TileMask: A Passive-Reflection-based Attack against mmWave Radar Object Detection in Autonomous Driving

Yi Zhu
University at Buffalo, the State University of New York
yzhu39@buffalo.edu

Chenglin Miao
Iowa State University
cmiao@iastate.edu

Hongfei Xue
University of North Carolina at Charlotte
hongfei.xue@charlotte.edu

Zhengxiong Li
University of Colorado Denver
zhengxiong.li@ucdenver.edu

Yunnan Yu
University at Buffalo, the State University of New York
yunnanyu@buffalo.edu

Wenyao Xu
University at Buffalo, the State University of New York
wenyaoxu@buffalo.edu

Lu Su
Purdue University
lusu@purdue.edu

Chunming Qiao
University at Buffalo, the State University of New York
qiao@buffalo.edu

ABSTRACT

In autonomous driving, millimeter wave (mmWave) radar has been widely adopted for object detection because of its robustness and reliability under various weather and lighting conditions. For radar object detection, deep neural networks (DNNs) are becoming increasingly important because they are more robust and accurate, and can provide rich semantic information about the detected objects, which is critical for autonomous vehicles (AVs) to make decisions. However, recent studies have shown that DNNs are vulnerable to adversarial attacks. Despite the rapid development of DNN-based radar object detection models, there have been no studies on their vulnerability to adversarial attacks. Although some spoofing attack methods are proposed to attack the radar sensor by actively transmitting specific signals using some special devices, these attacks require sub-nanosecond-level synchronization between the devices and the radar and are very costly, which limits their practicability in real world. In addition, these attack methods can not effectively attack DNN-based radar object detection. To address the above problems, in this paper, we investigate the possibility of using a few adversarial objects to attack the DNN-based radar object detection models through passive reflection. These objects can be easily fabricated using 3D printing and metal foils at low cost. By placing these adversarial objects at some specific locations on a target vehicle, we can easily fool the victim AV's radar object detection model. The experimental results demonstrate that the attacker can achieve the attack goal by using only two adversarial objects and conceal them as car signs, which have good stealthiness and flexibility. To the best of our knowledge, this is the first study on the passive-reflection-based attacks against the DNN-based radar object detection models using low-cost, readily-available and easily concealable geometric shaped objects.

CCS CONCEPTS

• **Security and privacy → Domain-specific security and privacy architectures**; • **Computer systems organization → Embedded and cyber-physical systems**.

KEYWORDS

Autonomous driving; radar perception; adversarial attack

ACM Reference Format:
Yi Zhu, Chenglin Miao, Hongfei Xue, Zhengxiong Li, Yunnan Yu, Wenyao Xu, Lu Su, and Chunming Qiao. 2023. TileMask: A Passive-Reflection-based Attack against mmWave Radar Object Detection in Autonomous Driving. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), November 26–30, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3576915.3616661

1 INTRODUCTION

Autonomous driving has been advocated as a future trend and many autonomous vehicles (AVs) have been deployed on public roads. The perception system, especially the object detection system, plays a critical role in helping the AVs make driving decisions and ensure road safety. As one of the most important sensors in AV's perception system, radar can work in harsh conditions such as severe weather and lighting conditions, which makes it more robust and reliable than camera and LiDAR. Autonomous vehicles typically use frequency modulated continuous wave (FMCW) radar that transmits millimeter-wave (mmWave) signals and receives the echo signals to detect objects on the roads.

To deal with the received mmWave signals, many signal processing operations such as Fast Fourier Transform (FFT) have been developed [59]. Although the outputs of these FFT operations can provide some raw measurements on the information (e.g., distances and angles) of potential objects, these raw radar measurements can

CCS '23, November 26–30, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0050-7/23/11…$15.00
https://doi.org/10.1145/3576915.3616661

be too noisy to be directly used for object detection in autonomous driving [9, 26, 41]. To deal with these problems and provide more robust and accurate detection results, deep neural networks have been used after the FFT pre-processing step to extract useful features for object detection. These DNN-based radar object detection models have been widely used in autonomous driving applications, and can make the detection results more accurate and more robust to noisy signals [6, 31, 68, 74, 75]. In addition, they can provide rich semantic information, such as the objects class labels (e.g., car, pedestrian or motorcycle), which is critical for AVs' decision making [12, 27, 75]. However, recent studies have demonstrated that DNNs can be fooled by *adversarial attacks* [30, 54], where an attacker can significantly change the outputs by slightly perturbing the model's input. Despite the rapid development of the DNN-based radar object detection models and their critical role in guaranteeing the safety of autonomous vehicles, there have been no studies so far on their vulnerability to adversarial attacks.

Although there have been a few studies showing that the mmWave radar sensors can be spoofed by attackers to manipulate the outputs of raw radar measurements [20, 39, 49, 69], these attacks rely on some special devices to actively transmit specific signals to the victim AV's radar. It is usually difficult to perform such attacks in practice because they either require some special devices to be placed at a specific distance to the victim radar, or require sub-nanosecond-level synchronization between the devices and the victim radar. Besides, the specially designed spoofing devices make the attacks very costly. In addition, our investigation shows that although these spoofing attack methods can change the outputs of raw radar measurements after FFT, they can not effectively fool the succeeding DNN models to eventually change the outputs of radar object detection models.

To address the above problems, in this paper, we investigate the possibility of using some adversarial objects to attack the DNN-based radar detection models through passive reflection. Our investigation shows that the amplitudes of the radar echo signals generated by different spots on a target vehicle are important for DNN-based radar detection models to learn the features of the target. If we can manipulate the amplitudes of the echo signals generated by different spots on the target, we can fool the radar detection model and change its detection result on the target. Towards this end, we leverage the characteristics of mmWave signal reflection on a metal surface and design a novel structure of adversarial objects, named as TileMask, which can be customized into specific geometric shape to generate desired signals in specific amplitudes through passive reflection. Figure 1 shows an example for the proposed adversarial objects. Each adversarial object is composed of a base and multiple reflective surfaces, called reflective tiles. These tiles are made of reflective materials such as metal foils, and different tiles are in different orientations (different angles of inclination θ_s). The intuition behind this design is that, the amplitude of the echo signal generated by each metal tile is determined by the inclination angle θ_s of the tile. By sticking some adversarial objects at some specific locations on a target vehicle and designing the value of θ_s (or d) for each tile, we can change the amplitudes of echo signals generated by the covered areas on the target and manipulate the superimposed signal received by the radar. This can further affect the features learned by the radar object detection models to make it failed to detect the target.

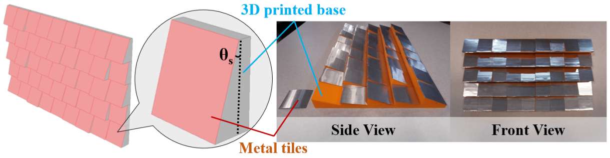


Figure 1: An example of adversarial objects.

Figure 2 shows an example that can be used to illustrate the proposed attack. The victim AV drives on a road and there is a target vehicle in front of it. The target vehicle could be parked on the road by the attacker intentionally. The attacker first generates the adversarial objects, and then sticks them at the derived locations on the target vehicle before the attack. As the victim AV drives towards the target vehicle, its radar perception system is fooled and fails to detect the target vehicle, which may lead to a rear-end collision. The proposed adversarial object can be easily fabricated at low cost. The base of the object can be fabricated using 3D printing techniques and the metal foils can be made as metal tiles, which do not require any other special materials. The average cost of fabricating the adversarial object is only $10. This example shows the proposed attack can be easily performed in practice. In addition, since mmWave signals can penetrate some thin papers or fabrics, the attacker can cover the adversarial objects with some advertisement posters to conceal them as car signs, as shown in Figure 2. In this way, the adversarial objects are hard to be recognized as malicious objects, making the attack more stealthy.

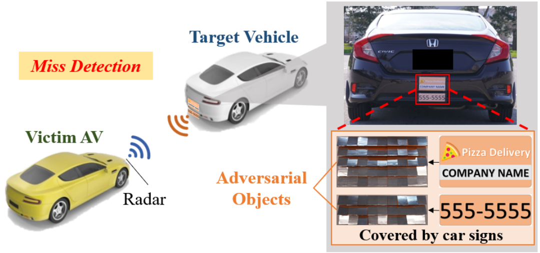


Figure 2: An example of the attack.

In the proposed attack, a challenging problem is how to generate the adversarial objects that are not only effective at achieving the attack goal but also in small sizes to make the attack stealthy and cost-efficient. To address this problem, we first characterize the adversarial objects with some parameters including the number of the adversarial objects, the locations and sizes of these objects, and the orientations of the tiles that compose these objects. Then we formulate an optimization problem to generate the desired adversarial objects. Since the number of the adversarial objects is a discrete value, it is hard to directly solve this optimization problem. To handle this challenge, we propose a two-step framework that can iteratively update the orientations of the tiles and the adversarial objects' number, locations, and sizes. In addition, our proposed attack framework can achieve environment independent attack where the generated adversarial objects can achieve the attack goal in various background environments.

Experiments are conducted in both the physical world and digital world. In the physical world experiments, we show that the attacker can use two adversarial objects concealed as car signs to continuously hide the target vehicle as the victim AV is approaching the target. The generated adversarial objects can achieve 93% attack success rate. And the experimental results demonstrate that the derived adversarial objects are environment independent. To the best of our knowledge, this is the first study on the passive-reflection-based attacks against the DNN-based radar object detection models using low-cost, readily-available and easily-concealable geometric shaped objects.

2 PRELIMINARY

2.1 DNN-based Radar Object Detection

DNN-based radar object detection models have been widely used for the perception systems in autonomous driving, and have achieved state-of-the-art performance. These models aim to understand the driving environments by transmitting mmWave FMCW signals and extracting features from the received signals. For state-of-the-art radar detection models, the common pipeline is summarized in Figure 3.

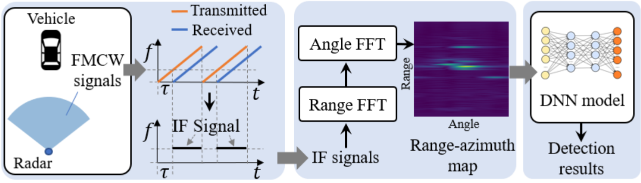


Figure 3: Radar object detection pipeline.

