Universal Targeted Adversarial Attacks Against mmWave-based Human Activity Recognition

Yucheng Xie*, Ruizhe Jiang*, Xiaonan Guo†, Yan Wang‡, Jerry Cheng§, Yingying Chen¶
*Indiana University-Purdue University Indianapolis, IN 46202, USA
†George Mason University, VA 22030, USA
‡Temple University, Philadelphia, PA 19122, USA
§New York Institute of Technology, New York, NY 10023, USA
¶Rutgers University, New Brunswick, NJ 08901, USA
Email: *{yx11, ruizjian}@iupui.edu
, †xguo8@gmu.edu, ‡y.wang@temple.edu, §jcheng18@nyit.edu, ¶yingche@scarletmail.rutgers.edu

IEEE INFOCOM 2023 - IEEE Conference on Computer Communications | 979-8-3503-3414-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/INFOCOM53939.2023.10228887

***Abstract*—Human activity recognition (HAR) systems based on millimeter wave (mmWave) technology have evolved in recent years due to their better privacy protection and enhanced sensor resolution. With the ever-growing HAR system deployment, the vulnerability of such systems has been revealed. However, existing efforts in HAR adversarial attacks only focus on untargeted attacks. In this paper, we propose the first targeted adversarial attacks against mmWave-based HAR through designed universal perturbation. A practical iteration algorithm is developed to craft perturbations that generalize well across different activity samples without additional training overhead. Different from existing work that only develops adversarial attacks for a particular mmWave-based HAR model, we improve the practicability of our attacks by broadening our target to the two most common mmWave-based HAR models (i.e., voxel-based and heatmap-based). In addition, we consider a more challenging black-box scenario by addressing the information deficiency issue with knowledge distillation and solving the insufficient activity sample with a generative adversarial network. We evaluate the proposed attacks on two different mmWave-based HAR models designed for fitness tracking. The evaluation results demonstrate the efficacy, efficiency, and practicality of the proposed targeted attacks with an average success rate of over 90%.**

***Index Terms*—Millimeter Wave, Human Activity Recognition, Adversarial Learning, Universal Targeted Attack, Black-box Attack**

I. Introduction

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 developed using various sensing modalities. Traditional camera-based [1], [2] and sensor-based [3], [4] HAR systems capture human activities using video cameras and body sensors, respectively. They usually intrigue privacy concerns or are not convenient. Recently, wireless signals (e.g., WiFi [5], [6], sound [7], [8], mmWave [9], [10]) have been proposed 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 [11]. Some researchers have proposed introducing minor perturbations that cause deep learning networks to make inaccurate predictions in image classification [12] and voice recognition [13]. Nevertheless, few studies have investigated the susceptibility of adversarial targeted attacks in mmWave-based HAR systems. Because mmWave-based HAR systems are usually integrated in many crucial applications such as older patients monitoring and user authentication [14], [15], we believe that studying adversarial attacks on these systems is critical and urgent. Most recently, Ozbulak et al. [16] have done an initial investigation with the untargeted adversarial attack on mmWave-based HAR. The proposed 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 [17], how to launch universal target adversarial attacks [18], or more challenging black-box attacks [19], are still worth further 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 work, we aim to systematically investigate and reveal the severe security issues of mmWave-based HAR models by developing the following effective adversarial attacks: *(1) Targeted and Untargeted attacks.* Unlike existing work that only studied the untargeted attack for a particular mmWave-base HR model, we successfully design both targeted and untargeted attacks for different mmWave-based HAR models. *(2) Universal Attack.* Both targeted and untargeted attacks need to train a unique adversarial perturbation for each activity sample [20], 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, which is ready to be used in real-time without additional training; *(3) Black-box Attack.* The existing adversarial attacks against mmWave-based HAR assume white-box settings, wherein the attacker has full knowledge

of the target model, including architecture and parameters. However, attackers may not have such information and need to conduct attacks under more realistic conditions (e.g., the target model is unavailable to the attacker). Therefore, we develop an effective method to enable black-box targeted attacks in such challenging scenarios.

Designing effective and practical adversarial attacks for different mmWave-based HAR models is nontrivial. Different from traditional replay attacks [21], 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 propose 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 for designing the universal adversarial attack is deriving an effective adversarial perturbation for any activity sample without online training. We propose 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 [22], 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 target 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.

