doppler-FFT, and Angle-FFT). Each point in the point clouds presents the x , y , and z coordinates of a mmWave signal reflected from the human body [6, 63, 78, 84, 91], which allows mmWave radars to generate a rough contour of the human body. However, point clouds are incompatible with neural network architecture as the number of points varies over time. Prior mmWave-based HAR research usually adopts voxelization to transform the point cloud into a constant amount of voxels [6, 63] for HAR. The other representation is the heatmap of the object-related information (e.g., distance, velocity, angle, and energy) extracted from the received mmWave signals. Many mmWave-based HAR systems have leveraged the heatmap to identify human activities (e.g., doppler-range map [39], micro-Doppler map [77], spatial spectrograms [57], spatial feature map [61], and projection heatmap [58]) because it is easy to achieve a good accuracy by applying pre-trained neural network models from the image domain to mmWave-base sensing.

In this chapter, we investigate the attacks on two typical mmWave-based HAR models using different types of representations, which brings more challenges to design a generic attack method because of their significant differences.

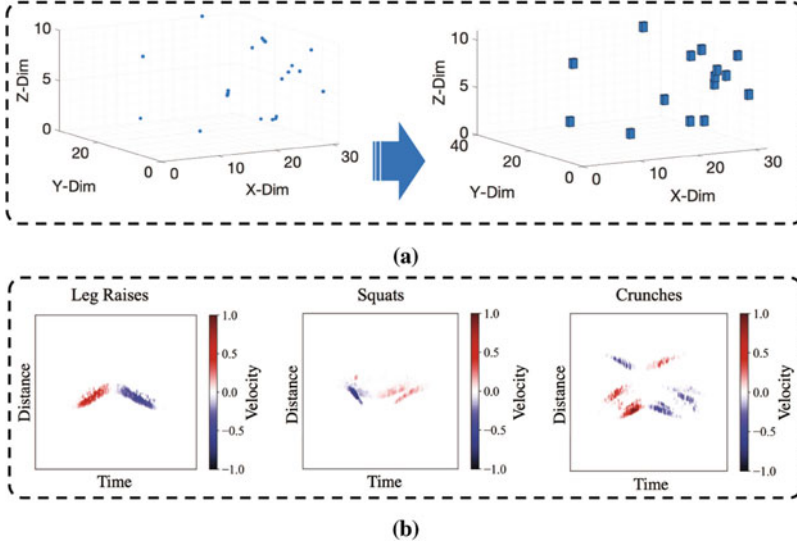


Fig. 3 Two typical data representations for mmWave-based HAR. (a) Voxels generation from the point cloud; (b) Spatial-Temporal heatmaps of three different activities

Voxel-Based Machine Learning Model We choose an existing mmWave-based HAR system [63] as a representative to study the vulnerability of voxel-based HAR models to adversarial attacks. This model has been utilized as a benchmark in numerous subsequent publications [3, 49]. In particular, the point clouds data is subjected to voxelization to address the non-uniformity issue in each frame, as shown in Fig. 3a. After the voxelization, the point clouds of each frame is transformed into a set of voxels in a three-dimensional space. A voxel is defined as $[x, y, z, v]$, where x, y, z are the spatial position of the voxel and v is the number of cloud points in the cube-shaped voxel with a designated size. Each activity sample is defined as t sets of voxels, where t is the time dimension. As for the machine learning model, they employ a Time-distributed CNN plus Bi-directional LSTM model. This model consists of 3 time-distributed convolutional layers followed by a bidirectional LSTM layer and an output layer. This model is directly trained on the input sample, which includes its temporal and spatial dimensions.

Heatmap-Based Machine Learning Model In addition to the voxel-based mmWave-based HAR system, we devise a heatmap-based HAR system to study its vulnerability to adversarial attacks. Similar to state-of-the-art mmWave-based HAR methods (e.g., [39]), we first derive the Doppler-range map of the users' activity by calculating Range-FFT and Doppler-FFT. Then, we generate heatmaps by accumulating the velocity of every distance in every denoised Doppler-range map together. Next, we normalize the derived velocity information and present the velocity-distance relationship in the time dimension. In this way, we transfer the

original instantaneous velocity-distance relationship to a more comprehensive spatial-temporal heatmap, which describes the process of a whole activity as shown in Fig. 3b. We utilize a CNN model for activity classification. In particular, this model consists of 3 convolutional layers, each followed by a max-pooling layer. A 64-dimensional feature map is created after 3 rounds of upsampling and downsampling. The feature map is then condensed into a one-dimensional array by integrating a flattened layer.

5 Threat Model

Adversarial attacks can be categorized into two categories: white-box and black-box. In order to investigate adversarial attacks against mmWave-based HAR, we first adopt the white-box assumption used by most of the previous research [7, 32, 87]. After that, we explore a more difficult black-box attack because black-box scenarios are more likely to occur in real-world applications. Exploring both white-box and black-box adversarial attacks allows us to evaluate and improve the security of mmWave-based HAR comprehensively and practically.

5.1 White-Box Attack

In the white-box scenario, the attackers have full knowledge of the machine learning model's input, architecture, and parameters. The adversary may also continuously access the victim model to produce adversarial samples. In addition, the adversary may be familiar with the HAR system's data preprocessing techniques in order to provide the proper perturbation. Based on that, white-box attacks are applicable to internal threats: they happen when someone within the organization with comprehensive knowledge of the system architecture and access to sensitive information acts maliciously. For example, attackers might modify the adversarial samples during the data preprocessing stage. The possibility of a white-box attack may be increased by a local adversary or information leakage.

The goal of this adversarial attack is to generate mmWave adversarial samples to confuse the mmWave-based HAR system. The mmWave-based HAR system can be conceptualized as a function f that receives mmWave signals as input and outputs the predicted activity class based on the probability score p for all the enrolled activity classes. Specifically, suppose there are n enrolled activities, where $p_i \in [0, 1]$ and $\sum_{i=1}^n p_i = 1$, the deep learning model f identifies the mmWave input as the class with the greatest probability score. In white-box scenarios, we study three adversarial attacks and formalize these attacks as follows:

Untargeted Attack In untargeted attacks, which is usually designed for a specific sample (sample-specific untargeted attack), the adversary aims to confuse the HAR

system by changing the output from the original activity prediction y to a different one y' . Specifically, given a machine learning model f and an activity sample x , a sample-specific untargeted attack can be formulated as $f(x + \delta) \neq y$, where x is the original activity sample, δ is the generated perturbation, and y is the original predicted activity of the classifier model. In order to achieve this, we should modify the activity sample by inserting δ to decrease the probability score of the original activity class p_y till it is lower than other activities.

Targeted Attack In targeted attacks for a specific sample (sample-specific targeted attack), the adversary aims to make the HAR systems output the desired class. The targeted attack can be formulated as $f(x + \delta) = z$, where $x + \delta$ is the adversarial sample and z is a pre-defined class. To enable this objective, we should modify the activity sample to increase the probability score of the desired activity p_z till it is higher than other enrolled activities.

Universal Attack To further improve the efficiency of targeted attacks and make it practical in time-constrained contexts, we develop universal attacks by generating a well-designed general perturbation. Then, we can insert it to different unseen activity samples directly without incurring additional training efforts. In particular, the activity data samples gathered at various times or under different conditions would often vary. Thus, the perturbation δ_1 designed to attack sample x_1 might not work for another sample x_2 , such as $f(x_2 + \delta_1) \neq z$. In addition, generating the perturbation for a high-dimensional mmWave activity sample (e.g., voxel-based data) is time-consuming, thus it is not always feasible to produce a sample-specific perturbation that is tailored for each activity sample. It is important to create some universal perturbations δ , such that $f(x_i + \delta) = z$, where x_i can be different samples from the same type of activity.

5.2 Black-Box Attack

Black-box is a more challenging scenario. It assumes that the target machine learning model is unavailable to the attacker. The adversary only knows the input and output of the model [34]. Thus, black-box attacks are usually launched by external actors who do not have prior knowledge of the system's inner workings. In black-box scenarios, it is possible for attackers to insert the adversarial sample right before the recognition phase, where activity data are sent as input to the machine learning model. In this case, the attackers can generate adversarial samples in advance and feed them into the machine learning models furtively. In addition, attacks can seize the original activity samples during the transmission from a local client to the server and then replace normal samples with adversarial ones due to the widespread usage of cloud computing and federated learning [40, 64]. We investigate our adversarial attacks on HAR in black-box settings, because they are more realistic than white-box settings. In black-box scenarios, we would explore whether a perturbation generated based on model f could still work on another

model f' , where f' has different structures and parameters from f . This attack can be formulated as $f'(x') = f(x')$, where x' is the adversarial sample generated based on model f .

In order to create a practical adversarial sample that is difficult for humans to identify, the distortion caused by the perturbation should be as minimal as possible. It can be formulated as $\min \|\delta\|_p$, s.t. $f(x + \delta) = z$. Additionally, it brings additional difficulties to produce reliable and undetectable adversarial perturbations due to the unique characteristics of mmWave signal representations (e.g., voxel-based data). Since we should consider both the effectiveness of our attack and the distortion of the perturbations, we formally define the objective function as minimize $\mathcal{D}(x, x + \delta)$, such that $f(x + \delta) = z$. $\mathcal{D}$ is the distance metrics $\|\delta\|_p$ that evaluates the magnitude of the generated perturbation. However, as discussed in previous work [10], directly solving this non-linear constrained non-convex problem is difficult. Thus, we reformulate the objective function as a gradient-based optimization instance:

$$\text{minimize } \mathcal{L}(x + \delta) + \lambda * \mathcal{D}(x, x + \delta), \quad (1)$$

where the first component $\mathcal{L}$ represent the adversarial loss which measures the possibility of launching adversarial attacks successfully and the second component $\mathcal{D}$ represents perturbation loss which constraints the perturbation size.