The radar first transmits the FMCW signal, which is a kind of continuous wave whose frequency increases uniformly with time. Then the echo signals are received and processed into Intermediate Frequency (IF) signals at the receiver, by measuring the difference of the instantaneous frequency of the received signal and transmitted signal, as shown in Figure 3. Suppose there is only one small object at a distance of d^* and angle-of-arrival (AOA) a^* from the radar transmitter, the IF signal at the k-th antenna of the receiver can be approximated as [52]: $s(t,k) \approx P_r \exp(-j2\pi(\gamma(2d^*/c)t + k\delta_a \cos a^*/\lambda))$, where P_r is the amplitude of the IF signal, γ is the slope of the chirp signal, and δ_a is the spacing between adjacent antennas of the receiver. $2d^*/c$ represents the propagation delay between the object and the radar, and $k\delta_a \cos a^*/\lambda$ represents the relative propagation delay between the receiver's antennas. The detection model preprocesses the obtained IF signals to generate the range-azimuth map X through two successive steps: range-FFT and angle-FFT. Range-FFT is applied on the IF signal at each antenna of the receiver along the time domain to estimate the range of the object, and angle-FFT is applied on the output of range-FFT along the receiver's k-th antenna to estimate the angle-of-arrival of the object. After the above two steps, the IF signals at multiple antennas of the receiver can be transformed to a range-azimuth map: $X(d,a) \approx P_r\delta(d-d^*)\delta(a-a^*)$, where $\delta()$ is the Dirac delta function.

The above formulas describe an ideal case where a small enough object can be treated as a single reflective spot (surface). In the physical world, a target object can be divided into a large number of small reflective surfaces and they all contribute to the final IF signal. The final IF signal S at the k-th antenna of the receiver can be formed as the summation of the IF signals from every small surface i:

$$S(t,k) \approx \sum_i P_{ri} \exp(-j2\pi(\gamma(2d_i^*/c)t + k\delta_a \cos a_i^*/\lambda)), \tag{1}$$

where d_i^* and a_i^* are the range and AOA of each surface i. P_{ri} is the amplitude of the IF signal generated by each surface i. The final range-azimuth map X is derived by applying the two successive FFT on the final IF signal S. According to Eq. (1), the amplitude P_{ri} of the IF signal from each surface determines the final IF signal and affects the range-azimuth map.

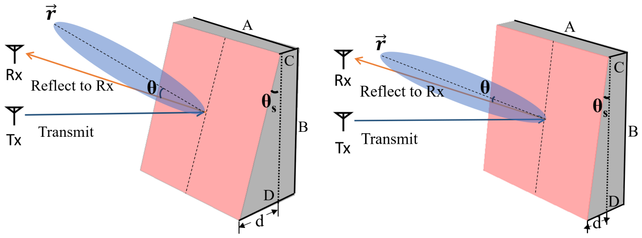
The range-azimuth map provides rich information about the driving environments and can be processed by convolutional neural networks (CNNs). It is widely adopted as the input of DNN-based radar object detection models. The detection model normally takes multiple range-azimuth maps X at consecutive timestamps as the input to encode the velocity information. The detection model learns the features from the input range-azimuth maps and outputs a set of detection result candidates. Each candidate contains the confidence score of each class and its location (distance and angle). The candidates whose confidence scores are smaller than a threshold are removed and the remaining candidates are merged to get the final detection results. The class of each detected object is the class that has the maximum confidence score.

2.2 Vulnerability of DNNs

DNNs have been demonstrated to be vulnerable to adversarial attacks where an attacker can fool the DNNs to generate wrong outputs by making small perturbations to their inputs [19, 30, 53]. Suppose M denotes a deep learning model. In adversarial attacks, an attacker aims to slightly modify the original input X to X' so that the model's output is significantly different from the ground truth y^*, i.e., $M(X') \neq y^*$. The modification to the original input can be achieved either in digital world by perturbing the input data directly [48, 51, 55, 56, 79] or in physical world by modifying the physical environment [13, 16, 44, 61, 73, 81]. The attacker usually derives the modification by solving an optimization problem.

3 MOTIVATIONS OF TILEMASK

According to Eq. (1), the input of the DNN-based radar detection model is equivalent to the summation of IF signals generated by every spot (small reflective surface i) on the target vehicle. Our investigation shows that the amplitudes of the IF signal (i.e., P_{ri}) generated by each surface i on the target are important for the DNN-based radar detection to learning the features. And the P_{ri} values of some specific surfaces on the target are especially important, which we will demonstrate in Section 7. P_{ri} is determined by the amplitude of the echo signal reflected from the surface i to the radar receiver. Thus, we can manipulate the amplitudes of echo signals generated by these specific surfaces to change the features learned by the radar detection model, making it can not detect the target. To manipulate the amplitude of the echo signals, an intuitive approach is to use some special materials [38] as the skin of the target vehicle, similar to the ideas in some military applications [8]. However, this approach is limited in its cost and practicability.



(a) Weak reflection to receiver **(b) Strong reflection to receiver**

Figure 4: Different θ_s values result in different mmWave reflection on the metal tile.

These materials can be expensive and hard to obtain, which makes the attack very costly. In addition, the reflection ratio (albedo) of a specific material is always the same. To achieve the attack goal, the attacker may need to make different spots on the target generate different amplitudes of echo signals, so the attacker has to use many types of special materials to cover the vehicle. This can be challenging in practice and may require complex manufacturing processes. The above challenges raise an important question: Is there an easier and low-cost way to manipulate the amplitudes of echo signals from the target?

mmWave signal reflection on a single metal tile. To answer this question, we leverage the characteristics of mmWave signal reflection on metal surfaces. Specifically, we found that the amplitude of the echo signal reflected from a metal surface to the radar receiver is determined by the surface's orientation, and the mmWave signals can barely penetrate the metal surface. If we stick a metal surface on the target vehicle, the original echo signal generated by the covered area on the target are replaced by the signal generated by the metal surface. And if we change the orientation of the metal surface, we can change the echo signal into different amplitudes. To stick a metal surface on the target with a specific orientation, we design a special structure of object. As shown in Figure 4, the object is composed of a base (grey color) and a metal tile (pink color). Each tile is a small metal rectangle surface. The value d is the difference between the length of edge C and D. If the length of edge A, B and C is fixed, the geometric parameter d determines the orientation (inclination angle θ_s) of the metal tile.

Next, we will discuss how the angle θ_s of the metal tile affects the amplitude of its echo signal to the radar receiver. Figure 4 shows the examples of the mmWave reflection on the metal tile in this structure. Theoretically, to calculate the reflected signal from one tile, we need to consider every single spot on the tile and superimpose the signals reflected from each spot. This is obviously infeasible. In practice, we study only the signal reflected from the geometric center of each tile (as shown in Figure 4). Given the small size of the tile (3cm * 3cm) compared with its distance to the radar, the signals from the other spots should have similar amplitudes and angles with respect to the tile surface. Thus, the amplitude of the superimposed signal reflected by one tile can be modeled as the product of the amplitude of the signal from the tile center and the area of the tile. The superimposed signal has the same direction as the signal from the tile center. To obtain a more accurate approximation of the superimposed signal, we can further divide the tile into smaller grids and study the signal from the geometric center of each grid. The more grids we divide, the better approximation can be achieved, and on the other hand, the more computations have to be done.

Based on the above approach, we then calculate the reflected signal on the metal tile. A perfect reflector would reflect the incident signals from the same direction into a single outgoing direction (referred as mirror-like reflection). However, the metal tile is not perfect in practice. It is usually modeled as a quasi-specular reflector, where the reflected signals on the tile surface are diffused in many directions [11, 40, 47, 66]. The amplitudes of the signals reflected to different directions are usually different, and their distribution can be modeled as the Gaussian distribution [40]. We use r to denote the direction where the reflected signal has the strongest amplitude. Suppose the amplitude of the transmitted signal is A_i, then the amplitude of the echo signal reflected back to the receiver can be modeled as:

$$A_r = \epsilon A_i \exp(-\theta^2/2\sigma^2), \tag{2}$$

where ϵ is a constant that is determined by the reflection ratio of the tile's material and the area of the tile, σ is a constant that is determined by the tile's material, and θ denotes the angle between the direction r and the direction to the receiver. Obviously, when the locations of transmitter, receiver, and the tile are fixed, the angle θ is determined by the inclination angle θ_s of the tile.

Thus, the amplitude of the echo signal reflected back to the receiver (i.e., A_r) is determined by the inclination angle θ_s of the metal tile. By changing the value of d, we can change the value of θ_s, and further manipulate the amplitude of the echo signal generated by the tile. Figure 4 shows two examples of mmWave reflection under different values of d and θ_s. The angle θ in Figure 4b is smaller than that in Figure 4a, which results in larger amplitude of echo signal to the receiver, according to Eq. (2).

Design of the adversarial object. To change the learned features of the radar detection system, we need to change the amplitudes of the echo signals generated by different spots on the target vehicle into some specific values, and these values for different spots could be different. Thus, we use the above structure as a basic unit and design the adversarial object in Figure 1. Each unit has a different value of d and θ_s, which makes all the metal tiles form into a specific surface pattern. By sticking some adversarial objects on the target vehicle at some specific locations and carefully designing the value of θ_s in each unit, we can manipulate the amplitudes of echo signals generated by different spots on the target vehicle, which can significantly change the learned features of radar detection systems. Ideally, to achieve the best attack performance, we hope to make every spot on the adversarial object to generate a specific amplitude of signal. This requires the size of each metal surface (edge A and B) to be very small. However, smaller value of edge A and B would increase the difficulty of 3D-printing the object. Based on the above considerations, we empirically set edge A and B to 3cm. The edge C has no effect on our attack, so we empirically set it to $0.5cm$. 3D printing techniques can be used to print the bases of all the units into one piece, and each metal tile is made of a stainless steel foil, as shown in Figure 1. The material of the base almost has no impact on the mmWave reflection because the mmWave signal can barely penetrate the metal tiles. In our experiments, we use polylactic acid (PLA) to print the base.

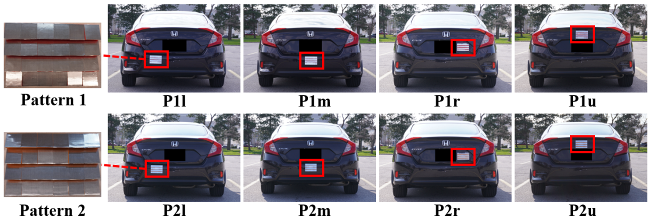


Figure 5: Adversarial objects with different surface patterns placed on different locations on target vehicle.

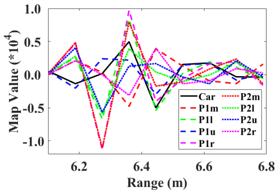


Figure 6: Real dimension of $X(d, a = 0)$.