We summarize the main contributions of this work as follows:

- We propose a comprehensive assessment of the challenges brought by adversarial attacks on various mmWave-based HAR systems, including both untargeted and targeted attacks. As far as we know, we are the first to implement targeted attacks against mmWave-based HAR systems, especially for voxel-based mmWave models.
- 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. We leverage knowledge distillation to address the information deficiency of the target model. We also develop a generative adversarial network to address the lack of training data.
- We assess our proposed attack methods on two representative mmWave-based HAR models and demonstrate the efficacy, efficiency, and practicality of the proposed attacks with a high attack success rate of over 90%.

II. Related Work

Because of its wide application, HAR has attracted great attention for the past decade. Many HAR systems use cameras, body sensors, acoustic sound, and WiFi signals to recognize and track human activities [5], [6]. Recently, due to the high resolution and bandwidth, mmWave has been proposed to perform HAR [9], [10], [14]. 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.* [11]. We discover that the majority of current adversarial attacks are proven in applications related to image recognition and speech authentication [12], [18], [23]–[26]. However, it has seldom been investigated how adversarial attacks will affect HAR systems based on mmWave. Yang *et al.* [20] 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.* [22] examine the vulnerability of radar-based HAR system 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 universal targeted attack on other types of mmWave-base HAR models are unknown. Besides, based on unique patterns of mmWave activity data, how to develop adversarial activity samples to assure their validity and make them unnoticeable are necessary but seldom be 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 propose 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.

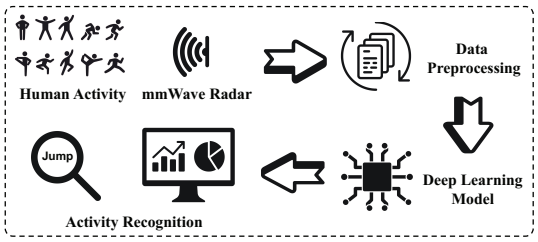


Fig. 1: Framework of mmWave-based human activity recognition.

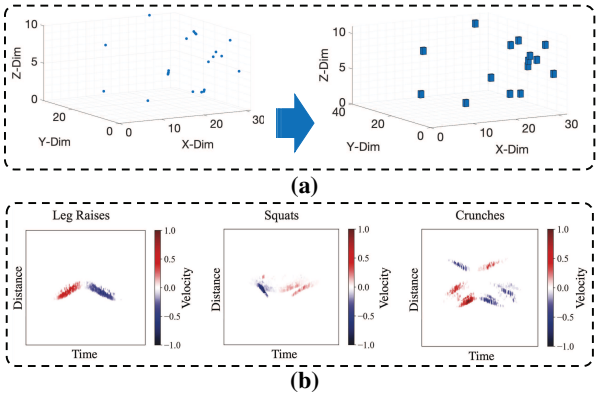


Fig. 2: Two typical data representations for mmWave-based HAR. (a) Voxels generation from the point cloud; (b) Spatial-Temporal heatmaps of three different activities.

III. Target Machine Learning Models for HAR

Background. The main goal of the mmWave-based human activity recognition system is to identify actions or gestures by examining the dynamics of mmWave signals [10], [14], [27]. As shown in Figure 1, a typical mmWave-based HAR system captures mmWave signals reflected from the human body. It performs signal processing to determine activity characteristics (e.g., velocity or posture) of users and then estimate the activity class using deep learning models.

Existing mmWave-based HAR systems are usually based on two different 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 [9], [14], [15], [28], [29], 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 [9], [14] 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 [10], micro-Doppler map [30], spatial spectrograms [27], spatial feature map [31], and projection heatmap [32]) 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 work, we investigate the attacks to two mmWave-based HAR models using different types of representations, which brings more challenges to design a generic attack method because of their significant differences.

Voxel-based Machine Learning Model. We choose an existing mmWave-based HAR system [9] as a representative to study the vulnerability of voxel-based HAR model to adversarial attacks. This model has been utilized as a benchmark in numerous subsequent publications [33], [34]. In particular, the point clouds data is subjected to voxelization to address the non-uniformity issue in each frame, as shown in Figure 2a. 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 number of cloud points in the cube-shape 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 as state-of-the-art mmWave-based HAR methods (e.g., [10]), 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 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 Figure 2b. 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.

IV. Threat Model and Problem Formalization

A. Threat Model

Possible Attack Scenarios. Our attack is applicable to attack scenarios where the attacker needs to create an adversarial example offline and then insert it to the HAR systems at the inference stage. Three insert points are taken into account for our potential adversarial attacks to HAR systems. Firstly, attackers might modify the adversarial samples during the data preprocessing stage. For example, as shown in Figure 1, a local adversary can launch attacks by modifying the generated voxel-based or heatmap-based activity sample directly. Moreover, it is also possible to insert the adversarial sample

right before the recognition phase, where activity data are sent as an input to the DNN model. In this case, the attackers can generate adversarial samples in advance and fed them into the machine learning models furtively. Furthermore, 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 [35], [36].