6 Attack Design

In this section, we explore algorithms for adversarial attacks, delineating both white-box and black-box strategies. Figure 4 graphically demonstrates the progression of our white-box methodology. To generate adversarial mmWave activity samples, we initially select loss functions based upon varying attack strength requirements, such as untargeted attacks or targeted attacks for mmWave-based human activity recognition systems. Enhancing the validity and efficacy of the adversarial mmWave samples demands the adoption of clipping and discretization processes, paired

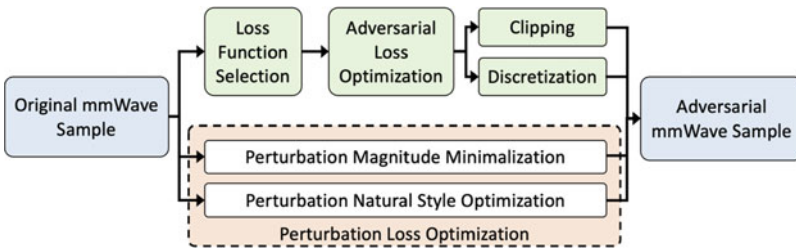


Fig. 4 The flow and components of the designed adversarial mmWave activity sample generation in white-box attack

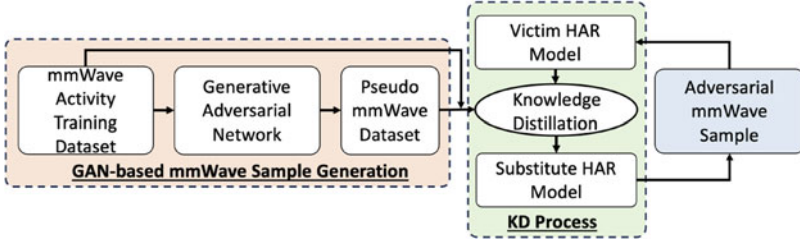


Fig. 5 The flow and components of the designed adversarial mmWave activity sample generation in black-box attack

with adversarial loss optimizations. We focus not only on minimizing the perturbation magnitude but also on examining the unique characteristics of mmWave activity data, thereby defining two distance metrics to foster more unnoticeable adversarial samples. Additionally, we contrive a practical and efficient universal targeted attack method, shaping a general perturbation during the offline training phase. This perturbation readily adapts to runtime mmWave activity samples for a targeted attack. Moreover, we delve into black-box attacks, proposing various methods to evaluate the feasibility of launching adversarial attacks under different circumstances. Figure 5 outlines the implemented black-box attack procedure. To facilitate the black-box attack, we first devise a knowledge distillation method for the creation of the substitute model. The adversarial mmWave activity sample is subsequently generated based on this substitute model. In more challenging cases where the available training data for creating this substitute model is inadequate, we use a GAN-based method to produce an abundant number of high-quality mmWave data, thus compensating for the insufficient mmWave training dataset.

6.1 White-Box Attack Implementation

6.1.1 Targeted and Untargeted Attack

Adversarial Loss We first describe the implementation of our method for specific mmWave activity sample. For sample-specific untargeted attacks, we define the objective function as $\mathcal{L} = \max(\mathcal{Z}(x + \delta)_s - \max_{i \neq s}(\mathcal{Z}(x + \delta)_i), -k)$, where $\mathcal{Z}(x + \delta)_s$ represent the possibility of estimating the activity as the original activity class (i.e., the predicted class without attack), and $\mathcal{Z}(x + \delta)_i$ represent the possibility of estimating the activity as another class (i.e., a class that is different from the original-predicted activity class). k is a configurable parameter that controls attack confidence. For sample-specific targeted attacks, we define $\mathcal{L} = \max(\max_{i \neq t}(\mathcal{Z}(x + \delta)_i) - \mathcal{Z}(x + \delta)_t, -k)$, where $\mathcal{Z}(x + \delta)_t$ is the possibility of estimating the activity as the class t we desired. By optimizing the above adversarial loss functions, we aim to make our attack method not only confuse the HAR

systems (untargeted attack), but also force the HAR system to output our desired class (targeted attack). In practice, by using different special-designed adversarial loss functions, the attacker could either launch an untargeted attack or a targeted attack according to different attack strength requirements, which makes our attack framework more powerful and dangerous than previous studies [48, 87].

Perturbation Loss Generally speaking, the perturbations are the difference between the original mmWave sample and the adversarial one. L_2 Norm, which calculates the Euclidean distance between two sets has been commonly used as a metric for adversarial perturbation evaluation [10, 43, 48]. In this project, we define the perturbation loss $\mathcal{D} = \|\delta\|_2^2$ and generate the perturbation with minimal magnitude by optimizing the perturbation loss. In order to ensure the effectiveness of the perturbation and improve the efficiency of perturbation generation, we set a dynamic threshold τ for each mmWave sample to ensure $\|\delta\|_2^2 < \tau$, s.t. $f(x + \delta) = z$. The threshold is derived by analyzing the deviation between the normal mmWave sample and other normal samples. For a specific sample, we calculate the average L_2 Norm between the sample and all other available samples of the same type of activity, and set it as the threshold τ for perturbation generation.

Parameter Selection The weight λ , which determines the balance between the adversarial loss $\mathcal{L}$ and perturbation loss $\mathcal{D}$, must be set to a suitable number in order to cause gradient descent to minimize both components concurrently, as opposed to optimizing over one term at a time. In practice, we do a 12-step binary search to identify the appropriate λ and its accompanying adversarial perturbation δ .

6.1.2 Perturbation Optimization

Clipping In order to ensure the validity of the adversarial sample, there should be a clipping process after each training iteration. The clipping process trims the value of the adversarial sample to fall inside a valid range $[\alpha, \beta]$, which should be chosen based on the data representation of the activity samples in the HAR system. For a voxel-based HAR (e.g., [63]), the range should be $[0, \infty]$, as the value of each voxel represents the number of points within its limit. For a heatmap-based HAR (e.g., [77]), the range is usually set to $[0, 255]$.

Discretization Discretization is a crucial process that is usually neglected in prior research [14, 43, 48]. However, due to the specific properties of mmWave data, we discover that perturbation discretization is necessary and cannot be disregarded. Specifically, the value of each pixel in a valid adversarial heatmap must be a discrete integer between 0 and 255, and a valid voxel often has a much lower upper limit value (e.g., 5) because of the sparse point clouds. Using the previous method that round the value of each adversarial voxel or heatmap to the nearest integer could eliminate minor perturbations and render the adversary's attack ineffective.

To handle this discrete optimization issue, we incorporate another loss function $\mathcal{L}_{model}(\lfloor x + \delta \rfloor)$, where $\lfloor x + \delta \rfloor$ represents the discrete adversarial sample. We mark the original $\mathcal{L}$ mentioned in Sect. 6.1.1 as $\mathcal{L}_{adv}$, and reformulate the final adversarial loss function as $\mathcal{L} = \mathcal{L}_{adv}(x + \delta) + \mathcal{L}_{model}(\lfloor x + \delta \rfloor)$. By simultaneously optimizing $\mathcal{L}_{adv}$ and $\mathcal{L}_{model}$, we could ensure the validity and efficiency of adversarial samples in different kinds of HAR systems.

Natural Style Optimization Furthermore, we discovered that the majority of existing approaches [10, 34, 48, 87] only focus on minimizing the perturbation magnitude by using a smaller L_2 Norm. However, little attention has been given to optimizing adversarial activity samples to have a natural style. In this study, we design a method to achieve natural style optimization by minimizing the radius of the generated perturbation and reducing the distance between the perturbation and the original mmWave sample. Our approach is important for mmWave activity samples because if the generated perturbation is too sparse or entirely separated from the original mmWave sample, it will be noticeable in heatmaps or voxels.

To minimize the radius of the generated perturbation, we reduce the pairwise Euclidean distance between elements inside the perturbation. We formulate it as $\mathcal{D}_{mean}(\delta) = \max_{m, n \in \delta} \|m - n\|_2$, where m, n are positions of any two elements (e.g., pixels in the heatmap) inside the generated perturbation δ . By reducing the average pairwise distance inside the perturbation, the radius of the perturbation can be reduced. We further reduce the distance between the perturbation and the original activity sample, allowing the perturbation to be concealed within the normal samples. In particular, we calculate *Chamfer Distance*, which seeks the nearest pairwise element euclidean distance between the generated perturbation and activity sample and takes the mean of all nearby element pair distances. It is expressed as $\mathcal{D}_{cf}(x, \delta) = \frac{1}{\|\delta\|_0} \sum_{m \in \delta} \min_{n \in x} \|m - n\|_2^2$, where x is the activity sample. By reducing the *Chamfer Distance*, the inserted perturbation is pushed nearer the activity sample. After integrating the above two functions, we reformulate the final perturbation loss function as follows:

$$\mathcal{D} = \mathcal{D}_{mag}(x, x + \delta) + \mathcal{D}_{mean}(\delta) + \mathcal{D}_{cf}(x, \delta), \quad (2)$$

where $\mathcal{D}_{mag}(\delta)$ controls the magnitude of the perturbation as mentioned in Sect. 6.1.1.