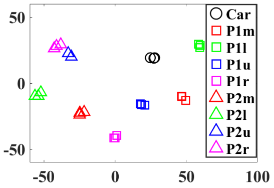
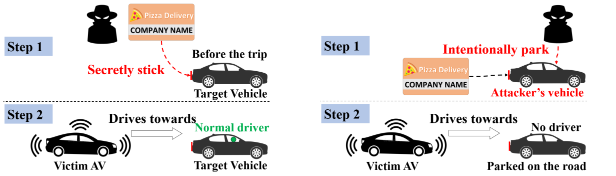


Figure 7: Learned features visualized by t-SNE.

Preliminary evaluation. We conduct a preliminary study to demonstrate the possibility of using our designed adversarial objects to perform the attack. As shown in Figure 5, we fabricate two adversarial objects with different surface patterns and stick each object at four different locations on a car. We use a TI AWR1843 board attached with a DCA1000 board as the mmWave radar, and place it at $6m$ behind the car to collect the IF signals. The victim radar detection model is RODNet [75]. We obtain the inputs of the radar detection model (i.e., range-azimuth maps $X(d, a)$) from the collected signals in the eight examples in Figure 5, denoted as P1l, P1m, P1r, P1u, P2l, P2m, P2r, and P2u. We also obtain the input of radar detection model when no object is placed on the car, denoted as Car. Figure 6 visualizes the values of the real dimension of $X(d, a = 0)$ in the nine examples. We can see the values of P1l, P1m, P1r, P1u, P2l, P2m, P2r, and P2u are quite different from that of Car. Also, the values of P1l, P1m, P1r, P1u, P2l, P2m, P2r, and P2u are quite different from each other. These show that adversarial objects with different surface patterns at different locations can change the inputs of the radar detection model to different values. In Figure 7, we use t-distributed stochastic neighbor embedding [71] to visualize the feature points learned by the radar detection model given the inputs in Figure 6. We can see that these different inputs can result in different learned features of the radar detection model. Thus, placing the proposed adversarial object can affect the input of radar detection models and further affect the features learned by the radar detection model. By designing the surface patterns and locations of the objects, it is possible to use them to change the output of the radar detection model.

4 PROBLEM SETTING

Attack goal and threat model. This paper focuses on the scenario where the AVs are equipped with mmWave radar and use DNN-based radar object detection models to detect objects (e.g., vehicles or pedestrians) on the roads. Specifically, we assume that the victim AV drives on a road and there is a vehicle in front of it (target vehicle). The goal of the attack is to continuously hide the target vehicle from the radar perception system of the victim AV, e.g., make the victim AV not able to detect the target vehicle in its collected radar frames as it drives towards the target vehicle. This type of attack may cause catastrophic consequences such as rear-end collisions. We assume that the attacker can generate and stick the adversarial objects on the target vehicle. For example, the attacker could use their own cars to launch the attack by intentionally parking a car on the road and sticking the adversarial objects on it (Figure 8b). Besides, the attacker could secretly stick the objects on someone else's car when the driver parks on the roadside temporarily, stops at a traffic light, or even before the driver starts a trip (Figure 8a). The attacker may have many types of motivations to launch this kind of attack, such as causing traffic accidents for insurance frauds, unfair competition between autonomous driving companies, or hurting the drivers and passengers in the vehicles. We consider a practical and challenging setting where the attacker can not obtain the original radar data collected by the victim AV. Besides, we consider a white-box setting and assume that the attacker has the full knowledge of the victim radar object detection system, which is also adopted in existing radar attack methods [20, 39, 49, 69]. This is reasonable because some autonomous driving companies launch open-source autonomous driving platforms [2, 4]. The attacker can also purchase the same model of AV as the victim AV and obtain such information through reverse engineering. The attack works as follows: the attacker first simulates the possible driving conditions of the victim AV in some random environments and generate the adversarial objects in an offline manner; then he stick the adversarial objects on the target vehicle to perform the attack, and he does not need to take any further actions after that. The adversarial objects can continuously hide the target vehicle from the victim AV as it drives forward in real-time, even under various background environments.



(a) Example 1 (b) Example 2

Figure 8: Examples of the attack.

Problem definition. To achieve the attack goal, the attacker needs to generate the adversarial objects and their locations by maximizing the attack effectiveness. In addition, to make these adversarial objects easy to fabricate, cost-efficient and stealthy, the attacker needs to minimize the total area of these objects and average angle θ_s of all tiles. Suppose the number of the adversarial objects used by the attacker is N. The sizes of these objects are denoted by $P_s = \{P_{s,n} | n = 1, 2, ..., N\}$, where $p_{s,n} = \{h_n, w_n\}$ denotes the size of the n-th adversarial object, and $\{h_n, w_n\}$ represents the height h_n and width w_n of this object. We use $P_l = \{P_{l,n} | n = 1, 2, ..., N\}$ to denote the locations of the adversarial objects. $P_{l,n} = \{x_n, y_n, z_n\}$ is the location of the n-th adversarial object, and $\{x_n, y_n, z_n\}$ represents the xyz-coordinates of this object. In addition, the surface patterns of the adversarial objects are denoted by $P_a = \{P_{a,n} | n = 1, 2, ..., N\}$, and $P_{a,n}$ denotes the set that contains every tile's angle (i.e., θ_s) in the n-th object. Then we formulate the problem of deriving the adversarial objects as the following optimization problem:

$$\begin{aligned} \min_{N,P} \quad & M(X'_e) + \alpha L_{area} + \beta L_{pattern} \\ \text{s.t.} \quad & X'_e = F(S'_e), \\ & S'_e = T_{V,e}(N, P_s, P_l, P_a), \end{aligned} \tag{3}$$

where $P = \{P_s, P_l, P_a\}$ and $M(X'_e)$ denotes the output detection confidence of the target vehicle in its corresponding class given the input range-azimuth map X'_e. $F(S'_e)$ denotes the two successive FFT on the obtained IF signal S'_e. S'_e is derived by the function $T_{V,e}$, which models the obtained IF signal given the target vehicle V in the environment e. L_{area} is the total area of the N adversarial objects, and $L_{pattern}$ is the average value of $\{P_{a,n}\}_{n=1}^{N}$, minimizing which makes the adversarial objects easy to fabricate, cost-efficient and stealthy. α and β are used to adjust the trade-off between the three terms in the objective function.

5 METHODOLOGY

To solve the above optimization problem, we first need to obtain the function $T_{V,e}()$ that models the IF signal at the radar. However, in practice, the final IF signal are not only generated by the echo signals reflected from the target vehicle and adversarial objects (foreground objects) but also generated by the echo signals reflected from other objects in the surrounding environment. It is difficult and time-consuming to simulate the echo signals from various surrounding environments by manually building the meshes of the surrounding environments. To address this challenge, we propose to decompose the IF signals into two parts, i.e., the signals generated by the foreground objects and the signals generated by the background environment. Then we separately simulate the two types of signals and finally combine them to generate the final IF signal obtained by radar. A new method is proposed to simulate the signals reflected from the background environment using point cloud data. The details are described in Section 5.1.

Another challenge when we solve the above optimization problem is that the second constraint in this problem (Eq. (3)) is non-differentiable because N is discrete, and this makes it difficult to directly solve this problem using gradient based methods. To address this challenge, we propose a novel solution in which the parameters are divided into two groups, i.e., the surface pattern-related parameters P_a and the object's coverage-related parameters $\{N, P_l, P_s\}$, and they are updated alternatively until the convergence criterion is satisfied. The proposed solution contains two steps: *pattern update* and *coverage update*. In the pattern update step, we fix the coverage-related parameters and use gradient descent to optimize P_a. In the coverage update step, we update the coverage-related parameters using a heuristic method. Specifically, we first calculate an importance score for each reflective tile of the adversarial objects, and this score reflects the importance of a particular reflective tile on determining the detection result. Then we update $\{N, P_l, P_s\}$ by removing the redundant tiles with small importance scores to reduce L_{area} without hurting the value of $M(X'_e)$ significantly. The details of the solution are described in Section 5.2 and Section 5.3.

5.1 mmWave Reflection Simulation

Reflected signal from the target. We first simulate the IF signal generated by the target vehicle without the background environment. Specifically, we first obtain the mesh of the target V through 3D model databases [1], mesh generation methods [36, 76], or manual building. Then we divide the target mesh into a large number of small reflective surfaces (triangles). The occluded surfaces of the target mesh (e.g., surfaces occluded by adversarial objects) are removed based on the parameters $\{N, P_l, P_s\}$ of the adversarial objects during the updating process. Based on the method proposed in [40], the IF signal at the receiver's k-th antenna can be represented as the summation of the IF signals generated by each surface i:

$$S'(t,k) = \sum_i \frac{\omega A_g A_m A_a}{(4\pi)^2 d_{Ti} d_{iR}} \exp(-j2\pi\gamma \frac{d_{Ti} + d_{iR}}{c} t), \tag{4}$$

where d_{Ti} is the distance between the transmitter and surface i, and d_{iR} is the distance between the surface i and the receiver's k-th antenna. Since the amplitude of the echo signal reflected from each surface to radar receiver is determined by its area, orientation, and material, a few matrices (i.e., A_a, A_m, and A_g) are used to measure these factors. A_a measures the area of each surface. A_m represents the reflective ratio of the mmWave signal on each surface. A_g models the relationship between the amplitude of the incident signal and the reflected signal towards the receiver's k-th antenna: $A_g = \exp(-\theta^2/2\sigma^2)$, where θ is the angle between the direction r_i that achieves the maximum reflection amplitude at surface i and the direction from the surface i to the Rx, as shown in Figure 4. This simulation method is lightweight, effective, and efficient. The average structured similarity [77] between the range-azimuth maps generated from the simulated signals and the real signals for various targets is over 90%.

Reflected signal from the background environment. To simulate the IF signals generated by background environments, we could build the environments' meshes and use the same method above. However, building the meshes for various environments is difficult and time-consuming in practice. To address this challenge, we propose to use 3D point clouds of the environments to generate their IF signals. Point cloud is a precise 3D representation of the environments, and can be easily collected using LiDAR or RGBD sensors, or from open-source datasets. Each point in the point cloud can be treated as the location of a virtual reflective surface. Each surface's orientation can be generated by estimating its normal vector. Each surface's area can be approximated by calculating the density of the points around it. Each surface's material can be approximated based on the point's label, which can be obtained from the datasets' label information. Thus, the IF signal generated from the environment can be calculated based on Eq.(4). The proposed method can conveniently and quickly derive the IF signals generated by the background environments without the need of building the environments' meshes. The average simulation time for each background environment is only 2.6s. And we found that these simulated signals can help achieve environment independent attacks by considering enough environments in the optimization problem.