Adversary Capability. The attacks can be classified as white-box and black-box attacks. 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 target 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. The possibility of a white-box attack may be increased by a local adversary or information leakage. In order to study adversarial attacks to mmWave-based HAR, we first make the white-box assumption as most previous studies [18], [20], [37]. 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 [19]. We investigate our adversarial attacks on HAR in black-box settings, because they are more realistic than white-box settings.

B. Problem Formalization

The adversarial goal of our work 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 output 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. We formalize various adversarial attacks as following:

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 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 context, we propose to 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.

Practical Black-box Attack. In order to launch adversarial attack in practical, the attack method should also be robust enough to work in more challenging scenarios, such as attacking a black-box machine learning model. Specifically, we would explore whether a perturbation generated based on model f could still work on other 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.

Unnoticeable Perturbation. In order to make the adversarial sample practical and it difficult for human to identify the attack, the distortion brought on by the perturbation should be as small as feasible. 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 $\text{minimize}\, \mathcal{D}(x, x + \delta)$, such that $f(x + \delta) = z$, where z can be different based on the attack objective (i.e., untargeted attack or targeted attack). D is the distance metrics $\|\delta\|_p$ that evaluate the magnitude of the generated perturbation. However, as discussed in previous work [24], directly solve this non-linear constrained non-convex problem is difficult. Thus we reformulate the objective function as a gradient-based optimization instance:

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

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.

V. Methodology

In this section, we discuss the detailed adversarial attack algorithms to enable targeted attack and untargeted attack for mmWave-based HAR systems. We also introduce how we balance the trade off between attack effectiveness and perturbation magnitude. We then analyze the unique characteristics of mmWave activity samples and provide two distance metrics

to further optimize the perturbation. Moreover, we explore an efficient and powerful universal attack approach for mmWave-based HAR systems. In addition, we study the feasibility of launching adversarial attacks on black-box scenarios.

A. Untargeted and Targeted Attack for Specific Samples

Adversarial Loss. For sample-specific untargeted attack, we define $\mathcal{L} = \max\left(\mathcal{Z}(x+\delta)_s - \max_{i \neq s}\left(\mathcal{Z}(x+\delta)_i\right), -k\right)$, 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 which controls attack confidence. For sample-specific targeted attack, we define $\mathcal{L} = \max\left(\max_{i \neq t}\left(\mathcal{Z}(x+\delta)_i\right) - \mathcal{Z}(x+\delta)_t, -k\right)$, where $\mathcal{Z}(x+\delta)_t$ is the possibility of estimating the activity as the class t we desired. By optimizing above adversarial loss functions, we aims to make our attack method not only confuse the HAR systems (untargted attack), but also force the HAR system output our desired class (targeted attack). In practice, by using different special-designed adversarial loss function, the attacker could either launch untargeted attack or targeted attack according to different attack strength requirements, which makes our attack framework more powerful and dangerous than previous studies [20], [22].

Perturbation Loss. Generally speaking, the perturbations are the difference between the original activity 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 [16], [24], [25]. 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 activity sample to ensure $\|\delta\|_2^2 < \tau,\ \text{s.t. } f(x+\delta) = z$. The threshold is derived by analyzing the deviation between the normal activity sample from 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 practical, we do a 12-step binary search to identify the appropriate λ and its accompanying adversarial perturbation δ.

B. Optimization for Unnoticeable Perturbation

Clipping. In order to assure 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., [9]), 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., [30]), the range is usually set to $[0, 255]$.

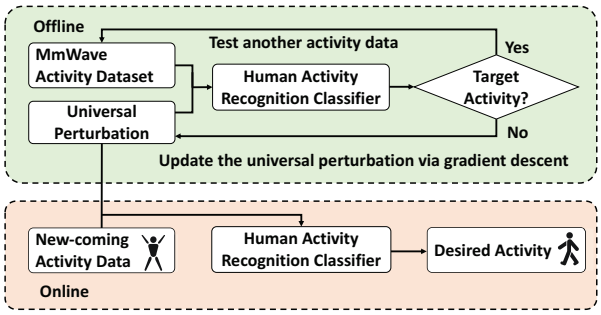


Fig. 3: Overview of universal attack.