6.1.3 Practical Universal Targeted Attack Design

In this part, we provide details on how to launch efficient targeted attacks against HAR through universal perturbation design. Our basic idea is to create universal perturbations δ , such that $f(x_i + \delta) = z$, where x_i can be any activity samples from the same type of activity. The designed universal perturbation generation method consists of an offline training phase in which a training activity set is utilized to produce a universal perturbation, and an online test phase in which the universal

Algorithm 1: Universal Perturbation Generation

Input: Training set $\Omega = \{\Omega_1, \Omega_2, \dots, \Omega_i\}$, HAR model f , targeted activity class z , desired perturbation magnitude τ , desired attack success rate ϵ on training set.

Output: Universal perturbation δ .

```

1: Initialize  $\delta \leftarrow 0$ .
2: while Success Rate( $\Omega$ ) <  $\epsilon$  do
3:    $\Omega_j \leftarrow RS(\Omega)$   $\triangleright$  Random Select a Sample
4:   if  $f(x_j + \delta) = z$  then
5:     Calculate the perturbation that satisfies:  $\delta \leq \tau$ .
6:   else
7:      $\Delta\delta_j \leftarrow \arg \min_{\Delta\delta_j} \mathcal{D}(\Delta\delta_j)$ 
       such that  $f(\Omega_j + \delta + \Delta\delta_j) = z$ .
8:      $\delta \leftarrow (\delta + \Delta\delta_j)$ .  $\triangleright$  Update the Perturbation
9:   endif
10: end while

```

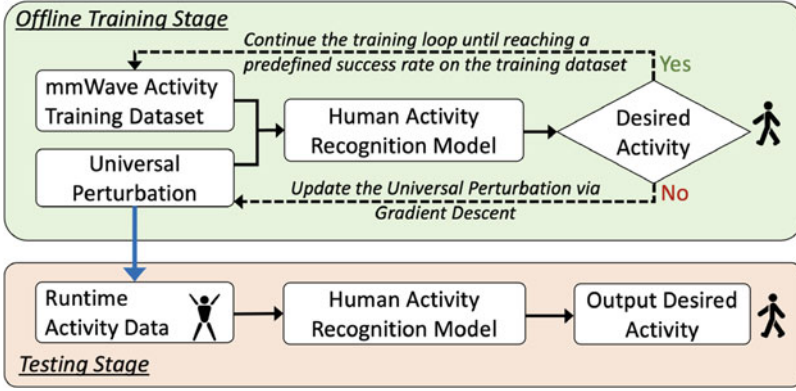


Fig. 6 The overview of our designed universal attack consists of an offline perturbation generation phase and an online testing phase

perturbation is directly applied to runtime activity data for a targeted attack. As shown in Fig. 6, we generate universal perturbations δ for each type of activity, such that when the perturbation is applied to the majority of activity data x from the same class, the HAR always recognizes it as our desired class z . We generate the perturbation for each activity sample in the training set using the same objective function (Eq. (1)). To make the adversarial perturbation work for the majority of activity examples in the training set, we iteratively adjust the universal perturbation.

Specifically, the adversarial perturbation is started with zeros and added to an mmWave activity sample. If the HAR's prediction does not match the desired activity class, the perturbation will be modified in the direction of gradient descent, in which the likelihood of the desired class increases. Otherwise, the current perturbation is applied to a fresh training activity sample. If the existing universal perturbation does not fit in the new sample, a minimal magnitude perturbation revision is calculated and added to the current universal perturbation. The iteration

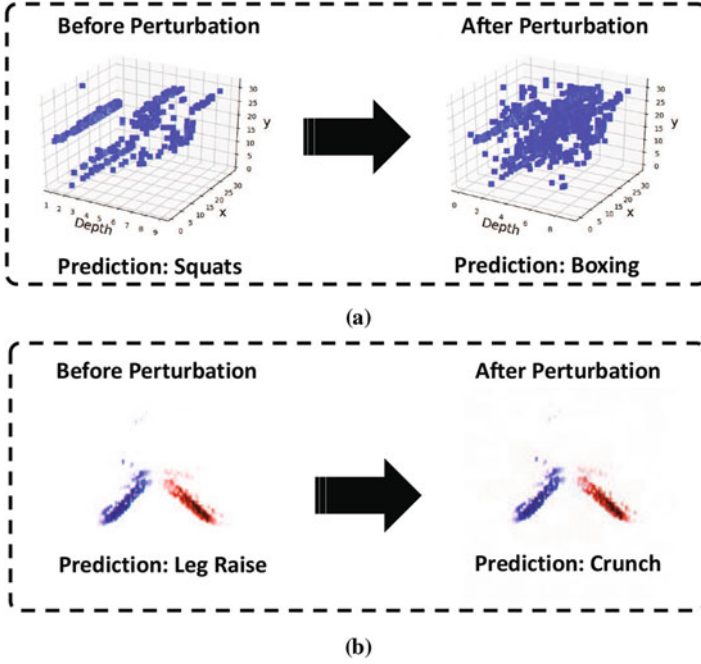


Fig. 7 Two representative adversarial samples generated by adding universal perturbations directly. (a) Adversarial voxel-based data generation with L_2 Norm of 21; (b) Adversarial heatmap-based data generation with L_2 Norm of 2083

process ends when the universal perturbation on the training dataset exceeds a predefined success rate (e.g., 70%). Notably, the objective of the technique is not to seek the smallest global perturbation that fools the majority of activity samples, but rather to select one that is sufficiently tiny. Figure 7 depicts the production of adversarial samples by directly applying universal perturbation. We observe that the adversarial instances deviate from the original sample only slightly in terms of L_2 Norm. The adversarial samples look natural, which makes them hard to be noticed by naked eye. However, adversarial examples enable HAR systems to efficiently predict the activity as desired. Compared with traditional sample-specific attacks methods, our universal perturbations would significantly shorten the attack launch time, which makes it more practical in time-constraint attack scenarios. Different from existing universal untargeted attack methods that need to insert padding frames between two successive activities [48], our method modifies the activity sample directly and thus broadens its applicability to various of mmWave-based HAR systems. In addition, our universal attack method is compatible with both untargeted and targeted attacks by merely modifying the adversarial loss function.

6.2 Black-Box Attack Implementation

Black-box attack is a more challenging scenario where the attacker usually cannot access the victim model but only the input and output of the model [48]. Thus, the adversarial perturbation cannot be created and updated by exploiting the gradient from the victim model. A potential approach to a black-box attack is to train a substitute model. Adversarial samples generated by the substitute model can be exploited to launch attacks towards the victim model, leveraging the transferability of the adversarial sample.

We begin with a basic black-box setting where the training data of the victim model is fully accessible. The key challenge is how to ensure the similarity between the target and the substitute model. Directly training the substitute model on the dataset usually gets poor performance since the structure of the substitute model is different [48]. To solve such a problem, we take advantage of Knowledge Distillation (KD) to learn a substitute model that can mimic the prediction of the victim model [21]. In black-box scenarios, although the inner structure of the victim model is inaccessible, its output class and soft logits indicate the class probability distribution for a given input is accessible [34]. Supposing the soft logits of the target and the substitute model are P_t and P_s , respectively. The predicted class of substitute model is Z and the ground truth is G . We formulate the KD process as the loss function $L = L_s + L_d$, where $L_s = \text{CrossEntropy}(Z, G)$ and $L_d = \text{KLDivergence}(P_t, P_s)$. By optimizing the loss function, we transfer the dark knowledge from the victim model to the substitute model [21].

To ensure the robustness of the black-box attack, we utilize a configurable parameter of k to control the confidence of the attack as mentioned in Sect. 6.1.1. With a larger k value, the possibility that the adversarial sample being misclassified by the victim model will increase. We set $k = 0$ in white-box scenarios and set a larger k in black-box scenarios. We evaluate the impact of k in Sect. 7.4.

We then move to a more challenging scenario where the original training data of the victim model is only partially accessible. To deal with the problem of insufficient training data, we develop a GAN to synthesize sufficient pseudo training samples. GAN has been proved to generate high-quality pseudo samples with a limited amount of real samples [77]. In this chapter, we implemented a GAN with a 3-layer generator and a 3-layer discriminator to generate sufficient activity samples using only 20% of the original training dataset of the victim model. Specifically, the generator seeks to learn the distribution of the real samples so as to have the ability of synthesizing pseudo samples. The discriminator tries to discriminate whether a sample is a real or pseudo one. The generator and discriminator are trained in turn to optimize each other by updating the parameters of their networks. The final state is a Nash equilibrium, where the synthesized pseudo samples are similar to the real ones, and the discriminator fails to identify whether the activity samples are real or not. After obtaining enough high-quality pseudo training samples, we exploit the KD method mentioned above to train the substitute model and launch black-box attacks toward the victim model.

7 Performance Evaluation

7.1 Experimental Setup

Equipment We have classified mmWave-based HAR systems into two distinct categories based on their unique data representations after conducting an exhaustive investigation. This approach to classification allowed us to select one representative dataset from each category in order to implement our designed adversarial attack methodologies. Specifically, we develop our own mmWave-based HAR system [83] with heatmap representations in order to examine its susceptibility to adversarial attacks. We construct our own HAR system (heatmap-based) using a single mmWave device, the TI AWR1642 [8], which incorporates a 2×4 antenna array. The frequency range of the device is between 77 GHz and 81 GHz. The sampling rate is fixed at 100 frames per second, with 17 chirps per frame. A TI DCA1000EVM [69] data capture card is employed to acquire data from the mmWave device and transmit it to a Dell laptop for deep model inference. In addition, we select an existing mmWave-based HAR system [63] as a representative to investigate the voxel-based HAR model's susceptibility to adversarial attacks. It utilizes TI IWR1443BOOST radar [70] to gather the new point cloud dataset. This radar also operates in the frequency range of 77 GHz to 81 GHz. With four receiver and three transmitter antennas, the radar is able to monitor multiple objects based on their distance and angle data. This antenna design allows for the estimation of both azimuth and elevation angles, enabling the detection of objects in a three-dimensional plane.