5.2 Pattern Update

Our solution for the proposed optimization problem contains two steps: pattern update and coverage update, which are alternatively conducted until some convergence criterion is satisfied. The goal of the first step is to optimize P_a while fixing the values of $\{N, P_l, P_s\}$. To achieve this goal, we divide the IF signal generated by the foreground objects into two parts: the IF signal generated by the target itself (S'_t) after removing the surfaces occluded by adversarial objects, and the IF signal generated by the adversarial objects (S'_a). Then, we have the following optimization problem:

$$\begin{aligned} \min_{P_a} \quad & M(X'_e) + \beta L_{pattern} \\ \text{s.t.} \quad & X'_e = F(S'_t + S'_a + S_e), \\ & S'_a = T(P_a). \end{aligned} \tag{5}$$

where S_e is the IF signal from the background environment e calculated by the proposed environment signal simulation method. S'_t can be calculated using the above target's signal simulation method. Since the parameters $\{N, P_l, P_s\}$ are fixed, the remaining surfaces of the target after occlusion are also fixed, so S'_t is a constant in this step. T is a function that is used to model the IF signals generated by the adversarial objects when fixing the parameters $\{N, P_l, P_s\}$. T is differentiable and can be derived by modeling the IF signals generated by each reflective tile based on Eq.(2) and Eq.(4). Gradient descent is used to optimize P_a.

5.3 Coverage Update

Importance score. To update the parameters N, P_l, and P_s, we propose to remove the unimportant tiles of the adversarial objects. Removing these unimportant tiles can help reduce the total area of the adversarial objects L_{area} without affecting the detection confidence $M(X'_e)$. To achieve this, we introduce an importance score associated with each small reflective surface on the target and each tile on adversarial objects to study its importance to radar object detection. In the rest of this section, we use "surface" to refer to both the small surface on the target and the tile on adversarial objects for convenience. Based on the intuition in Section 3, our proposed attack is achieved by manipulating the amplitude of the echo signal generated by each small surface on the object. We define the importance score of each small reflective surface as the effect of changing its echo signal amplitude on the detection confidence. To measure the effect of changing each surface's echo signal amplitude, we leverage the characteristics of 2D FFT. According to Section 2.1, the amplitude of echo signal generated by each surface affects the magnitude of the corresponding pixel in the range-azimuth map. Larger amplitude of its echo signal results in larger magnitude of the corresponding pixel in the range-azimuth map. Based on this, we define the effect of changing the amplitude of echo signal generated by the surface i as: $(F(s_i + \delta s_i) - F(s_i)) * G(F(\sum_i s_i))$, where s_i is the IF signal generated by surface i, $G()$ calculates the gradients of M for each input pixel given the input range-azimuth map $F(s_i)$, and δs_i is the change of IF signal after increasing/decreasing the amplitude of echo signal from surface i. In the physical world, there are constraints on the maximum and minimum echo signal amplitude generated by each surface, which limits the amplitude of IF signal generated by each surface. We define the IF signal upper bound for each surface $\hat{S} = \{\hat{s}_i\}$ as the IF signal generated by assuming each surface reflect all the incident signal back to the radar receiver, i.e., the amplitude of the echo signal is the same as that of the incident signal ($A_g = A_m = 1$). Obviously, the IF signal lower bound for each surface is zero: $\check{S} = \{\check{s}_i\} = 0$, i.e., no signal is reflected back to the radar receiver. We define the importance score of each surface by measuring the effect of decreasing its echo signal amplitude to its lower bound and increasing its echo signal amplitude to its upper bound. The effect of decreasing the echo signal amplitude of surface i is calculated by the integrated effect on the detection confidence:

$$W_{dec} = F(s_i) * \sum_{b=0}^{B} G(F(\sum_i \frac{b}{B} * s_i))/B, \tag{6}$$

so does the effect of increasing the reflected signal amplitude of surface i:

$$W_{inc} = (F(\hat{s}_i) - F(s_i)) * \sum_{b=0}^{B} G(F(\sum_i s_i^b))/B, \tag{7}$$

where $s_i^b = (s_i + \frac{b}{B} * (\hat{s}_i - s_i))$ and B is the number of integrated steps. The total effect of changing surface i is calculated by considering the effect of both increasing and decreasing its echo signal amplitude: $W = W_{inc} + W_{dec}$. The positive value of W indicates positive correlation between its echo signal amplitude and the detection confidence (i.e., increasing its echo signal amplitude will also increase the detection confidence of the target), while the negative value of W indicates negative correlation. The final importance score of surface i is given by the absolute value of W. Manipulating the echo signal amplitude of surfaces with high importance score has large impact on the detection results, while manipulating the echo signal amplitude of surfaces with low importance score has a small impact on the detection results.

Parameters Update To update $\{N, P_l, P_s\}$, we propose to remove the tiles with low importance score to decrease the total L_{area}. Specifically, after calculating the importance scores of all the tiles of the initial adversarial objects. The tiles with low importance scores (smaller than threshold w_{thre}) are removed and the remaining tiles are clustered based on their geometric locations. The clusters for which the number of tiles is smaller than a threshold are removed. A new adversarial object is then generated from each of the remaining clusters, so the number N of the adversarial objects is the number of the remaining clusters. The location $P_{l,n}$ of each adversarial object is the location of the corresponding cluster, and the size of the objects $P_{s,n}$ is generated based on the minimum bounding box of the corresponding cluster. For the feasibility of 3D printing, the adversarial objects are kept in rectangle shape during the updating process.

5.4 Attack Algorithm

Before conducting the above two steps, we first initialize the adversarial objects by calculating the importance score of each small surface on the target vehicle. Then we cluster the small surfaces and sort the clusters according to their average importance scores. As we discussed, changing the amplitudes of signals reflected from the surfaces with high importance scores tends to have a large impact on the detection results. Thus, we select the clusters with high average importance scores. The number of the adversarial objects is the number of the selected clusters. The locations of the initial adversarial objects are generated based on the locations of the selected clusters. The sizes of the adversarial objects are generated based on the minimum bounding box of the selected clusters. The initial surface patterns are randomly generated within a given range. After the initialization, the Pattern Update step and Coverage Update step are alternatively conducted to generate the adversarial objects' parameters N, P_s, P_l, P_a.

5.5 Environment Independent and Continuous Attack

We then propose to generate the adversarial objects that are environment independent, i.e., they can achieve the attack goal in various environments. In addition, we propose to continuously hide the target under various possible positions, orientations, and speed between the target and radar, and consider the possible location and orientation errors when placing the adversarial objects.

Perform the attack in practice. To achieve the above goals, the adversarial objects and their locations are generated in an offline manner. Specifically, for a given target, the attacker first simulates various possible driving conditions and generates the target meshes in these conditions, such as different positions, orientations and speed. The attacker randomly samples some environments from a public dataset and generates their background signals S_e using the method in Section 5.1. Based on the above algorithm, the attacker derives the adversarial objects and their locations by summing the objective values in Eq. (3) for all the simulated conditions and the randomly sampled background signals. Random perturbations on the locations and orientations of the adversarial objects are also added during the generation process. Finally, the attacker can perform the attack by placing these adversarial objects at the derived locations. And the adversarial objects can continuously hide the target vehicle under various driving conditions and be robust to location and orientation errors. And they can achieve the attack goal under various background environments even when they are not included in the sampled environments.

6 PERFORMANCE EVALUATION

6.1 Experimental Setting

We first use one of the state-of-the-art radar object detection models, i.e., RODNet [75], as our target model. It first preprocesses the IF signals received by radar into range-azimuth maps, and then uses CNNs to learn the features from the range-azimuth maps, as discussed in Section 2.1. The outputs of RODNet contain the locations and class labels of the detected objects (e.g., vehicle, pedestrian, and bicycle). We train RODNet with CRUW dataset [5] according to [75]. In our experiments, we consider the scenarios where the victim AV is driving toward the target vehicle, i.e. a black Honda sedan. The attack goal of the attacker is to continuously hide the target vehicle from the victim AV's radar perception system as the victim AV drives forwards. In addition, we intend to achieve the attack goal under various background environments to achieve environment independent attack. To achieve these goals, we generate the adversarial objects using the proposed framework in Section 5.5. The mesh of the target vehicle is obtained from the public 3D model database with slight modifications.

Physical world evaluation. To evaluate the proposed attacks in the physical world, we use a TI AWR1843 board attached with a DCA1000 board as the mmWave radar, which is a widely adopted radar for object detection in autonomous driving [7]. It is also the same radar used by the victim detection model, i.e., RODNet. We mount the radar on the front of the Lincoln MKZ as the victim AV testbed, as shown in Figure 9. The height of the radar is around 0.5m. The radar configurations are the same as described in RODNet, which aims to ensure the input range-azimuth map has the same size as that of the original model. 3D printing techniques are adopted to print the base of the adversarial objects, and stainless steel foils used as the metal tiles. Based on the proposed algorithm, we generate two adversarial objects that are enough to hide the target vehicle, as shown in Figure 10. Due to the fact that mmWave signal can penetrate some thin papers and fabrics, the two adversarial objects can be concealed as some car signs by covering them with some posters. The attacker can stick these car signs on the rear of the target vehicle to continuously hide it from the victim AV. To evaluate the attack performance, we drive the victim AV towards the target vehicle in Figure 10 and collect the data as the victim AV is approaching the target vehicle from 20m to 2m. To demonstrate the environment independent attack, we collect the data in four different real-world environments. Please note that when generating the adversarial objects, none of the four environments are considered in the optimization problem.

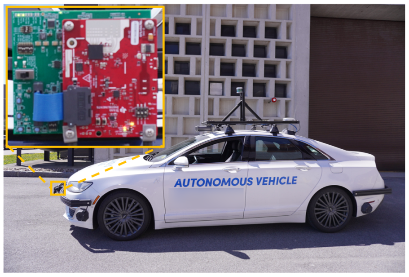


Figure 9: mmWave radar testbed

Digital world evaluation. To further evaluate the attack in more scenarios, we also perform the attack in the digital world. We randomly sample 200 scenes from the SemanticKITTI dataset [15]. In each scene, we consider the same attack scenarios as that in the physical-world experiments, i.e., the victim AV drives towards the target vehicle from 20m to 2m. The adversarial objects and their locations are the same as that in the physical world evaluations. To evaluate the attack performance under these scenes, we simulate the signals reflected from these background environments and the IF signals received by the radar using the proposed method in Section 5.1. Please note that when generating the adversarial objects, none of the above environments are considered in the optimization problem.

For the evaluation metric, we use the *Attack Success Rate* (ASR), which is defined as the percentage of the examples (radar frames) that are successfully attacked among all the collected examples. An example is successfully attacked when the target is successfully hidden from the radar detection. We also calculate the detection Recall (i.e., the percentage of the examples where the target vehicle is successfully detected) before and after the attack, referred to as *Recall-benign* and *Recall-attack*, respectively.