Discretization. Discretization is a crucial processing that usually be neglected in prior research [16], [23], [25]. 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 previous method that simply rounding the value of each adversarial voxel or heatmap to the nearest integer could eliminate the 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 Section V-A 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 assure 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 [16], [19], [20], [24] only focus on minimizing perturbation magnitude. How to optimize adversarial activity samples into natural styles by constraining the form and position of the generated perturbation is less studied or neglected. In this study, we suggest minimizing the radius of the generated perturbation to make it unnoticeable in heatmaps or voxels. We realize it by minimizing 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 pair-wise 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*

Algorithm 1: Universal Perturbation Generation

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

Output: Universal perturbation δ.

Initialize $\delta \leftarrow 0$.
while Success Rate$(\Omega) < \epsilon$ **do**
  $\Omega_j \leftarrow RS(\Omega)$ ▷ Random Select a Sample
  if $f(x_j + \delta) = z$ **then**
    Calculate the perturbation that satisfies: $\delta \leq \tau$.
  else
    $\Delta\delta_j \leftarrow \arg\min_{\Delta\delta_j} \mathcal{D}(\Delta\delta_j)$ such that $f(\Omega_j + \delta + \delta_j) = z$.
    $\delta \leftarrow (\delta + \Delta\delta_j)$. ▷ Update the Perturbation
  end if
end while

Distance, the inserted perturbation is pushed nearer the activity sample. After integrating 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 Section V-A.

C. Universal Perturbation Generation

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 proposed 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 perturbation is directly applied to incoming activity data for a targeted attack. As shown in Figure 3, 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 (Equation 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 a 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 process ends when the universal perturbation on the training dataset exceeds an predefined success rate (e.g., 70%). Notably, the objective of the technique is not to seek

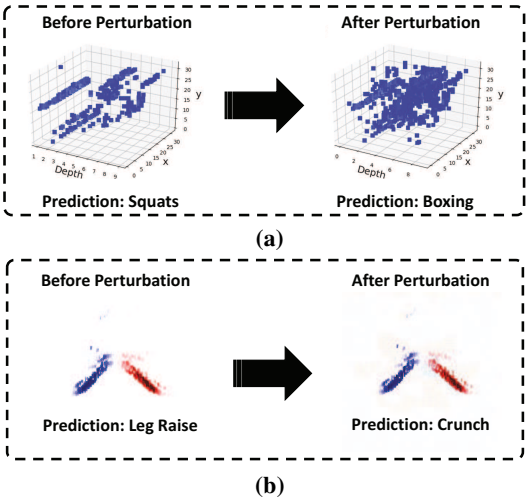


Fig. 4: Two representative adversarial samples generated by adding universal perturbations directly. (a) Adversarial voxel-based data generation with L2 Norm of 21; (b) Adversarial heatmap-based data generation with L2 Norm of 2083.

the smallest global perturbation that fools the majority of activity samples, but rather to select one that is sufficiently tiny. Figure 4 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 L2 Norm. The adversarial samples looks natural which makes it hard to be noticed by naked eyes. However, adversarial examples enable HAR systems to efficiently predict the activity as our desired. Compared with traditional sample-specific attack methods, our universal perturbations would significantly shorten the attack launch time, which make it more practical in time-constraint attack scenarios. Different from existing universal untargeted attack method that need to inserting padding frames between two successive activities [22], our method modifies the activity sample directly and thus broaden its applicability to various of mmWave-based HAR systems. In addition, our universal attack method is compatible to both untargeted and targeted attacks by merely modifying the adversarial loss function.

D. Practical Attack in Black-box Scenario

Black-box attack is a more challenging scenario where the attacker usually cannot access the target model but only the input and output of the model [22]. Thus, the adversarial perturbation cannot be created and updated by exploiting the gradient from the target model. A potential approach of black-box attack is to train a substitute model. Adversarial samples generated by the substitute model can be exploited to launch attacks towards the target model leveraging the transferability of the adversarial sample.

We begin with a basic black-box setting where the training data of the target 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 get poor performance since the structure

of the substitute model is different [22]. To solve such a problem, we take advantage of Knowledge Distillation (KD) to learn a substitute model that can mimic the prediction of the target model [38]. In black-box scenarios, although the inner structure of the target model is inaccessible, it's output class and soft logits indicating the class probability distribution for a given input is accessible [19]. Supposing the soft logits of the target and substitute model is P_t and P_s, respectively. The predicted class of substitute model is Z and the ground truth is G. We formulate KD process as loss function $L = L_s + L_d$, where $L_s = Cross\_Entropy(Z, G)$ and $L_d = KL\_Divergence(P_t, P_s)$. By optimizing the loss function, we transfer the dark knowledge from the target model to the substitute model [38].

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 Section V-A. With larger k value, the possibility that the adversarial sample being misclassified by the target 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 Section VI-D.