Data Collection Two datasets are used in our experiment. The public voxel-based human activity dataset includes 156,355 samples representing 5 distinct activities. The radar is mounted on a tripod platform at a height of 1.3 m for data collection. They have gathered information from two users. Five distinct activities are performed by the participants in front of the radar. These include walking, jumping, jumping jacks, squatting, and boxing. For a subject performing the same activity, data is collected in continuous intervals of approximately 20 seconds. They have collected a total of 93 minutes of data. The captured point clouds include spatial coordinates (x,y,z in meters), velocity (in meters/second), range (distance of the point from the radar) in meters, intensity (dB), and bearing angle (degrees). The radar's sampling rate is 30 pixels per second.

They separated collected data files into train and test files, with the train containing 71.6 minutes of data and the test containing 21.4 minutes of data. To surmount the nonuniform number of points in each frame, the point clouds were converted into voxels with dimensions of $10 \times 32 \times 32$ (depth=10), making the magnitude of the input constant regardless of the number of elements in the frame. By evaluating their efficacy, they empirically determined these dimensions. The value of each voxel in the voxel representation is the number of data points contained within its boundaries. Due to the fact that activities are carried out over a

Table 1 14 common in-home full-body activities

W1	Crunches	W8	Squats
W2	Elbow plank and reach	W9	Burpees
W3	Leg raise	W10	High knees
W4	Lunges	W11	Turning kicks
W5	Mountain climber	W12	Chest squeezes
W6	Punches	W13	Side leg raise
W7	Push ups	W14	Side to side chops

period of time, the time window from activities is derived to encapsulate temporal dependencies. They generate 2-second windows (60 frames) with a 0.33-second sliding factor (10 frames). Finally, they receive 12,097 in training samples and 3538 in testing samples. 20% of the training samples are used for validation. In the voxelized representation of the time window, each sample has a $60 \times 10 \times 32 \times 32$.

For our human activity heatmap-based dataset. We enlist 11 volunteers aged 20 to 44, ranging in height from 162 to 185 centimeters and weight from 50 to 86 kilograms. The volunteers are required to complete 14 typical indoor exercises, as outlined in Table 1. Four distinct environments (such as the lounge, corridor, small classroom, and large classroom) are utilized to capture exercise data. We position the mmWave device on a 60cm-high table and record the ground truth recordings using a camera. During an 8-month study, we ask each participant to complete at least 20 repetitions of each type of exercise (half of the segments are used to train the model and the rest segments are used to assess performance). We collect over 7000 segments from volunteers in total.

Victim Models Deep learning models are utilized by both the voxel-based and heatmap-based HAR systems due to their superior performance compared to traditional machine learning models. The authors train a convolution neural network (CNN) combined with long short-term memory (LSTM) for the voxel-based victim HAR model. CNN layers are applied to every temporal segment of the input data by time-distributed CNN. Time-distributed CNN + Bi-directional LSTM classifier consists of three time-distributed convolutional modules (convolution layer + convolution layer + maxpooling layer) followed by a bi-directional LSTM layer and an output layer. The network has 291,000 trainable parameters overall. This classifier is trained directly on the input sample’s time and spatial dimensions. They use the sklearn GridSearchCV function to optimize the SVM’s hyperparameters (C and gamma). Deep learning classifiers were trained using Adam optimizer with a learning rate of 0.001. After training for 30 epochs, the models with the lowest loss on validation data were preserved. The original classification accuracy of the voxel-based deep learning model is 90.47%.

For the heatmap-based victim HAR model, we use a convolution neural network. The input of the deep learning model is an array with a size of $224 \times 224 \times 3$. The array is generated by quantifying all the pixels in the Spatial-Temporal Heatmap into 3-D tuples representing RGB value. The feature extractor has three convolutional

layers, each with a 3×3 filter and a ReLU activation function. The convolutional layers are used to up-sample the image by extracting high-dimensional features from the spatial correlation of pixels. Each convolutional layer is followed by a max-pooling layer with a stride of 2 and a filter size of 2×2 to downsample and reduce data redundancy. After the process of 3 rounds of up-sampling and down-sampling, a 64-dimension feature map is obtained. Then, a flatten layer is integrated to reduce the feature map into a one-dimension array. Given an input data D , the feature extractor produces feature representations $R = F(D, \Theta_f)$, where F represents the feature extractor and Θ_f represents its trainable parameters. Based on the derived feature representation R , a neural network consisting of two dense layers is followed to classify the inputs into several classes (e.g., different types of workouts). Given the input representation R , the classifier predicts the label as Y_c . We optimize the classifier by minimizing the cross-entropy loss between the ground truth $\tilde{Y}_c$ and the predicted label Y_c as $L_c = L_{CE}(Y_c, \tilde{Y}_c)$, where L_{CE} represents the cross-entropy loss function. The original classification accuracy of the heatmap-based deep learning model is 97%. In addition, for the implementation of our adversarial attacks in this project, the prototype of our designed attack methods is implemented using Python along with TensorFlow.

Evaluation Metrics We use three metrics to evaluate the performance of our attack scheme. (1) *Success Rate (SR)*: it represents the number of succeeded adversarial attacks over the total number of attack attempts. In the untargeted attack, we report a success when the predicted class is different from the original class while in the targeted attack, we only report success if the predicted class matches the desired target class; (2) *L_2 Norm*: it indicates the Euclidean distance between the adversarial sample and original sample; Smaller L_2 Norm values indicate that the adversarial sample is similar to the original activity sample, therefore, harder to be noticed by human eyes. (3) *Confusion Matrix*: Each cell in the matrix indicates an original-target class pair that the actual class in the row is classified as the target class in the column. The value of each cell represents the average SR and L_2 Norm of the corresponding universal attack on the testing set.

7.2 Evaluation of White-Box Attack

First, we evaluate the efficacy of sample-specific targeted attack and untargeted attack in white-box scenarios for mmWave-based HAR systems. In the white-box scenario, the attackers have complete knowledge of the inputs, architecture, and parameters of the victim deep learning model. The adversary may also access the victim model continuously to generate adversarial samples. In addition, the adversary may be familiar with the data preprocessing procedures of the HAR system in order to provide the appropriate perturbation. To thoroughly evaluate the efficacy of sample-specific attacks, we randomly select 200 and 100 activity

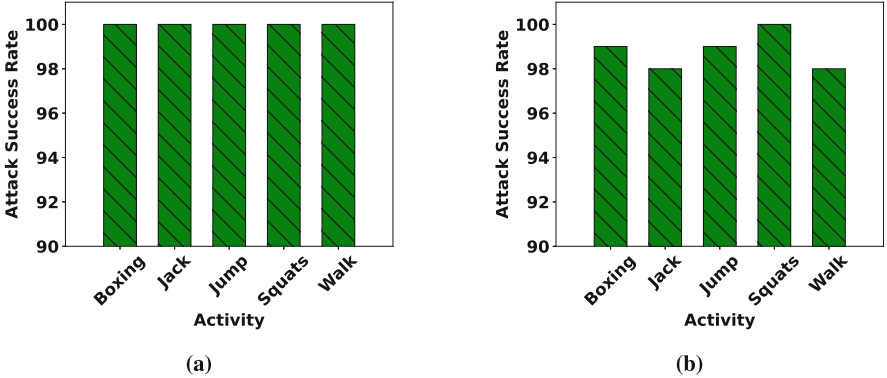


Fig. 8 (a) Success rate of sample-specific untargeted attacks on voxel-based dataset; (b) Success rate of sample-specific targeted attacks on voxel-based dataset (x-axis represents the original classes)

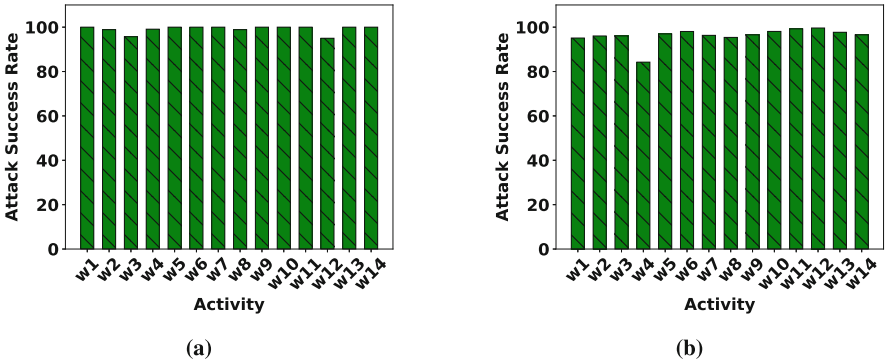


Fig. 9 (a) Success rate of sample-specific untargeted attacks on heatmap-based dataset; (b) Success rate of sample-specific targeted attacks on heatmap-based dataset (x-axis represents the original classes)

samples for each activity type from the voxel-based and heatmap-based testing sets, respectively. To evaluate the efficacy of universal attacks in white-box scenarios for mmWave-based HAR systems, half of the selected samples are used to generate universal perturbations (universal attack training set), while the other half are used to evaluate the universal attack (universal attack testing set).

Untargeted Attack In untargeted attacks, the adversary attempts to confound the HAR system by modifying the output from the original activity prediction to a different prediction. In another word, we report a success untargeted attack when the predicted class is different from the original class. Figures 8a and 9a demonstrate the attack success rate of our untargeted attacks on the voxel-based HAR dataset and heatmap-based dataset, respectively. We can learn that our method achieves nearly 100% attack SR for all 5 original classes in the voxel-based dataset and all

14 original classes in the heatmap-based dataset, indicating that almost all samples tested are class-flipped from the original class under our attack scheme. Because the goal of untargeted attacks is just to cause the HAR models to make mistakes, these high attack success rates demonstrate the effectiveness of our methods.