6.2 Overall performance

As shown in the Figure 10, the attacker camouflages the adversarial objects as a pizza delivery advertisement, and sticks them at the derived locations on the target vehicle. The total area of the

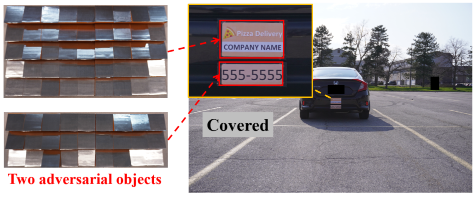


Figure 10: The adversarial objects are concealed as car signs and continuously hide the front car as victim AV approaches.

two adversarial objects is $0.06m^2$. We first evaluate the attack performance in four different types of environments in the physical world, where the victim AV are driving on the road surrounded by trees (Figure 11b), buildings (Figure 11c) and passing-by vehicles and pedestrians (Figure 11d). Before the attack, we drive the victim AV towards the target vehicle 10 times in each environment. Each time, we collect a sequence of the radar frames, i.e., a sequence of IF signals received by the radar at each timestamp. Without the attack, we collect 40 sequences that contain 2243 radar frames in total, and he target vehicle is detected in 95% of the frames. During the attack, we also drive the victim AV towards the target vehicle 10 times in each environment and collect 40 sequences that contain 2316 frames. On average, each sequence contains 58 frames and the target vehicle is detected in only 4 frames. Among all the collected frames, the target vehicle is detected in 7% of the frames, and the average attack success rate is 0.93. For the frames where the target is not detected, the average detection confidence is 0.12. Table 1 shows the average attack success rates and detection Recalls before and after the attack in these four different environments. We can see that the proposed attack can achieve similar success rates in the four environments. The results demonstrate that the derived adversarial objects are environment independent.

Table 1: Performance in different environments

Environments	Env-A	Env-B	Env-C	Env-D
Recall-benign	0.93	0.94	0.98	0.93
Recall-attack	0.09	0.06	0.09	0.04
ASR	0.91	0.94	0.91	0.96

Our proposed attack can continuously hide the target when the victim AV approaches. To further evaluate the effect of the distance between the victim AV and target, we divide the collected radar frames into different groups according to the distance between victim AV and target vehicle. The average attack success rates of different groups are shown in Table 2. The target vehicle is completely hidden from the victim AV when the distance is larger than $8m$, which can be smaller than the minimum braking distance of the vehicle [32]. In addition, even when the target vehicle is occasionally detected when the distance is smaller than $8m$, the occasionally detection in a few frames may be identified as false alarms and ignored by the victim AV.

To demonstrate the attack effectiveness in more scenarios, we also evaluate the attack performance in the digital world. Figure 12 shows two examples of the selected environments in SemanticKITTI dataset. The average attack success rate in all 200 environments is 0.92. We summarize the attack performance under eight selected environments in Table 3. The attack can achieve similar attack success rates in the eight environments. These results demonstrate that the derived adversarial objects are environment independent. This is because the adversarial objects have dominant effects on the detection results by considering some randomly sampled environments when generating the adversarial objects. To demonstrate this, we calculate the importance score of the environment (background) and the target vehicle with adversarial objects (foreground). Specifically, we adopt the method in Section 5.3, and calculate the average importance score of each small reflective surface in the background and foreground, respectively. Table 3 reports the ratio between the background importance score and the foreground important score in the eight selected environments. We find that the background environments are much less important than the foreground, which demonstrates that the derived adversarial objects have dominant effects on the detection.

Table 2: Performance w.r.t. different distances.

Distance (m)	$2-8m$	$8-14m$	$14-20m$
Recall-benign	0.97	0.95	0.89
Recall-attack	0.20	0.00	0.00
ASR	0.80	1.00	1.00

Table 3: Performance in different environments.

Environments	$E1$	$E2$	$E3$	$E4$	$E5$	$E6$	$E7$	$E8$
Recall-benign	0.91	0.95	0.92	0.90	0.94	0.92	0.97	0.94
Recall-attack	0.06	0.09	0.09	0.06	0.09	0.07	0.10	0.06
ASR	0.94	0.91	0.91	0.94	0.91	0.93	0.90	0.94
Importance ratio	0.02	0.06	0.07	0.05	0.04	0.03	0.06	0.04

6.3 Comparison to Baselines

In this section, we compare the proposed method with other baseline methods. We consider two types of baseline methods, object-based attack and spoofing attack. In object-based attack, we also leverage the proposed idea that metal surfaces can manipulate the amplitudes of echo signals to change the detection model's output, and consider a naive attack method where the attacker randomly stick two pieces of metal board (metal foil attached on cardboard) in random inclination angle on the rear of the target vehicle, denoted as *Obj-random*. We also use Differential Evolution algorithm [67] to optimize the sizes, locations, and angles of the two metal boards, and this method is denoted as *Obj-optimize*. For spoofing attack, we adopt the method in [49], which has the same attack goal as our proposed attack. It can change the outputs of raw radar measurements by using a special device to inject a spoofed object at a specific location. Since the raw radar measurement method always identifies the strongest reflection point around an area as a potential object, the spoofed object that has strong reflection can make the radar unable to detect the target and detect the spoofed object instead [49]. To perform such an attack, we generate the spoofing signal using the method in [49] and inject it into the received IF signals. We consider the same attack scenarios as that in Figure 11, where the attacker intentionally parks a car on the road and make the victim AV fail to detect it. In addition, to demonstrate the effectiveness of the proposed two-step optimization framework, we consider a baseline method that solves the optimization problem in Eq. (3) directly using Differential Evolution algorithm. The

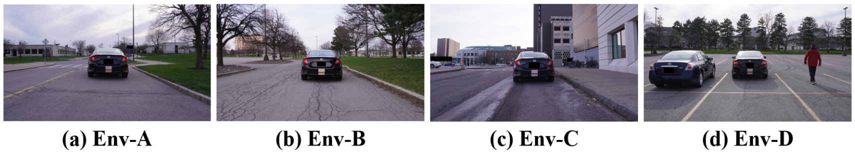


Figure 11: Attack scenarios in physical world.

Figure 12: Examples of selected backgrounds in SemanticKITTI.

number of adversarial objects is set to 2. This method is denoted as *DE*. Table 4 shows the performance of the proposed attack method and baseline methods. The proposed *TileMask* achieves the highest success rate when the number of objects and total area of objects are similar to that of baselines. *Obj-random* and *Obj-optimize* are not effective because the structure of the metal board is too simple to create a malicious pattern of echo signal amplitudes to fool the DNN model. We can also find that the spoofing attack can not fool the DNN-based radar detection models. This is because the DNN-based radar detector relies on the reflection pattern (i.e., pattern of the amplitudes of echo signals generated by different spots on the target) to identify the object, and the spoofing attack can not change the reflection pattern of the target vehicle. Compared with raw radar measurements, the DNN-based radar detection is more robust to the spoofing attack. The results also show that TileMask can achieve better performance than *DE*, which demonstrates the effectiveness of the proposed two-step framework when generating the parameters of the adversarial objects.

Table 4: Comparison to baselines

Method	Recall-attack	ASR	Num of objects	Total area (m^2)
Obj-random	0.93	0.07	2	0.12
Obj-optimize	0.70	0.30	2	0.11
Spoofing	0.85	0.15	-	-
DE	0.57	0.43	2	0.14
TileMask	**0.07**	**0.93**	2	0.06

In addition to the attack effectiveness, we also provide more comparisons between TileMask and spoofing attacks. Existing spoofing attacks require the attacker to use special devices to aim at the victim radar. As shown in Table 5, they either require the devices to be placed at a specific distance to the victim radar [49], or require sub-nanosecond-level synchronization between the devices and the victim radar [39]. So, in their experiments, they only attacked a stationary radar or used a wired link to connect their devices to radar. And their specially-designed spoofing devices are much more expensive than our proposed adversarial objects. Compared with these attacks, our proposed attack is more effective, and does not have requirements on the radar's distance and synchronization. The adversarial objects can be easily fabricated using 3D printing techniques, at an average cost of $10.

Table 5: Comparison to spoofing attacks.

	Constant distance	Wired link	Cost
Spoofing attack [49]	Yes	No	$1920
Spoofing attack [39]	No	Yes	$540
TileMask	**No**	**No**	**$10**

6.4 Hiding Different Types of Targets

In this section, we demonstrate the possibility of hiding different types of targets using the proposed attack framework. We consider a scenario where the attacker aims to continuously hide a person from the vicim AV's radar detection. Since mmWave signals can penetrate paper and thin fabric, the attacker can conceal the adversarial objects with a notebook cover and secretly place them inside a backpack, as shown in Figure 13. When the target person wears this backpack and walks on the road, the victim AV's radar can not see him, which may cause severe traffic accidents. We generate the adversarial objects using the proposed attack framework, and conduct the experiments in the physical world, as shown in Figure 13. The total area of the two adversarial objects is $0.04m^2$. Before the attack, we repeat the experiments 20 times and collect 2321 radar frames in total. The average detection Recall is 0.89. After the attack, we also repeat the experiments for 20 times and collect 2024 frames in total. The average detection Recall is 0.05 and the average attack success rate is 0.95. For each time of the experiment, we collect 101 frames on average and the target person is detected in only 5 frames on average. On the frames where the target is not detected, the average detection confidence is 0.08. Apart from the sedan (Sedan-A) in Figure 10 and person (Person-A) in Figure 13, we perform the attacks to hide more targets in digital world, including two more sedans, a SUV and two more persons. Table 6 shows the average attack success rate, the average value of the angles θ_s for all tiles, and the total area of the derived adversarial objects when hiding different targets. We can see that the derived adversarial objects can achieve over 89% attack success rate for all these targets.

Table 6: Performance on hiding different targets.

Targets	Sedan-B	Sedan-C	SUV	Person-B	Person-C
ASR	0.94	0.90	0.93	0.92	0.89
Average θ_s (°)	23	24	20	19	23
Total area (m^2)	0.09	0.06	0.12	0.04	0.02

Transfer attack on different targets. In some scenarios, the mesh of the target vehicle may not be available to the attacker, so we investigate the possibility of using the adversarial objects generated from one vehicle to hide other vehicles. Specifically, we use the derived adversarial objects for Sedan-A in Figure 10 to hide other vehicles. As shown in Figure 14, we stick the same adversarial objects as that in Figure 10 at the same locations on a blue Nissan sedan and drive the victim AV towards it 20 times. In total, we collected 1186 frames, and the average attack success rate is 0.88. On the frames where the target is not detected, the average detection confidence is 0.11. Table 7 summarizes the attack success rates when using the adversarial objects designed for Sedan-A to hide Sedan-B, Sedan-C and a SUV in digital world. We can find that

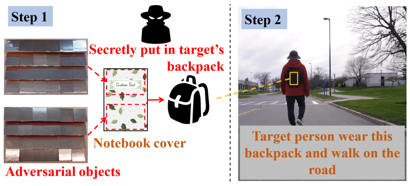

Step 1
Secretly put in target's backpack
Notebook cover
Adversarial objects
Step 2
Target person wear this backpack and walk on the road

Figure 13: Attacker conceals adversarial objects and secretly put them in a backpack. When the target wears this backpack, he is continuously hidden from the AV's radar.