We then move to a more challenging scenario where the original training data of the target model is only partially accessible. To deal with the problem of insufficient training data, we develop GAN to synthesize sufficient pseudo training samples. GAN has been proved to generate high-quality pseudo samples with limited amount of real samples [30]. In this work, 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 target model. Specifically, the generator seeks to learn the distribution of the real samples so as to have the ability of synthesizing pseudo sample. 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 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 towards the target model.

VI. Performance Evaluation

A. Experimental Setup

Equipment. Our own dataset (i.e., heatmap-based) is collected using TI AWR1642 mmWave radar [39], while the public dataset (i.e., voxel-based) is collected using IWR1443 mmWave radar [9]. Both mmWave radar work at the frequency in the range of $77 \sim 81GHz$. The prototype of our proposed attack method is implemented using Python along with TensorFlow.

Data Collection. Two datasets are used in our experiment. The public voxel-based human activity dataset contains 15635 samples from 5 different activities. Our own heatmap-based fitness activity dataset contains 8760 samples from 14 typical workouts. Both dataset is splitted into training and testing sets with a ratio of 7 to 3. For untargeted and targeted attack, we randomly select 200 and 100 activity samples of each type of activity from the voxel-based and heatmap-based testing set, respectively. For the universal attack, half of the selected samples are utilized for universal perturbation generation (universal attack training set) and the others for evaluation (universal attack testing set), respectively. The original classification accuracy of voxel-based and heatmap-based machine learning model is 90.47% and 97%, respectively.

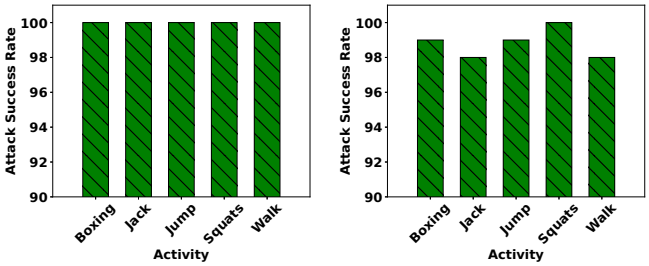


Fig. 5: (a) Success rate of sample-specific untargeted attacks on voxel-based dataset; (b) Success rate of sample-specific targeted attacks on voxel-based dataset (x-axis represents the original classes).

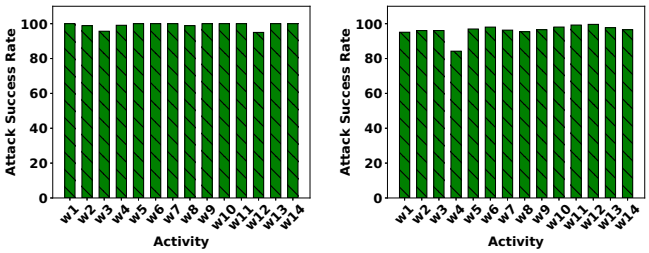
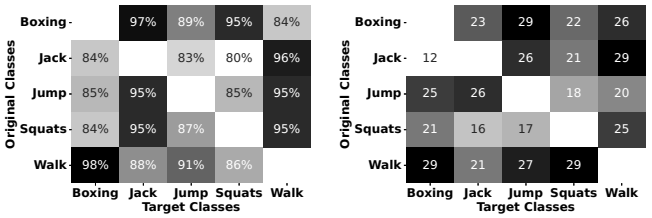


Fig. 6: (a) Success rate of sample-specific untargeted attacks on heatmap-based dataset; (b) Success rate of sample-specific targeted attacks on heatmap-based dataset (x-axis represents the original classes).

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 untargeted attack, we report a success when the predicted class is different from the original class while in targeted attack, we only reported a success if the predicted class matches the desired target class; (2) *L2 Norm:* it indicates the euclidean distance between the adversarial sample and original sample; Smaller L2 Norm values indicate that the adversarial sample is similar to the original activity sample and 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 represent the average SR and L2 Norm of corresponding universal attack on the testing set.

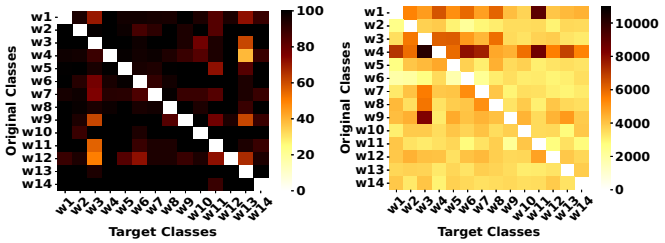
B. Evaluation of Attack Effectiveness

Untargeted Attack. Figure 5a and Figure 6a demonstrate the attack success rate of untargeted attacks on the voxel-based



(a) (b)

Fig. 7: (a) Success rate of universal targeted attacks on voxel-based dataset; (b) L2 Norm of generated universal perturbations on voxel-based dataset.