Targeted Attack In targeted attacks, the adversary seeks to cause HAR systems to predict a particular class. To achieve this objective, we must modify the sample of activities so that the probability score of the preferred activity in the victim model is greater than that of all other enrolled activities. This work is more difficult because we only report a successful targeted attack if the predicted target class matches the intended target class. Figures 8b and 9b demonstrate the attack success rate of sample-specific targeted attacks on the two datasets, respectively. The x-axis indicates the initial classifications. Each sample from the original class is modified to be recognized as the other activities, i.e., the other 4 activities in a voxel-based dataset or the other 13 activities in a heatmap-based dataset. Our method achieves an average SR of 96% on both datasets. Note that attacks towards some target classes have relatively lower SR (i.e., jack and walk from the voxel-based dataset; and *w4* (lunges) from the heatmap-based dataset). This is because those classes have more different patterns from other activities, making the attack relatively harder. But even the lowest SR in targeted attack is still higher than 80%, proving the effectiveness of our targeted attack method.

Universal Attack We design the universal attacks to enhance the efficacy of targeted attacks and make them applicable in time-constrained environments. This would permit the use of targeted attacks in situations with limited time. By launching a universal attack, we will be able to directly insert the universal perturbation into previously unseen activity samples without incurring any additional training expenses. Launching universal attacks is not a simple task due to the fact that activity data samples gathered at different times or under different conditions frequently differ, meaning that the perturbation designed to attack one sample may not be effective against another sample, even if they are from the same activity type. Launching universal attacks is necessary. It is not always possible to generate a sample-specific perturbation that is tailored to each activity sample due to the time-consuming nature of generating the perturbation for a particularly high-dimensional mmWave activity sample. It is more practicable to generate universal perturbations that can be applied to distinct samples from the same type of activity. To illustrate the efficacy of our universal attacks, we conduct targeted attacks against each activity type and present the results using a confusion matrix. Each cell in the matrix represents a pair of original-target classes, where the actual class in the row corresponds to the target class in the column. The value of each cell represents the SR of the corresponding universal attack on the testing set. The SR of universal attacks represents the percentage of activity samples in the testing set that are classified as the desired class when subjected to the identical perturbation.

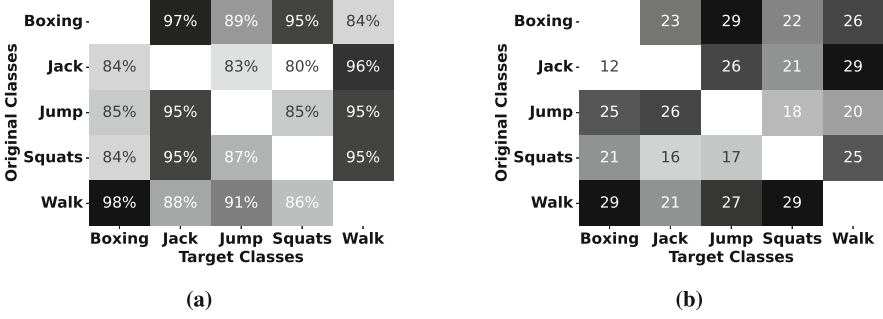


Fig. 10 (a) Success rate of universal targeted attacks on voxel-based dataset; (b) L_2 Norm of generated universal perturbations on voxel-based dataset

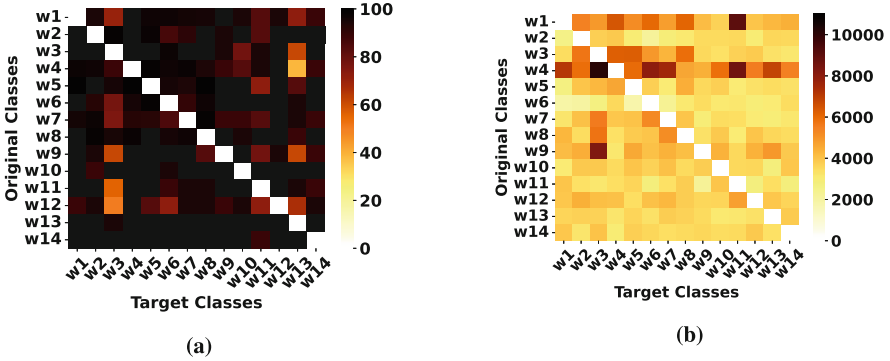


Fig. 11 (a) Success rate of universal targeted attacks on heatmap-based dataset; (b) L_2 Norm of generated universal perturbations on heatmap-based dataset

The performance of our universal attacks over the voxel-based HAR dataset is demonstrated in Fig. 10a. We can learn that all universal attacks achieve over 80% SR, with the highest SR reaching 98% (98% of walk samples in the testing set has been classified as boxing using the same universal perturbation). We note that the SR of some original-target pairs (e.g., jack-squats) is relatively low. This is because the samples of the target class vary a lot from the original class, making it harder to launch targeted attacks. Despite this, our method still achieves an overall SR of 90%. For the heatmap-based dataset, as is shown in Fig. 11a, attacks on most original-target pairs achieve higher than 90% SR. Few pairs (e.g., $w4-w13$, $w12-w3$) have relatively low SR due to large differences between original and target samples. But our method still reaches an average SR of 94% over 182 original-target pairs on the heatmap-based dataset.

7.3 *Impact of Perturbation Magnitude*

There are a number of reasons why it's important for us to minimize the impact of adversarial perturbations and create adversarial activity samples that are highly similar to the original samples. First, we aim to subtly alter the activity samples so that the perturbations are virtually undetectable while preserving the original sample's inherent properties. This stealthiness is essential in a hostile environment, where noticeable changes could signal the presence of an attack. Second, we strive to preserve the original classification of an activity from a human standpoint, despite causing misclassification in the recognition system. This highlights the disparity between human and HAR model perception, highlighting the vulnerabilities of the HAR system. Using minimal perturbations, we demonstrate that even minor deviations from the original mmWave data can result in substantial misinterpretations by the victim system. Our research aims to make a clear illustration of the HAR system's security issues, spurring the creation of more durable mmWave-based HAR systems.

In this research, we optimize the L_2 Norm in order to minimize the magnitude of the perturbation. In order to guarantee the efficacy of the perturbation and increase the efficacy of its generation, a dynamic threshold is set for each activity sample. For a particular sample, we compute the average L_2 Norm between the sample and all other available samples of the same type of activity and use this value as the perturbation generation threshold.

Untargeted Attack To evaluate the impact of perturbation magnitude on our white-box attacks, we first evaluate the impact of perturbation magnitude on untargeted attacks. Figure 12a demonstrates the L_2 Norm of untargeted attacks on the voxel-based dataset. Note that each red line indicates the average value of the thresholds of each activity type. We can learn that the median L_2 Norm of untargeted adversarial samples on all 5 original classes are below 10 and the maximum L_2 Norm values are all lower than 30, far below the 5 average thresholds, which are all around 40–50. For the heatmap-based dataset, as is shown in Fig. 13a, the median L_2 Norms between adversarial and original samples are around 2000–2500. Adversarial samples towards one workout, $w4$, have relatively higher L_2 Norm distribution due to high-specific features of original heatmaps. But the highest L_2 Norm is still lower than 4000, far below the average threshold of 14,000 for $w4$.

Targeted Attack We then evaluate the impact of perturbation magnitude on targeted attacks. Figure 12b demonstrates the L_2 Norm of targeted attacks on the voxel-based dataset. We find that the median L_2 Norm values of samples towards all 5 target classes are still around 20. But there is a significant increase in the maximum L_2 Norm value on all 5 classes compared with untargeted attacks. This is because in targeted attacks, we not only need to flip the class but also need to turn the class to the required one. Thus, for some samples a larger perturbation magnitude is needed. But even the maximum L_2 Norm values are still below the corresponding average threshold (i.e., red lines in the Figure). On the heatmap-based dataset, as is shown in

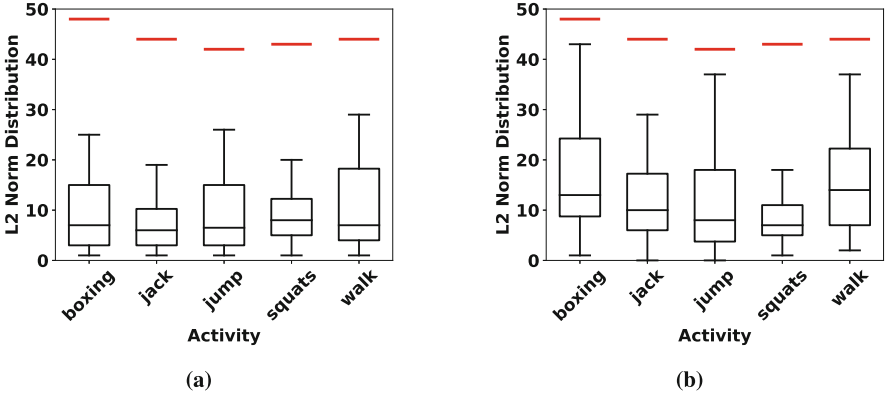


Fig. 12 (a) L_2 Norm of perturbations generated in sample-specific untargeted attacks on voxel-based dataset (The red line represents the average threshold of all attack samples); (b) L_2 Norm of perturbations generated in sample-specific targeted attacks on voxel-based dataset (x-axis represents the original class)