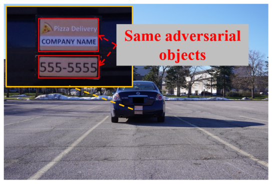
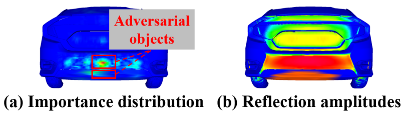


Figure 14: Hiding a different sedan using the same objects.



(a) Importance distribution (b) Reflection amplitudes

Figure 15: Importance distributions on the rear of target vehicle.

the adversarial objects can achieve good success rates when being used to attack the same type of vehicles, i.e., sedans. This is because although the shapes of different sedans are not exactly the same, they are still similar, which results in similar reflection patterns. So the derived adversarial objects can still change their reflection pattern and significantly affect the learned features of DNNs. This enables the attacker to perform the attack even though he could not obtain the mesh of the target: he could generate the adversarial objects designed for other vehicles to hide the target vehicle.

Table 7: Transferability on different targets.

Targets	Sedan-B	Sedan-C	SUV
ASR	0.85	0.89	0.38

6.5 Attacking Different Radar Detection Models

In this section, we evaluate the performance of the proposed attack on different state-of-the-art radar object detection models, including RODNet-hgwi [75], RODCSN [31], and SENet [68].

RODNet-hgwi. RODNet-hgwi uses the temporal inception convolution layers to extract different lengths of temporal features from the input, which is able to achieve better detection performance compared with origianl RODNet.

RODCSN. RODCSN is a computationally efficient radar object detection model. A densely connected residual block is proposed to better deliver the gradient flow from the loss function to improve the feature representation ability.

SENet. SENet improves the detection performance by proposing a noisy detection approach and a weighted location fusion strategy, which ranks as the 3rd place on the 2021 Radar Object Detection Challenge leaderboard [6].

Table 8: Performance on different radar detection models.

Models	RODNet-hgwi	RODCSN	SENet
ASR	0.87	0.85	0.95
Average θ_s (°)	23	15	23
Total area (m^2)	0.14	0.11	0.15

We generate the adversarial objects and their locations for each detection model using the proposed framework. Please note that when generating and evaluating the adversarial objects for each detection model, the radar configurations are set to the same values as that required by each model. We calculate the average attack success rate under the same settings in Section 6.2. Table 8 reports the performance on attacking different radar detection models. We can see that the proposed attack can achieve over 85% success rate on all these models. These results demonstrate the effectiveness of the proposed attack on different radar object detection models.

Transfer attack on different radar detection models. In this section, we investigate the possibility of performing black-box attack. We aim to use the adversarial objects generated for one radar detection model to attack other models. Specifically, we use the adversarial objects generated for RODNet to attack other state-of-the-art detection models. The experimental setting is the same as that in Section 6.2, and the average attack success rates are reported in Table 9. We can find that the adversarial objects for attacking RODNet can achieve good performance when being used to attack RODNet-hgwi and SENet. The attack performance for RODCSN is not as good as the others, but its average success rate is 50% which is still dangerous for autonomous driving. This transferability of the proposed attack can be used for black-box attacks: the attacker can generate the adversarial objects based on other detection models to attack the victim model even though he does not have full knowledge of the victim model. This transferability is because the amplitudes of echo signals generated by some specific areas on the target play important roles in most of the state-of-the-art radar detection models, and the derived adversarial objects can significantly change the echo signal amplitudes of these areas.

Table 9: Transferability on different radar detection models.

Models	RODNet-hgwi	RODCSN	SENet
ASR	0.65	0.50	0.72

7 VULNERABILITY INTERPRETATION

As discussed in Section 3, our proposed attack is performed by manipulating the amplitudes of echo signals generated by some specific areas on target. In Section 5.3, we divide the target into many small surfaces and propose an importance score for each small surface to study the effect of changing the amplitude of echo signal from this surface. In this section, we aim to provide an explanation for the vulnerability of radar detection models by studying the distribution of importance scores on the target vehicle. We propose importance distribution, which is generated by calculating the values of W and importance score $||W||$ for each small surface on the target. Figure 15a shows the importance distribution of the target vehicle in Figure 10 when the distance between the target and victim AV is $9m$. The red color indicates high importance score at that surface, while the blue color indicates low importance score. The red rectangles indicated the locations of the derived adversarial objects in Figure 10.

We can find that the surfaces with high importance scores are always concentrated in a few areas. This shows that there are some important areas on the target, from which the echo signals play critical roles in detecting this target. This characteristic makes the radar perception vulnerable to our proposed attacks. This is because the proposed attack framework can always generate some adversarial objects to cover these important areas (as shown by the red rectangles in Figure 15a) and manipulate the echo signals generated by these areas, which can significantly affect the features learned by the radar perception model. We also find that the important areas are more likely to be around the areas that generate large amplitudes of echo signals. Figure 15b shows the amplitudes of echo signals generated by different surfaces on the target vehicle. The red color indicates that the amplitude of the echo signal generated by this surface is large, while the blue color indicates small amplitude. We can see that the important areas in Figure 15a are around the areas that generate large amplitudes of echo signals in Figure 15b. This is because the amplitude of the echo signal from each surface can be treated as the weight of its IF signal according to Eq. (1), and the signal with larger amplitude contributes more to the summed IF signal, i.e., the input of the radar detection model. Thus, the radar perception model will be trained to rely on the echo signals generated by these important areas to learn the features. Once the echo signals from these areas are manipulated, the learned features will be significantly distorted. Our investigation shows that this vulnerability exists in various victim radar perception models.

8 DISCUSSION

8.1 Potential Defense Strategies

Sensor fusion. A straightforward defense strategy is to use additional sensors such as camera and LiDAR to help detect the target. However, recent studies have proved that camera and LiDAR can also be fooled by physically realizable adversarial attacks [34, 85, 86]. The attacker could combine them with our proposed attacks to fool all the sensors simultaneously to achieve the attack goal [17, 70].

Defense based on importance score. As discussed in Section 7, the echo signals generated by some important areas play critical roles in learning the features. And since the important areas are always concentrated, the attacker can derive some adversarial objects to cover these areas and significantly affect the learned features. To defend the proposed attack, we propose to make the important areas less concentrated and make it difficult for the attacker to find the effective locations of the adversarial objects. Towards this end, we propose to modify the training procedure and force the model to learn the features from the echo signals generated by other unimportant areas. Specifically, we inject the simulated target vehicle into the training data, by adding its generated IF signals to the original mmWave signals. And we divide the surfaces of the vehicle into important surfaces I and unimportant surfaces U. The surfaces I are the surfaces whose important scores of the original model are among the top 30%, while the remaining surfaces are surfaces U. We define the training loss as: $L + \gamma|W_I - W_U|$, where L is the original training loss, and the latter term is the difference between the average important score of important surfaces W_I and that of the unimportant surfaces W_U. This aims to scatter the distribution of important scores. γ is a hyper parameter to balance the two terms, which is set to 0.01 in our experiments. In the training process, we also randomly change the amplitudes of echo signals generated by the important surfaces I to force the model to learn features from unimportant surfaces U. To evaluate the effectiveness of the defense, we select RODNet as the victim model and train a new model using the above defense strategy, denoted as *RODNet-def*. We then perform the proposed attack against *RODNet-def* and evaluate the attack success rate. The attack scenarios are the same as that in previous experiments in the digital world. For various targets, the proposed defense reduces the ASR from 92% to 34% on average. We also evaluate the effectiveness of the defense when it is applied to different radar detection models. We use the proposed defense strategy to re-train various state-of-the-art radar object detection models. Table 10 summarizes the attack success rate of the proposed attack on different models after the defense. The ASR on all the radar detection models is reduced to 30%-40%. The above results demonstrate the effectiveness and generalizability of the proposed defense strategy. However, an attack with a 34% success rate is still dangerous. More robust radar detection model will be explored in our future work based on this defense strategy.

Table 10: Defense performance for different models.

Models	RODNet-def	RODNet-hgwi-def	RODCSN-def	SENet-def
ASR	0.34	0.30	0.35	0.46

8.2 Limitations and Future Works

Multi-sensor fusion. In this paper, we mainly focus on investigating a new type of attack against radar perception systems. As discussed in Section 8.1, the perception system that uses other sensors (e.g., camera and LiDAR) to perform sensor fusion can mitigate this attack. However, based on our proposed attack, it is possible to develop an attack method that can fool all the sensors simultaneously. In our future work, we will study how to extend the proposed attack to sensor fusion systems. For example, a potential attack against camera-LiDAR-radar systems is to cover the adversarial objects with papers painted in specific color patterns. As demonstrated in [81], the specific color pattern can change the pixel values in camera images and affect the camera detection results. And by placing some adversarial objects at some adversarial locations, the LiDAR perception systems can be fooled, as demonstrated in [86]. Thus, all three types of sensors can be attacked simultaneously to change the output of multi-sensor systems. In addition, in some severe weather conditions such as foggy weather, the camera and LiDAR may not provide reliable perception results, and the sensor fusion system may only rely on radar. In such weather conditions, our proposed TileMask can fool multi-sensor fusion systems by attacking its radar perception.

Universal attack. Another limitation of our attack is that the derived adversarial objects for one type of vehicles (e.g., sedan) may not be useful to hide another type of vehicles (e.g., SUV), as shown in Table 7. Thus, the attacker needs to fabricate new adversarial objects if he wants to hide a new type of vehicle. However, since the cost of the adversarial object is small, the attacker can easily manufacture many different adversarial objects for different types of vehicles, and simply choose the corresponding objects for the target vehicle type each time he performs the attack. To further generate the adversarial objects that can hide any types of vehicles, we can take various types of vehicles into the optimization problem by summing their objective values in Eq. (3). Further study on the universal adversarial objects will be our future work.