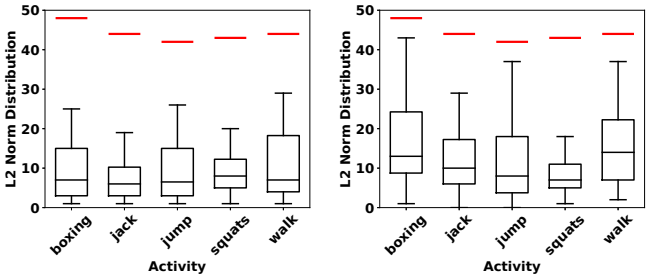
(a) (b)

Fig. 8: (a) Success rate of universal targeted attacks on heatmap-based dataset; (b) L2 Norm of generated universal perturbations on heatmap-based dataset.

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.

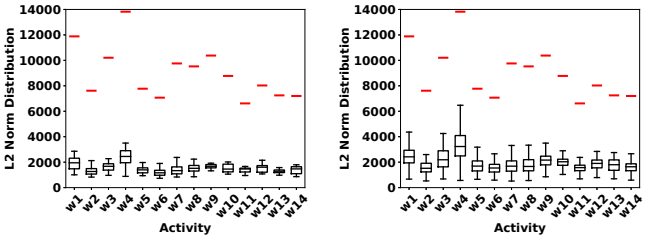
Targeted Attack. Figure 5b and Figure 6b demonstrate the attack success rate of sample-specific targeted attacks on the two dataset, respectively. Our method achieves an average SR of 96% on both datasets. Note that attacks towards some target classes have relatively lower SR (i.e., jack and walk from the voxel dataset; and $w4$ (lunges) from the heatmap-based dataset), this is because those classes have more different patterns from others, making the attack relatively harder. But even the lowest SR in targeted attack is still higher than 80%, proving the effectiveness of our attack scheme.

Universal Attack. The performance of universal attacks over the voxel-based HAR dataset is demonstrated in Figure 7a. 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 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 Figure 8a, 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.



(a) (b)

Fig. 9: (a) L2 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) L2 Norm of perturbations generated in sample-specific targeted attacks on voxel-based dataset (x-axis represents the original class).



(a) (b)

Fig. 10: (a) L2 Norm of perturbations generated in sample-specific untargeted attacks on heatmap-based dataset; (b) L2 Norm of perturbations generated in sample-specific targeted attacks on heatmap-based dataset.

C. Impact of Perturbation Magnitude

Untargeted Attack. We first evaluate the impact of perturbation magnitude on untargeted attack. Figure 9a demonstrates the L2 Norm of untargeted attacks on the voxel-based dataset. Note that each red line indicate the average value of the threshold mentioned in Section V-A. We can learn that the median L2 Norm of untargeted adversarial samples on all 5 original classes are below 10 and the maximum L2 Norm values are all lower than 30, far below the 5 average thresholds, which are all around $40 \sim 50$. For the heatmap-based dataset, as is shown in Figure 10a, the median L2 Norms between adversarial and original samples are around $2000 \sim 2500$. Adversarial samples towards one workout, $w4$, have relatively higher L2 Norm distribution due to high-specific features of original heatmaps. But the highest L2 Norm is still lower than 4000, far below the average threshold of 14000 for $w4$.

Targeted Attack. We then evaluate the impact of perturbation magnitude on targeted attack. Figure 9b demonstrates the L2 Norm of targeted attacks on the voxel-based dataset. We find that the median L2 Norm values of samples towards all 5 target classes are still around 20. But there is a significant increase in the maximum L2 Norm value on all 5 classes compared with untargeted attacks. This is because in targeted attack, we not only need to flip the class but also need to turn the class to the required one. Thus, for some samples larger perturbation magnitude is needed. But even the maximum