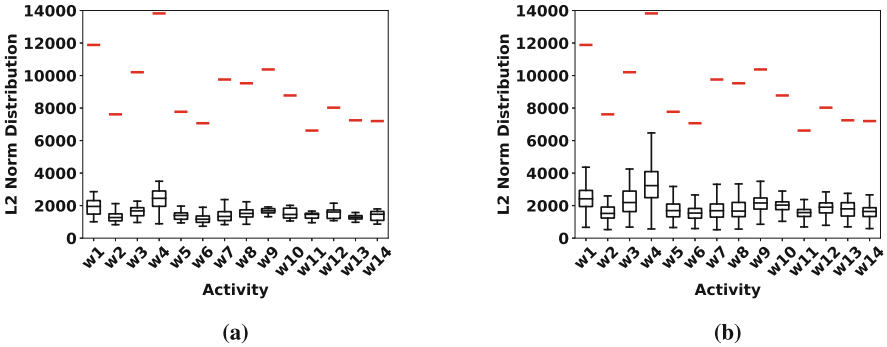


Fig. 13 (a) L_2 Norm of perturbations generated in sample-specific untargeted attacks on heatmap-based dataset; (b) L_2 Norm of perturbations generated in sample-specific targeted attacks on heatmap-based dataset

Fig. 13b, we also notice a larger L_2 Norm distribution compared with the result of untargeted attacks. Samples aiming at the target class of w_4 have a relatively higher maximum L_2 Norm value due to the highly-specific features of the original heatmap from this class. But even the maximum perturbation (i.e., 6300) of attack samples (i.e., w_4) still does not exceed the corresponding average threshold.

Universal Attack We next evaluate the impact of perturbation magnitude on universal attacks. The confusion matrix of universal L_2 Norm on voxel-based dataset and heatmap-based dataset are shown in Figs. 10b and 11b, respectively. We can learn that the average L_2 Norm for the adversarial samples towards voxel-based dataset is between 12 to 29, which is far below the average threshold of 5 classes

(i.e., around 40–50). The average L_2 Norm of universal samples towards heatmap-based dataset over 182 original-target pairs is 4000. Though some pairs (e.g., $w4-w3$, $w1-w11$) have relatively higher perturbation magnitude due to relatively large differences in heatmap patterns, these values are still below the average threshold of corresponding original classes.

7.4 Evaluation of Black-Box Attack

In this research, we delve deeper into black-box attacks for several crucial reasons. First, black-box attacks are a more probable form of threat. In real-world applications, attackers typically do not have access to the HAR model’s architecture, parameters, or training data. By examining black-box attacks, we hope to better prepare these HAR systems for actual threats. The successful execution of black-box attacks against mmWave-based HAR systems can reveal fundamental vulnerabilities that are independent of the system’s design. This enables researchers to comprehend the robustness of the HAR system in a more general sense, thereby guiding the development of more secure systems. Lastly, investigating black-box attacks can provide valuable insights into the design of defensive mechanisms that do not rely on modifying the HAR system’s internals, but instead emphasize external measures, such as activity sample validation or anomalous activity detection. Exploring black-box adversarial attacks permits us to comprehensively and practically evaluate and improve the security of mmWave-based HAR.

All black-box experiments are taken on the heatmap-based dataset due to the page limit. We begin with the basic settings where the victim models are inaccessible but we assume that the attacker has full access to the training data set. We use KD to train a substitute model to generate adversarial samples and launch attack towards the victim model. Our substitute model is a 2-layer CNN network with 3.2M trainable parameters. We also trained the substitute model directly on the training set without KD for comparison. As mentioned in Sect. 6.2, we exploit a confidence value of k to ensure the robustness of our attack method. We change the k value from 0 to 40 with a step size of 5 to study the impact of k . Figures 14a, b and 15a demonstrate the average SR and L_2 Norm under basic black-box settings for untargeted, targeted, and universal attacks, respectively. We can learn that the substitute model trained with KD outperforms the directly-trained model for all k larger than 0 in all types of attacks. When $k = 40$, the substitute model can achieve over 80% SR for untargeted and targeted attack as well as an SR of 75% for universal attack. We can also notice a trade-off between SR, L_2 Norm and k values. As k increases, we can obtain a higher SR but the L_2 Norm also increases accordingly, meaning the adversarial samples have more significant perturbations. However, our method still maintains the L_2 Norm value of less than 6500 for all three types of attacks even when $k = 40$.

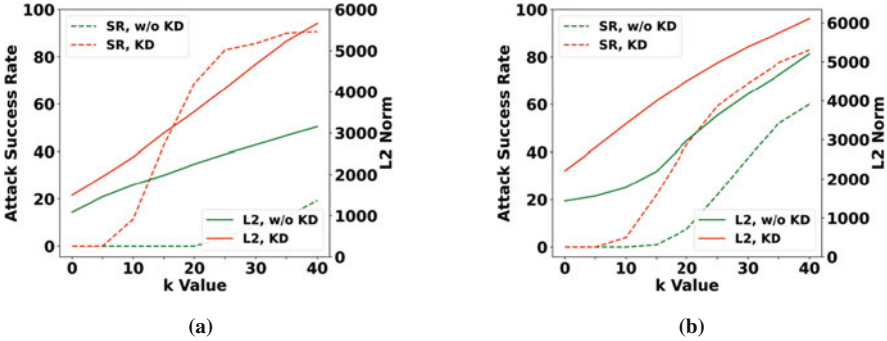


Fig. 14 (a) Success rate and L_2 Norm of untargeted black-box attack; (b) Success rate and L_2 Norm of targeted black-box attack

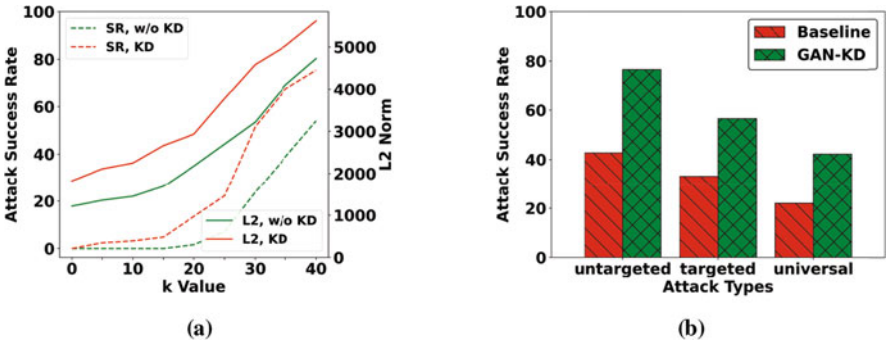


Fig. 15 (a) Success rate and L_2 Norm of universal black-box attack; (b) Success rate of GAN-KD-based black-box attack

We next move to a more challenge setting where the adversary can only access part of the training data used by the victim model. To overcome the training data deficiency issues, we exploit the GAN method mentioned in Sect. 6.2 to generate a pseudo training set with a size similar to that of the original training set using only 20% of original training data. The substitute model is trained using KD and the generated training set. We set confidence value $k = 40$ since previous results have proven that this confidence value can obtain relatively robust performance. For comparison, we trained a baseline model without KD and GAN using only 20% of the original training data, similar to the black-box model used in [48]. As is shown in Fig. 15b, the GAN-KD trained substitute model outperforms the baseline model for all 3 types of attacks, with the highest SR of 76.5% for the untargeted attack. Due to higher requirements for the adversarial samples, SR of targeted and universal attacks using the GAN-KD method are relatively low (i.e., 56 and 42%), but the SR still outperforms the baseline model with an increase of 23.4 and 20.22%, respectively.

8 Conclusion

Artificial intelligence (AI) has arisen as a swiftly advancing subfield of computer science with the goal of emulating and augmenting human intelligence. This expansion has been fueled by ultra-high-performance computing technology and the introduction of deep learning, resulting in a significant evolution of traditional AI technology. In recent years, deep learning has marked a substantial surge forward in the development of artificial intelligence in areas such as computer vision, speech recognition, and text comprehension. It is capable of detecting concealed non-linear correlations in data, supporting new file types, and identifying unidentified threats, which significantly improves cybersecurity defense. However, even the most sophisticated AI technologies, such as deep learning, are susceptible to adversarial attacks that can result in incorrect classification or prediction outcomes.

Due to their convenience, adaptability, and capacity to transmit data over long distances without the need for physical connections, wireless signals have become widely adopted for communication. The prevalence of wireless signals in our everyday electronic devices enables complex networks of reflected rays in indoor environments. Advances in wireless technologies and sensing methodologies have facilitated the use of wireless signals, such as millimeter wave signals, for wireless sensing tasks. The fundamental premise of these applications is that human movement affects wireless signal propagation, making it possible to track human motions by examining the received signals. This breakthrough has led to various applications, including human activity recognition, human localization, and vital sign monitoring.

Among these, Human Activity Recognition (HAR) has gained considerable recognition as a key technology in numerous Internet of Things (IoT) and security applications, such as health monitoring and user authentication. However, the widespread deployment of HAR systems has uncovered vulnerabilities in security, specifically their susceptibility to adversarial attacks. Since the majority of HAR systems employ deep learning models for activity identification due to their high accuracy and ability to handle interference in the real world, a gap exists in research exploring the vulnerability of HAR systems to adversarial attacks. In domains such as healthcare, security, and smart home applications, false human activity detection can have significant repercussions. Given the extensive integration of HAR systems in these critical applications, a thorough investigation into the variety of adversarial attacks that target HAR models is crucial. This chapter focuses on adversarial attacks on AI-empowered HAR, aiming to increase awareness about their susceptibility and shed light on the security challenges they present.