Precision of placing the adversarial objects. To discuss the difficulty of placing the adversarial objects precisely on the vehicle, we consider two cases based on the attack scenarios. (1) If the target vehicle belongs to the attacker (Figure 8b), the attacker could always place the adversarial objects very precisely before the attack because he has enough time to set up the target vehicle. (2) If the target vehicle does not belong to the attacker (Figure 8a), it is also not difficult to place the adversarial objects precisely enough to achieve the attack goal. Even if the adversarial objects are not placed on the desired locations with 100% precision, they can still achieve good performance. Specifically, we found that when the adversarial objects are placed $7cm$ away from the derived locations, the average attack success rate is still around 80%. This is because, when generating the adversarial objects, we propose to add random perturbations (within 5cm) on their locations in the optimization process, according to our attack framework in Section 5.5. In our real-world experiments, we do not use any precise measuring tools when placing the adversarial objects, and we found that the errors of placing the adversarial objects is always within 5cm. Thus, even when the target vehicle does not belong to the attacker, it is not difficult for the attacker to place the adversarial objects precisely enough to achieve the attack goal.

9 RELATED WORK

9.1 Security of Radar Perception Systems

There are many prior works on the security of autonomous driving systems [14, 23, 24, 28, 29, 37, 45, 46, 57, 58, 62, 65]. Since radar is an important sensor adopted by AVs, the security of radar perception systems has also been studied. There are some existing works that propose to actively transmit signals to the radar and change the outputs of raw radar measurements. The authors in [20] inject false data into the victim radar and change its distance measurement using a physical cable connected to the radar. Some works propose to transmit some specific signals to the radar using software-defined radio techniques, which can create fake objects [69], or range and velocity measurements of an object [39]. Among the existing works, [49] is more relevant to our work because it also proposes to hide a target from radar. They design a spoofing device to introduce a frequency shift on the incoming FMCW signals and transmit the signals to radar. However, the above attacks either require the devices to be placed at a specific distance to the victim radar, or require sub-nanosecond-level synchronization between the devices and the victim radar. So, in their experiments, they only attacked a stationary radar or used a wired link to connect their devices to the radar. And the specially designed spoofing devices are usually very costly. In addition, we found that although these spoofing attack methods can change the outputs of raw radar measurements, they can not effectively fool the succeeding DNN models to eventually change the outputs of DNN-based radar object detection in autonomous driving, as shown in Table 4. [21] proposes to use some tags that are made of a special mmWave absorbing material to hide an object from raw radar measurements. However, the special material can only be used for a specific mmWave radar frequency range, i.e., 18–40 GHz. And today's DNN-based mmWave radar object detection systems in autonomous driving normally operate in a much higher frequency such as 77 GHz [3, 7, 75]. So their method can not be used to attack state-of-the-art DNN-based radar object detection in autonomous driving.

Different from the above works, we leverage the characteristics of mmWave signal reflection on metal surfaces and design a novel structure of adversarial objects to perform the attack through passive reflection. The generated adversarial objects do not require any special materials and can be easily fabricated at low cost. And they can be used for any mmWave radar frequency.

9.2 Adversarial Attacks

Adversarial examples have been proposed to fool the deep neural networks for camera images [22, 25, 33, 35, 50, 61, 81, 83], LiDAR point cloud [18, 64, 78], text [42, 84], and audio signals [10, 63, 72, 80, 80, 82]. Due to the different sensing principles of mmWave FMCW radar and other sensors, these attacks against other sensors can not be adopted to attack mmWave radar. Few recent studies have investigated the adversarial examples in radio signals to fool the radio signal classification [43, 60]. However, radio signal classification intends to generate a label for a given radio signal to identify its modulation type (such as OFDM), which is a different task from radar object detection whose goal is to find locations and classes of the objects on the road. Besides, the above adversarial radio examples are mainly generated in the digital world, and it is difficult to generate them in the physical world. In this paper, we study the characteristics of feature learning in DNN-based radar detection and leverage the characteristics of mmWave signal reflection, and design a novel structure of adversarial objects to perform the adversarial attacks, which is effective, low-cost, and easy to fabricate in the real world.

10 CONCLUSIONS

In this paper, we study the vulnerability of the DNN-based radar object detection in autonomous driving. We leverage the characteristics of mmWave signal reflection on metal surfaces and design a novel structure of adversarial objects to hide a target vehicle. These adversarial objects can be easily fabricated at low cost and can be concealed as car signs. Experiments in both the physical and digital world are conducted to evaluate the attack effectiveness. The real-world evaluations show that the radar object detection model can be attacked continuously when the victim AV approaches the target under various environments, by using only two small adversarial objects.

11 ACKNOWLEDGMENTS

We thank our anonymous reviewers for their insightful comments and suggestions on this paper. We thank Prof. Chi Zhou from University at Buffalo, the State University of New York for printing the adversarial objects. And we thank Yan Zhang from Iowa State University for helping with some experiments. This work was supported in part by the US National Science Foundation under grant CNS-2120369, CNS-2154059, and ECCS-2028872.

REFERENCES

[1] 2012. Sketchfab. https://sketchfab.com/

[2] 2015. Autoware Foundation. https://www.autoware.org/

[3] 2019. Why Automotive Companies Should Adopt RADAR-based ADAS Systems. https://www.einfochips.com/blog/why-automotive-companies-should-adopt-radar-based-adas-systems/

[4] 2020. Baidu Apollo. https://developer.apollo.auto/

[5] 2021. CRUW Dataset. https://www.cruwdataset.org/

[6] 2021. ROD2021 Challenge. https://www.cruwdataset.org/rod2021

[7] 2021. TI AWR1843. https://www.ti.com/product/AWR1843/

[8] 2023. Stealth Aircraft. https://en.wikipedia.org/wiki/Stealth_aircraft

[9] Fahad Jibrin Abdu, Yixiong Zhang, Maozhong Fu, Yuhan Li, and Zhenmiao Deng. 2021. Application of deep learning on millimeter-wave radar signals: A review. *Sensors* 21, 6 (2021), 1951.

[10] Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick Traynor. 2021. Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems. In *2021 IEEE symposium on security and privacy (SP)*. IEEE, 730–747.

[11] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–13.

[12] Aleksandar Angelov, Andrew Robertson, Roderick Murray-Smith, and Francesco Fioranelli. 2018. Practical classification of different moving targets using automotive radar and deep neural networks. *IET Radar, Sonar & Navigation* 12, 10 (2018), 1082–1089.

[13] Giovanni Apruzzese, Hyrum Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. 2022. Position:"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice. In *IEEE Conference on Secure and Trustworthy Machine Learning*. IEEE.

[14] Maria Assunta, Giovanna Di Marzo Serugendo, Anne-Francoise Cutting-Decelle, and Martin Strohmeier. 2021. A semantic-based approach to analyze the link between security and safety for Internet of Vehicle (IoV) and Autonomous Vehicles (AVs). In *CARS 2021 6th International Workshop on Critical Automotive Applications: Robustness & Safety*.

[15] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9297–9307.

[16] Hamid Bostani and Veelasha Moonsamy. 2021. EvadeDroid: A practical evasion attack on machine learning for black-box Android malware detection. *arXiv preprint arXiv:2110.03301* (2021).

[17] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. 2021. Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks. *arXiv preprint arXiv:2106.09249* (2021).

[18] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. 2019. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2267–2281.

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

[20] Ruchir Chauhan. 2014. *A platform for false data injection in frequency modulated continuous wave radar*. Utah State University.

[21] Xingyu Chen, Zhengxiong Li, Baicheng Chen, Yi Zhu, Chris Xiaoxuan Lu, Zhenyu Peng, Feng Lin, Wenyao Xu, Kui Ren, and Chunming Qiao. 2023. MetaWave: Attacking mmWave Sensing with Meta-material-enhanced Tags. In *Network and Distributed Systems Security (NDSS) Symposium*.

[22] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. 2022. Physical attack on monocular depth estimation with optimal adversarial patches. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*. Springer, 514–532.

[23] Hongjun Choi, Sayali Kate, Yousra Aafer, Xiangyu Zhang, and Dongyan Xu. 2020. Cyber-physical inconsistency vulnerability identification for safety checks in robotic vehicles. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. 263–278.

[24] Hongjun Choi, Sayali Kate, Yousra Aafer, Xiangyu Zhang, and Dongyan Xu. 2020. Software-based Realtime Recovery from Sensor Attacks on Robotic Vehicles.. In *RAID*. 349–364.

[25] Ka-Ho Chow, Ling Liu, Margaret Loper, Juhyun Bae, Mehmet Emre Gursoy, Stacey Truex, Wenqi Wei, and Yanzhao Wu. 2020. Adversarial objectness gradient attacks in real-time object detection systems. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 263–272.

[26] Han Cui and Naim Dahnoun. 2021. High precision human detection and tracking using millimeter-wave radars. *IEEE Aerospace and Electronic Systems Magazine* 36, 1 (2021), 22–32.

[27] Xu Dong, Pengluo Wang, Pengyue Zhang, and Langechuan Liu. 2020. Probabilistic oriented object detection in automotive radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 102–103.

[28] Amrita Ghosal, Subir Halder, and Mauro Conti. 2022. Secure over-the-air software update for connected vehicles. *Computer Networks* 218 (2022), 109394.

[29] Jairo Giraldo, Sahand Hadizadeh Kafash, Justin Ruths, and Alvaro A Cardenas. 2020. Daria: Designing actuators to resist arbitrary attacks against cyber-physical systems. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 339–353.

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

[31] Chih-Chung Hsu, Chieh Lee, Lin Chen, Min-Kai Hung, Yu-Lun Lin, and Xian-Yu Wang. 2021. Efficient-ROD: Efficient Radar Object Detection based on Densely Connected Residual Network. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 526–532.

[32] Jack D Jernigan, Meltem F Kodaman, et al. 2001. *An investigation of the utility and accuracy of the table of speed and stopping distances specified in the Code of Virginia*. Technical Report. Virginia Transportation Research Council.

[33] Xiaoyu Ji, Yushi Cheng, Yuepeng Zhang, Kai Wang, Chen Yan, Wenyuan Xu, and Kevin Fu. 2021. Poltergeist: Acoustic Adversarial Machine Learning against Cameras and Computer Vision. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 160–175.

[34] Yunhan Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Hao Chen, Zhenyu Zhong, and Tao Wei. 2019. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In *International Conference on Learning Representations*.

[35] Pengfei Jing, Qiyi Tang, Yuefeng Du, Lei Xue, Xiapu Luo, Ting Wang, Sen Nie, and Shi Wu. 2021. Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*. 3237–3254.

[36] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7122–7131.