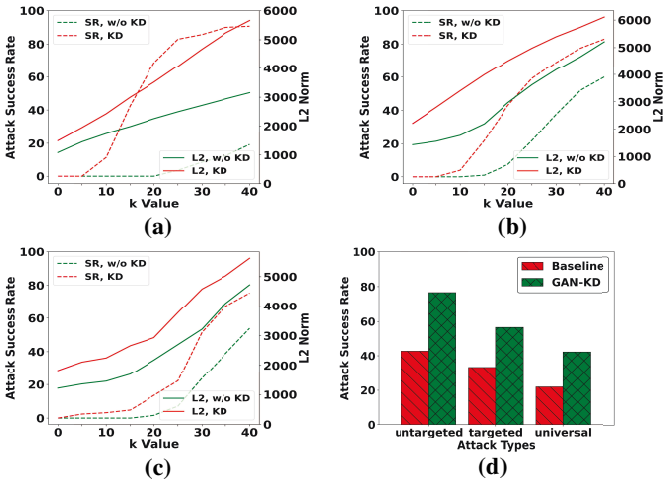


Fig. 11: (a) Success rate and L2 Norm of untargeted black-box attack; (b) Success rate and L2 Norm of targeted black-box attack; (c) Success rate and L2 Norm of universal black-box attack; (d) Success rate of GAN-KD-based black-box attack.

L2 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 Figure 10b, we also notice a larger L2 Norm distribution compared with the result of untargeted attacks. Samples aiming at the target class of $w4$ have a relatively higher maximum L2 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., $w4$) still does not exceed the corresponding average threshold.

Universal Attack. We next evaluate the impact of perturbation magnitude on universal attack. The confusion matrix of universal L2 Norm on voxel-based dataset and heatmap-based dataset are shown in Figure 7b and 8b, respectively. We can learn that the average L2 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 $40 \sim 50$). The average L2 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 difference in heatmap patterns, these values are still below the average threshold of corresponding original classes.

D. Evaluation of Black-box Attack

All black-box experiments are taken on the heatmap-based dataset due to the page limit. We begin with the basic settings where the target model 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 target 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 Section V-D, 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. Figure 11a, 11b and 11c demonstrate the average SR and L2 Norm under basic black-box settings for untargeted, targeted and universal attacks, respectively. We can learn that substitute model trained with KD outperforms directly-trained model for all k larger than 0 in all types of attacks. When $k = 40$, attacks using substitute model achieves higher than 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, L2 Norm and k values. As the k increases, we can obtain higher SR but the L2 Norm will also increase accordingly, meaning the adversarial samples will have relatively larger perturbations. But our method still maintain lower than 6500 L2 Norm value for all three types of attacks even when $k = 40$.

We next move to a more challenge setting where the adversary can only access part of the training data used by the target model. We exploit the GAN method mentioned in section V-D 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 [22]. As is shown in Figure 11d, 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 GAN-KD method are relatively low (i.e., 56% and 42%), but the SR still outperforms baseline model with an increase of 23.4% and 20.22%, respectively.

VII. Conclusion

In this paper, we propose a comprehensive study of adversarial attacks against mmWave-based HAR. Unlike existing work that only explore the feasibility of untargeted attacks, we are the first to design and investigate universal, yet practical perturbations to enable targeted adversarial attack against various mmWave-based HAR. We generate universal perturbation via an iteration algorithm to make it generalizes very well across different activity samples. We also assure the validity of mmWave-based adversarial sample and tailor them into natural style. In addition, we develop KD to address the information deficiency of the machine learning model for HAR and GAN to address the lack of training data in black-box scenarios. Extensive experiments on two typical mmWave-based HAR models demonstrate the efficacy, efficiency, and practicality of the proposed targeted attacks with an average success rate of over 90%.

VIII. Acknowledgment

This work was partially supported by the National Science Foundation Grants CNS2304766, CNS2145389, CNS2120276, CCF2000480, CNS2114220, CCF1909963, CCF2211163, CNS2120396, CCF2028873, CCF1909963, CNS2120350.

REFERENCES

[1] O. Çeliktutan, C. B. Akgul, C. Wolf, and B. Sankur, "Graph-based analysis of physical exercise actions," in *Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare*, 2013, pp. 23–32.

[2] E. Ghorbel, R. Boutteau, J. Boonaert, X. Savatier, and S. Lecoeuche, "Kinematic spline curves: A temporal invariant descriptor for fast action recognition," *Image and Vision Computing*, vol. 77, pp. 60–71, 2018.

[3] E. A. Akpa, M. Fujiwara, Y. Arakawa, H. Suwa, and K. Yasumoto, "Gift: glove for indoor fitness tracking system," in *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2018, pp. 52–57.

[4] Y. Meng, S.-H. Yi, and H.-C. Kim, "Health and wellness monitoring using intelligent sensing technique," *Journal of Information Processing Systems*, vol. 15, no. 3, pp. 478–491, 2019.

[5] X. Guo, J. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, "Device-free personalized fitness assistant using wifi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, pp. 1–23, 2018.

[6] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "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*, 2014, pp. 617–628.