In this chapter, we take a pioneering approach toward adversarial attacks on mmWave-based HAR systems, going beyond the prevalent research focus on untargeted attacks. Our unique contribution lies in the design and investigation of practical and universal perturbations to enable targeted adversarial attacks. To ensure wide applicability, these perturbations are generated through an iterative algorithm. Additionally, we affirm the validity of mmWave-based adversarial

samples and shape them into a more natural style. In order to address the information deficiency and scarcity of training data in black-box scenarios, we also incorporate techniques such as knowledge distillation and generative adversarial networks. Through our thorough investigation, we have classified mmWave-based HAR systems into two distinct categories based on their specific data representation. This classification approach enabled us to select representative datasets from each category for applying our designed adversarial attack methodologies. We are confident in this research approach as we believe these methodologies can be extended to a broader range of datasets that share similar data representation characteristics.

Physical adversarial attacks on mmWave-based HAR systems remain a relatively unexplored area of research, and thus, future investigation into physical adversarial attacks will significantly contribute to the understanding of the overall security dynamics of these systems. This understanding will undoubtedly be invaluable in developing more resilient defensive strategies.

References

1. Abdelnasser H, Harras KA, Youssef M (2015) Ubibreathe: a ubiquitous non-invasive wifi-based breathing estimator. In: Proceedings of the 16th ACM international symposium on mobile Ad Hoc networking and computing
2. Abdelnasser H, Youssef M, Harras KA (2015) Wigest: a ubiquitous wifi-based gesture recognition system. In: 2015 IEEE conference on computer communications (INFOCOM), pp 1472–1480
3. Ahuja K, Jiang Y, Goel M, Harrison C (2021) Vid2doppler: synthesizing doppler radar data from videos for training privacy-preserving activity recognition. In: Proceedings of the 2021 CHI conference on human factors in computing systems, pp 1–10
4. Akpa EAH, Fujiwara M, Arakawa Y, Suwa H, Yasumoto K (2018) Gift: glove for indoor fitness tracking system. In 2018 IEEE international conference on pervasive computing and communications workshops (PerCom workshops)
5. Akpa EAH, Fujiwara M, Arakawa Y, Suwa H, Yasumoto K (2018) Gift: glove for indoor fitness tracking system. In: 2018 IEEE international conference on pervasive computing and communications workshops (PerCom workshops), pp 52–57
6. Alam MAU, Rahman MM, Widberg JQ (2021) Palmar: towards adaptive multi-inhabitant activity recognition in point-cloud technology. In IEEE INFOCOM 2021-IEEE conference on computer communications, pp 1–10
7. Ambalkar H, Wang X, Mao S (2021) Adversarial human activity recognition using wi-fi csi. In: 2021 IEEE Canadian conference on electrical and computer engineering (CCECE), pp 1–5
8. Awr1642 data sheet, product information and support | ti.com. <https://www.ti.com/product/AWR1642>. Accessed 12 Jun 2023
9. Bo H, Xu L, Hao L, Dou Y, Zhao L, Yu W (2016) A single-channel non-orthogonal i/q rf sensor for non-contact monitoring of vital signs. Appl Comput Electromagn Soc J 31(6)
10. Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy (sp), pp 39–57
11. Çelikütan O, Akgul CB, Wolf C, Sankur B (2013) Graph-based analysis of physical exercise actions. In: Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare, pp 23–32

12. Chu D, Lane ND, Lai TT-T, Pang C, Meng X, Guo Q, Li F, Zhao F (2011) Balancing energy, latency and accuracy for mobile sensor data classification. In: Proceedings of the 9th ACM conference on embedded networked sensor systems, pp 54–67
13. Ding H, Han J, Shangguan L, Xi W, Jiang Z, Yang Z, Zhou Z, Yang P, Zhao J (2017) A platform for free-weight exercise monitoring with passive tags. *IEEE Trans Mob Comput* 16(12):3279–3293
14. Duan R, Ma X, Wang Y, Bailey J, Qin AK, Yang Y (2020) Adversarial camouflage: hiding physical-world attacks with natural styles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1000–1008
15. Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. *Pers Ubiquitous Comput* 10:255–268
16. Ghorbel E, Boutteau R, Boonaert J, Savatier X, Lecoecue S (2018) Kinematic spline curves: a temporal invariant descriptor for fast action recognition. *Image Vision Comput* 77:60–71
17. Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. Preprint, arXiv:1412.6572
18. Gu F, Chung M-H, Chignell M, Valaee S, Zhou B, Liu X (2021) A survey on deep learning for human activity recognition. *ACM Comput Surv* 54(8):1–34
19. Guo X, Liu J, Shi C, Liu H, Chen Y, Chuah MC (2018) Device-free personalized fitness assistant using wifi. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2(4):1–23
20. Hassana MM, Uddin Z, Mohamed A, Almogrena A (2018) A robust human activity recognition system using smartphone sensors and deep learning. *Future Gener Comput Syst* 81:307–313
21. Hinton G, Vinyals O, Dean J, et al (2015) Distilling the knowledge in a neural network. 2(7). Preprint, arXiv:1503.02531
22. Hui X, Kan EC (2018) Monitoring vital signs over multiplexed radio by near-field coherent sensing. *Nat Electron* 1(1):74–78
23. Hussain Z, Sagar S, Zhang WE, Sheng QZ (2019) A cost-effective and non-invasive system for sleep and vital signs monitoring using passive rfid tags. In: Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, pp 153–161
24. Iso T, Yamazaki K (2006) Gait analyzer based on a cell phone with a single three-axis accelerometer. In: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services, pp 141–144
25. Jin F, Sengupta A, Cao S (2020) mmfall: fall detection using 4-d mmwave radar and a hybrid variational rnn autoencoder. *IEEE Trans Autom Sci Eng* 19(2):1245–1257
26. Kaltiokallio O, Yigitler H, Jäntti R, Patwari N (2014) Non-invasive respiration rate monitoring using a single cots tx-rx pair. In: IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks
27. Khamis A, Chou CT, Kusy B, Hu W (2018) Cardiofi: Enabling heart rate monitoring on unmodified cots wifi devices. In: Proceedings of the 15th EAI international conference on mobile and ubiquitous systems: computing, networking and services, pp 97–106
28. Su C-J, Chiang C-Y, Huang J-Y (2014) Kinect-enabled home-based rehabilitation system using dynamic time warping and fuzzy logic. *Appl Soft Comput* 22:652–666
29. Könönen V, Mäntyjärvi J, Similä H, Pärkkä J, Ermes M (2010) Automatic feature selection for context recognition in mobile devices. *Pervasive Mob Comput* 6(2):181–197
30. Kwon SM, Yang S, Liu J, Yang X, Saleh W, Patel S, Mathews C, Chen Y (2019) Demo: hands-free human activity recognition using millimeter-wave sensors. In: 2019 IEEE international symposium on dynamic spectrum access networks (DySPAN)
31. Laput G, Ahuja K, Goel M, Harrison C (2018) Ubicoustics: plug-and-play acoustic activity recognition. In: Proceedings of the 31st annual ACM symposium on user interface software and technology, pp 213–224
32. Li Z, Wu Y, Liu J, Chen Y, Yuan B (2020) Advpulse: universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. In Proceedings of the 2020 ACM SIGSAC conference on computer and communications security, pp 1121–1134

33. Li G, Ge Y, Wang Y, Chen Q, Wang G (2022) Detection of human breathing in non-line-of-sight region by using mmwave fmcw radar. *IEEE Trans Instrum Meas* 71:1–11
34. Lin Y, Zhao H, Tu Y, Mao S, Dou Z (2020) Threats of adversarial attacks in dnn-based modulation recognition. In *IEEE INFOCOM 2020 - IEEE conference on computer communications*
35. Lin Z, Xie Y, Guo X, Ren Y, Chen Y, Wang C (2020) Wbeat: fine-grained device-free eating monitoring leveraging wi-fi signals. In *2020 29th international conference on computer communications and networks (ICCCN)*.
36. Lin Z, Xie Y, Guo X, Wang C, Ren Y, Chen Y (2020) Wi-fi-enabled automatic eating moment monitoring using smartphones. In: *IoT technologies for HealthCare: Proceedings of the 6th EAI international conference, HealthyIoT 2019, Braga, Portugal, December 4–6, 2019*, pp 77–91
37. Liu H, Gan Y, Yang J, Sidhom S, Wang Y, Chen Y, Ye F (2012) Push the limit of wifi based localization for smartphones. In: *Proceedings of the 18th annual international conference on mobile computing and networking (ACM MobiCom)*.
38. Liu J, Wang Y, Chen Y, Yang J, Chen X, Cheng J (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*.
39. Liu H, Wang Y, Zhou A, He H, Wang W, Wang K, Pan P, Lu Y, Liu L, Ma H (2020) Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 4(4):1–28
40. López-Medina MA, Espinilla M, Cleland I, Nugent C, Medina J (2020) Fuzzy cloud-fog computing approach application for human activity recognition in smart homes. *J Intell Fuzzy Syst* 38(1):709–721
41. Mahmood Khan U, Kabir Z, Hassan SA (2017) Wireless health monitoring using passive wifi sensing. In: *2017 13th international wireless communications and mobile computing conference (IWCMC)*
42. Meng Y, Yi S-H, Kim H-C (2019) Health and wellness monitoring using intelligent sensing technique. *J Inf Process Syst* 15(3):478–491
43. Moosavi-Dezfooli S-M, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 2574–2582
44. Moosavi-Dezfooli S-M, Fawzi O, Frossard P (2017) Universal adversarial perturbations. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1765–1773
45. Mpitiopoulos A, Gavlas D, Konstantopoulos C, Pantziou G (2009) A survey on jamming attacks and countermeasures in wsns. *IEEE Commun Surv Tutor* 11(4):42–56
46. Mun M, Estrin D, Burke J, Hansen M (2008) Parsimonious mobility classification using gsm and wifi traces. In: *Proceedings of the fifth workshop on embedded networked sensors (HotEmNets)*, pp 1–5
47. Oguntala GA, Abd-Alhameed RA, Ali NT, Hu Y, Noras JM, Eya NN, Elfergani I, Rodriguez J (2019) Smartwall: novel rfid-enabled ambient human activity recognition using machine learning for unobtrusive health monitoring. *IEEE access*
48. Ozbulak U, Vandersmissen B, Jalalvand A, Couckuyt I, Van Messem A, De Neve W (2021) Investigating the significance of adversarial attacks and their relation to interpretability for radar-based human activity recognition systems. *Comput Vis Image Underst* 202:103111
49. Palipana S, Salami D, Leiva LA, Sigg S (2021) Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 5(1):1–27
50. Pegoraro J, Meneghello F, Rossi M (2020) Multiperson continuous tracking and identification from mm-wave micro-doppler signatures. *IEEE Trans Geosci Remote Sens* 59(4):2994–3009
51. Pries R, Yu W, Fu X, Zhao W (2008) A new replay attack against anonymous communication networks. In *2008 IEEE international conference on communications*, pp 1578–1582
52. Pu Q, Gupta S, Gollakota S, Patel S (2013) Whole-home gesture recognition using wireless signals. In: *Proceedings of the 19th annual international conference on Mobile computing & networking*, pp 27–38