[37] Taegyu Kim, Chung Hwan Kim, Junghwan Rhee, Fan Fei, Zhan Tu, Gregory Walkup, Xiangyu Zhang, Xinyan Deng, and Dongyan Xu. 2019. RVFuzzer: Finding Input Validation Bugs in Robotic Vehicles through Control-Guided Testing.. In *USENIX Security Symposium*. 425–442.

[38] Rina Kinugawa, Kenta Imoto, Yuhei Futakawa, Shoma Shimizu, Rei Fujiwara, Marie Yoshikiyo, Asuka Namai, and Shin-ichi Ohkoshi. 2021. Highly Efficient Broadband Millimeter-Wave–Absorbing Ultrathin Films. *Advanced Engineering Materials* 23, 4 (2021), 2001473.

[39] Rony Komissarov and Avishai Wool. 2021. Spoofing attacks against vehicular FMCW radar. In *Proceedings of the 5th Workshop on Attacks and Solutions in Hardware Security*. 91–97.

[40] Belal Korany, Chitra R Karanam, Hong Cai, and Yasamin Mostofi. 2019. XModal-ID: Using WiFi for through-wall person identification from candidate video footage. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–15.

[41] Jihoon Kwon and Nojun Kwak. 2017. Human detection by neural networks using a low-cost short-range Doppler radar sensor. In *2017 IEEE Radar Conference (RadarConf)*. IEEE, 0755–0760.

[42] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271* (2018).

[43] Yun Lin, Haojun Zhao, Ya Tu, Shiwen Mao, and Zheng Dou. 2020. Threats of adversarial attacks in DNN-based modulation recognition. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2469–2478.

[44] Giulio Lovisotto, Henry Turner, Ivo Sluganovic, Martin Strohmeier, and Ivan Martinovic. 2021. {SLAP}: Improving Physical Adversarial Examples with {Short-Lived} Adversarial Perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*. 1865–1882.

[45] Mulong Luo, Andrew C. Myers, and G. Edward Suh. 2020. Stealthy tracking of autonomous vehicles with cache side channels. In *29th USENIX Security Symposium (USENIX Security 20)*.

[46] Mulong Luo and G. Edward Suh. 2022. WIP: Interrupt Attack on TEE-Protected Robotic Vehicles.

[47] Jianjun Ma, Rabi Shrestha, Wei Zhang, Lothar Moeller, and Daniel M Mittleman. 2019. Terahertz wireless links using diffuse scattering from rough surfaces. *IEEE Transactions on Terahertz Science and Technology* 9, 5 (2019), 463–470.

[48] Jiaming Mu, Binghui Wang, Qi Li, Kun Sun, Mingwei Xu, and Zhuotao Liu. 2021. A Hard Label Black-box Adversarial Attack Against Graph Neural Networks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 108–125.

[49] Prateek Nallabolu and Changzhi Li. 2021. A Frequency-Domain Spoofing Attack on FMCW Radars and Its Mitigation Technique Based on a Hybrid-Chirp Waveform. *IEEE Transactions on Microwave Theory and Techniques* 69, 11 (2021), 5086–5098.

[50] Dudi Nassi, Raz Ben-Netanel, Yuval Elovici, and Ben Nassi. 2019. MobilBye: attacking ADAS with camera spoofing. *arXiv preprint arXiv:1906.09765* (2019).

[51] AKM Iqtidar Newaz, Nur Imtiazul Haque, Amit Kumar Sikder, Mohammad Ashiqur Rahman, and A Selcuk Uluagac. 2020. Adversarial attacks to machine learning-based smart healthcare systems. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 1–6.

[52] John Nolan, Kun Qian, and Xinyu Zhang. 2021. RoS: passive smart surface for roadside-to-vehicle communication. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*. 165–178.

[53] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2016. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697* 1, 2 (2016), 3.

[54] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 372–387.

[55] Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, and Minhui Xue. 2022. Fingerprinting Deep Neural Networks Globally via Universal Adversarial Perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13430–13439.

[56] Huy Phan, Miao Yin, Yang Sui, Bo Yuan, and Saman Zonouz. 2023. CSTAR: Towards Compact and Structured Deep Neural Networks with Adversarial Robustness. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*.

[57] Raul Quinonez, Jairo Giraldo, Luis Salazar, Erick Bauman, Alvaro Cardenas, and Zhiqiang Lin. 2020. {SAVIOR}: Securing autonomous vehicles with robust physical invariants. In *29th USENIX Security Symposium (USENIX Security 20)*. 895–912.

[58] Raul Quinonez, Sleiman Safaoui, Tyler Summers, Bhavani Thuraisingham, and Alvaro A Cardenas. 2021. Shared Reality: Detecting Stealthy Attacks Against Autonomous Vehicles. In *Proceedings of the 2th Workshop on CPS&IoT Security and Privacy*. 15–26.

[59] Sandeep Rao. 2017. Introduction to mmWave sensing: FMCW radars. *Texas Instruments (TI) mmWave Training Series* (2017).

[60] Meysam Sadeghi and Erik G Larsson. 2018. Adversarial attacks on deep-learning based radio signal classification. *IEEE Wireless Communications Letters* 8, 1 (2018), 213–216.

[61] Esha Sarkar, Hadjer Benkraouda, Gopika Krishnan, Homer Gamil, and Michail Maniatakos. 2021. FaceHack: Attacking Facial Recognition Systems using Malicious Facial Characteristics. *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2021).

[62] Neetesh Saxena, Santiago Grijalva, Victor Chukwuka, and Athanasios V Vasilakos. 2017. Network security and privacy challenges in smart vehicle-to-grid. *IEEE Wireless Communications* 24, 4 (2017), 88–98.

[63] Ali Shahin Shamsabadi, Francisco Sepúlveda Teixeira, Alberto Abad, Bhiksha Raj, Andrea Cavallaro, and Isabel Trancoso. 2021. Foolhd: Fooling speaker identification by highly imperceptible adversarial disturbances. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6159–6163.

[64] Hocheol Shin, Dohyun Kim, Yujin Kwon, and Yongdae Kim. 2017. Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications. In *International Conference on Cryptographic Hardware and Embedded Systems*. Springer, 445–467.

[65] Hossein Shirazi, Indrakshi Ray, and Charles Anderson. 2020. Using machine learning to detect anomalies in embedded networks in heavy vehicles. In *Foundations and Practice of Security: 12th International Symposium, FPS 2019, Toulouse, France, November 5–7, 2019, Revised Selected Papers 12*. Springer, 39–55.

[66] Trade Shows and EDICON China. [n. d.]. mmWave Channel Modeling with Diffuse Scattering in an Office Environment. *Channels* 5 ([n. d.]), 6G.

[67] Rainer Storn. 1996. On the usage of differential evolution for function optimization. In *Proceedings of north american fuzzy information processing*. Ieee, 519–523.

[68] Pengliang Sun, Xuetong Niu, Pengfei Sun, and Kele Xu. 2021. Squeeze-and-Excitation network-Based Radar Object Detection With Weighted Location Fusion. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 545–552.

[69] Zhi Sun, Sarankumar Balakrishnan, Lu Su, Arupjyoti Bhuyan, Pu Wang, and Chunming Qiao. 2021. Who is in control? Practical physical layer attack and defense for mmWave-based sensing in autonomous vehicles. *IEEE Transactions on Information Forensics and Security* 16 (2021), 3199–3214.

[70] James Tu, Huichen Li, Xinchen Yan, Mengye Ren, Yun Chen, Ming Liang, Eilyan Bitar, Ersin Yumer, and Raquel Urtasun. 2021. Exploring Adversarial Robustness of Multi-Sensor Perception Systems in Self Driving. *arXiv preprint arXiv:2101.06784* (2021).

[71] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[72] Qian Wang, Baolin Zheng, Qi Li, Chao Shen, and Zhongjie Ba. 2020. Towards query-efficient adversarial attacks against automatic speech recognition systems. *IEEE Transactions on Information Forensics and Security* 16 (2020), 896–908.

[73] Xiaying Wang, Rodolfo Octavio Siller Quintanilla, Michael Hersche, Luca Benini, and Gagandeep Singh. 2022. Physically-Constrained Adversarial Attacks on Brain-Machine Interfaces. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

[74] Yizhou Wang, Jenq-Neng Hwang, Gaoang Wang, Hui Liu, Kwang-Ju Kim, Hung-Min Hsu, Jiarui Cai, Haotian Zhang, Zhongyu Jiang, and Renshu Gu. 2021. ROD2021 Challenge: A Summary for Radar Object Detection Challenge for Autonomous Driving Applications. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 553–559.

[75] Yizhou Wang, Zhongyu Jiang, Yudong Li, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. 2021. RODNet: A real-time radar object detection network cross-supervised by camera-radar fused object 3D localization. *IEEE Journal of Selected Topics in Signal Processing* 15, 4 (2021), 954–967.

[76] Yichao Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. A wifi vision-based 3D human mesh reconstruction. In *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking*. 814–816.

[77] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[78] Chong Xiang, Charles R Qi, and Bo Li. 2019. Generating 3d adversarial point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9136–9144.

[79] Shangyu Xie, Han Wang, Yu Kong, and Yuan Hong. 2022. Universal 3-dimensional perturbations for black-box attacks on video recognition systems. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1390–1407.

[80] Yi Xie, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan. 2021. Enabling fast and universal audio adversarial attack using generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14129–14137.

[81] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. 2020. Adversarial t-shirt! evading person detectors in a physical world. In *Proceedings of the European Conference on Computer Vision*. Springer, 665–681.

[82] Wenyuan Xu, Chen Yan, Weibin Jia, Xiaoyu Ji, and Jianhao Liu. 2018. Analyzing and enhancing the security of ultrasonic sensors for autonomous vehicles. *IEEE Internet of Things Journal* 5, 6 (2018), 5015–5029.

[83] Chen Yan, Zhijian Xu, Zhanyuan Yin, Xiaoyu Ji, and Wenyuan Xu. 2022. Rolling Colors: Adversarial Laser Exploits against Traffic Light Recognition. *arXiv preprint arXiv:2204.02675* (2022).

[84] Zihan Zhang, Mingxuan Liu, Chao Zhang, Yiming Zhang, Zhou Li, Qi Li, Haixin Duan, and Donghong Sun. 2021. Argot: generating adversarial readable chinese texts. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2533–2539.

[85] Yi Zhu, Chenglin Miao, Foad Hajiaghajani, Mengdi Huai, Lu Su, and Chunming Qiao. 2021. Adversarial Attacks against LiDAR Semantic Segmentation in Autonomous Driving. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 329–342.

[86] Yi Zhu, Chenglin Miao, Tianhang Zheng, Foad Hajiaghajani, Lu Su, and Chunming Qiao. 2021. Can We Use Arbitrary Objects to Attack LiDAR Perception in Autonomous Driving?. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 1945–1960.