[7] G. Laput, K. Ahuja, M. Goel, and C. Harrison, "Ubicoustics: Plug-and-play acoustic activity recognition," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018, pp. 213–224.

[8] J. M. Sim, Y. Lee, and O. Kwon, "Acoustic sensor based recognition of human activity in everyday life for smart home services," *International Journal of Distributed Sensor Networks*, vol. 11, no. 9, p. 679123, 2015.

[9] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "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*, 2019, pp. 51–56.

[10] H. Liu, Y. Wang, A. Zhou, H. He, W. Wang, K. Wang, P. Pan, Y. Lu, L. Liu, and H. Ma, "Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 4, no. 4, pp. 1–28, 2020.

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[12] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[13] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International conference on machine learning*. PMLR, 2019, pp. 5231–5240.

[14] M. A. U. Alam, M. M. Rahman, and J. Q. Widberg, "Palmar: Towards adaptive multi-inhabitant activity recognition in point-cloud technology," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[15] P. Zhao, C. X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "mid: Tracking and identifying people with millimeter wave radar," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2019, pp. 33–40.

[16] U. Ozbulak, B. Vandersmissen, A. Jalalvand, I. Couckuyt, A. Van Messem, and W. De Neve, "Investigating the significance of adversarial attacks and their relation to interpretability for radar-based human activity recognition systems," *Computer Vision and Image Understanding*, vol. 202, p. 103111, 2021.

[17] Z. Sun, S. Balakrishnan, L. Su, A. Bhuyan, P. Wang, and C. Qiao, "Who is in control? practical physical layer attack and defense for mmwave-based sensing in autonomous vehicles," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3199–3214, 2021.

[18] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, "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*, 2020, pp. 1121–1134.

[19] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of adversarial attacks in dnn-based modulation recognition," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020, pp. 2469–2478.

[20] Z. Yang, Y. Zhao, and W. Yan, "Adversarial vulnerability in doppler-based human activity recognition," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[21] R. Pries, W. Yu, X. Fu, and W. Zhao, "A new replay attack against anonymous communication networks," in *2008 IEEE International Conference on Communications*. IEEE, 2008, pp. 1578–1582.

[22] U. Ozbulak, B. Vandersmissen, A. Jalalvand, I. Couckuyt, A. V. Messem, and W. D. Neve, "Investigating the significance of adversarial attacks and their relation to interpretability for radar-based human activity recognition systems," *Computer Vision and Image Understanding*, vol. 202, p. 103111, jan 2021.

[23] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1000–1008.

[24] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57.

[25] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[26] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "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)*. IEEE, 2020, pp. 1738–1742.

[27] P. S. Santhalingam, A. A. Hosain, D. Zhang, P. Pathak, H. Rangwala, and R. Kushalnagar, "mmasl: Environment-independent asl gesture recognition using 60 ghz millimeter-wave signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–30, 2020.

[28] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su, "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*, 2021.

[29] Y. Wang, H. Liu, K. Cui, A. Zhou, W. Li, and H. Ma, "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)*. IEEE, 2021.

[30] J. Wang, L. Zhang, C. Wang, X. Ma, Q. Gao, and B. Lin, "Device-free human gesture recognition with generative adversarial networks," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7678–7688, 2020.

[31] C. Shi, L. Lu, J. Liu, Y. Wang, Y. Chen, and J. Yu, "mpose: Environment- and subject-agnostic 3d skeleton posture reconstruction leveraging a single mmwave device," *Smart Health*, vol. 23, p. 100228, 2022.

[32] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal*, 2020.

[33] S. Palipana, D. Salami, L. A. Leiva, and S. Sigg, "Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–27, 2021.

[34] K. Ahuja, Y. Jiang, M. Goel, and C. Harrison, "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*, 2021, pp. 1–10.

[35] K. Sozinov, V. Vlassov, and S. Girdzijauskas, "Human activity recognition using federated learning," in *2018 IEEE ISPA/IUCC/BDCloud/SocialCom/SustainCom*. IEEE, 2018, pp. 1103–1111.

[36] M. López-Medina, M. Espinilla, I. Cleland, C. Nugent, and J. Medina, "Fuzzy cloud-fog computing approach application for human activity recognition in smart homes," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 1, pp. 709–721, 2020.

[37] H. Ambalkar, X. Wang, and S. Mao, "Adversarial human activity recognition using wi-fi csi," in *2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2021, pp. 1–5.

[38] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[39] A. Pirkani, S. Pooni, and M. Cherniakov, "Implementation of mimo beamforming on an ots fmcw automotive radar," in *2019 20th International Radar Symposium (IRS)*. IEEE, 2019, pp. 1–8.