53. Qin Y, Carlini N, Cottrell G, Goodfellow I, Raffel C (2019) Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: International conference on machine learning, pp 5231–5240
54. Rachuri KK, Musolesi M, Mascolo C, Rentfrow PJ, Longworth C, Aucinas A (2010) Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In: Proceedings of the 12th ACM international conference on Ubiquitous computing (ACM Ubicomp)
55. Ren Y, Lu J, Beletchi A, Huang Y, Karmanov I, Fontijne D, Patel C, Xu H (2021) Hand gesture recognition using 802.11 ad mmwave sensor in the mobile device. In: 2021 IEEE wireless communications and networking conference workshops (WCNCW)
56. Rong Y, Bliss DW (2018) Direct rf signal processing for heart-rate monitoring using uwb impulse radar. In: 2018 52nd asilomar conference on signals, systems, and computers
57. Santhalingam PS, Hosain AA, Zhang D, Pathak P, Rangwala H, Kushalnagar R (2020) mmasl: Environment-independent asl gesture recognition using 60 ghz millimeter-wave signals. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 4(1):1–30
58. Sengupta A, Jin F, Zhang R, Cao S (2020) mm-pose: real-time human skeletal posture estimation using mmwave radars and CNNs. *IEEE Sensors J* 20(17):10032–10044
59. Shastri A, Valecha N, Bashirov E, Tataria H, Lentmaier M, Tufvesson F, Rossi M, Casari P (2022) A review of millimeter wave device-based localization and device-free sensing technologies and applications. *IEEE Commun Surv Tutor* 24(3):1708–1749
60. Sheen DM, McMakin DL, Hall TE (2007) Near field imaging at microwave and millimeter wave frequencies. In: 2007 IEEE/MTT-S international microwave symposium.
61. Shi C, Lu L, Liu J, Wang Y, Chen Y, Yu J (2022) mpose: Environment-and subject-agnostic 3d skeleton posture reconstruction leveraging a single mmwave device. *Smart Health* 23:100228
62. Sim JM, Lee Y, Kwon O (2015) Acoustic sensor based recognition of human activity in everyday life for smart home services. *Int J Distrib Sensor Netw* 11(9):679123
63. Singh AD, Sandha SS, Garcia L, Srivastava M (2019) Radhar: human activity recognition from point clouds generated through a millimeter-wave radar. In: Proceedings of the 3rd ACM workshop on millimeter-wave networks and sensing systems, pp 51–56
64. Sozinov K, Vlassov V, Girdzijauskas S (2018) Human activity recognition using federated learning. In: 2018 IEEE ISPA/IUCC/BDCloud/SocialCom/SustainCom, pp 1103–1111
65. Steele BG, Belza B, Cain K, Warms C, Coppersmith J, Howard J, et al (2003) Bodies in motion: monitoring daily activity and exercise with motion sensors in people with chronic pulmonary disease. *J Rehabil Res Develop* 40(5; SUPP/2):45–58
66. Sun Z, Balakrishnan S, Su L, Bhuyan A, Wang P, Qiao C (2021) Who is in control? practical physical layer attack and defense for mmwave-based sensing in autonomous vehicles. *IEEE Trans Inf Forensics Secur* 16:3199–3214
67. Sze V, Chen Y-H, Yang T-J, Emer JS (2017) Efficient processing of deep neural networks: A tutorial and survey. *Proc IEEE* 105(12):2295–2329
68. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. Preprint, arXiv:1312.6199
69. Texas Instruments (2024) DCA1000 Evaluation Module for Real-Time Data Capture and Streaming. <https://www.ti.com/tool/DCA1000EVM>. Accessed 23 Mar 2024
70. Texas Instruments (2024) IWR1443 BoosterPack™ evaluation module for single-chip 77GHz mmWave sensor. <https://www.ti.com/tool/IWR1443BOOST>. Accessed 23 Mar 2024
71. Tiwari G, Gupta S (2021) An mmwave radar based real-time contactless fitness tracker using deep cnns. *IEEE Sensors J* 21(15):17262–17270
72. Wang Y, Liu J, Chen Y, Gruteser M, Yang J, Liu H (2014) E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures. In: Proceedings of the 20th annual international conference on mobile computing & networking.
73. Wang Y, Liu J, Chen Y, Gruteser M, Yang J, Liu H (2014) E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures. In: Proceedings of the 20th annual international conference on Mobile computing and networking, pp 617–628

74. Wang X, Yang C, Mao S (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)
75. Wang F, Feng J, Zhao Y, Zhang X, Zhang S, Han (2019) Joint activity recognition and indoor localization with wifi fingerprints. *IEEE Access* 7:80058–80068
76. Wang J, Zhao Y, Ma X, Gao Q, Pan M, Wang H (2020) Cross-scenario device-free activity recognition based on deep adversarial networks. *IEEE Trans Veh Technol* 69(5):5416–5425
77. Wang J, Zhang L, Wang C, Ma X, Gao Q, Lin B (2020) Device-free human gesture recognition with generative adversarial networks. *IEEE Internet Things J* 7(8):7678–7688
78. Wang Y, Liu H, Cui K, Zhou A, Li W, Ma H (2021) m-activity: accurate and real-time human activity recognition via millimeter wave radar. In ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP).
79. Wu C, Zhang F, Wang B, Liu KJR (2020) mmtrack: passive multi-person localization using commodity millimeter wave radio. In: IEEE INFOCOM 2020
80. Xia K, Wang H, Xu M, Li Z, He S, Tang Y (2020) Racquet sports recognition using a hybrid clustering model learned from integrated wearable sensor. *Sensors* 20(6):1638
81. Xie Y, Shi C, Li Z, Liu J, Chen Y, Yuan B (2020) Real-time, universal, and robust adversarial attacks against speaker recognition systems. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1738–1742
82. Xie Y, Jiang R, Guo X, Wang Y, Cheng J, Chen Y (2022) mmeat: Millimeter wave-enabled environment-invariant eating behavior monitoring. *Smart Health* 23:100236
83. Xie Y, Jiang R, Guo X, Wang Y, Cheng J, Chen Y (2022) mmfit: Low-effort personalized fitness monitoring using millimeter wave. In: 2022 international conference on computer communications and networks (ICCCN), pp 1–10
84. Xue H, Ju Y, Miao C, Wang Y, Wang S, Zhang A, Su L (2021) mmmesh: towards 3d real-time dynamic human mesh construction using millimeter-wave. In: Proceedings of the 19th annual international conference on mobile systems, applications, and services
85. Yang Z, Pathak PH, Zeng Y, Liran X, Mohapatra P (2016) Monitoring vital signs using millimeter wave. In: Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing, pp 211–220
86. Yang X, Liu J, Chen Y, Guo X, Xie Y (2020) Mu-id: multi-user identification through gaits using millimeter wave radios. In: IEEE INFOCOM 2020-IEEE conference on computer communications, pp 2589–2598
87. Yang Z, Zhao Y, Yan W (2020) Adversarial vulnerability in doppler-based human activity recognition. In: 2020 international joint conference on neural networks (IJCNN), pp 1–7
88. Zeng Y, Pathak PH, Yang Z, Mohapatra P (2016) Poster abstract: human tracking and activity monitoring using 60 ghz mmwave. In: 2016 15th ACM/IEEE international conference on information processing in sensor networks (IPSN)
89. Zhang J, Wu F, Wei B, Zhang Q, Huang H, Shah SW, Cheng J (2020) Data augmentation and dense-lstm for human activity recognition using wifi signal. *IEEE Internet Things J* 8(6):4628–4641
90. Zhang L, Hua Y, Cotton SL, Yoo SK, Da Silva CRCM, Scanlon WG (2020) An rss-based classification of user equipment usage in indoor millimeter wave wireless networks using machine learning. *IEEE Access* 8:14928–14943
91. Zhao P, Lu CX, Wang J, Chen C, Wang W, Trigoni N, Markham A (2019) mid: Tracking and identifying people with millimeter wave radar. In: 2019 15th international conference on distributed computing in sensor systems (DCOSS), pp 33–40
92. Zhou B, Yang J, Li Q (2019) Smartphone-based activity recognition for indoor localization using a convolutional neural network. *Sensors* 19(3):621