Adversarial Waveform Design for Wireless Transceivers Towards Intelligent Eavesdropping

Zhenju Zhang, Mingqian Liu, *Member, IEEE*, Yunfei Chen, *Senior Member, IEEE*, Nan Zhao, *Senior Member, IEEE*, Jie Tang, *Senior Member, IEEE*, Kai-Kit Wong, *Fellow, IEEE*, and George K. Karagiannidis *Fellow, IEEE*

***Abstract*—In wireless communications, the communication channel between the transmitter and receiver can be monitored by an eavesdropper. The eavesdropper uses deep learning (DL) to quickly identify the modulation parameters of signals and further disrupt legitimate communications. Since DL has been proven to be vulnerable to adversarial attacks, this paper proposes to attack the eavesdropper's model by designing adversarial waveforms, preventing the eavesdropper from correctly identifying the modulation schemes used by legitimate users, and thereby preventing the eavesdropper from interfering with normal communications. This paper proposes an attention-based black-box attack method, which uses the prediction of different networks in the ensemble model to assign adversarial attention factors to each network. This greatly improves the transmission attack performance of the designed adversarial examples. In addition, by analysing the influence of the channel on the adversarial waveform, we further design the adversarial waveform that can be transmitted in the channel to improve the practicability of the attack algorithm. Finally, we theoretically derive the bounds of the adversarial risk increase that the attack brings to the target model. Simulation results show that the proposed method can improve the success rate of the attack on the eavesdropper's modulation detection model, cause the model to misidentify the signal modulation type, and improve the security and reliability of legitimate transceivers in wireless communication systems.**

***Index Terms*—Adversarial waveform design, deep learning, modulation recognition, black-box attack, transferability.**

I. Introduction

MODULATION recognition is a key technology in wireless communication systems. It helps signal demodulation and information recovery by identifying the communication parameters and the modulation mode of the received signal, which helps to alleviate the shortage of spectrum resources [1]–[3]. Traditional modulation recognition is based on maximum likelihood estimation and statistical patterns, relying on prior knowledge of the signals and manually extracted features. With the development of deep neural networks (DNNs), deep learning (DL) has been widely used in wireless systems, such as integrated sensing and communication [4], channel estimation [5], signal recognition [6], semantic communication [7] and specific emitter identification [8]. The modulation recognition scheme based on deep learning can automatically extract the complex features of the signal, and has extremely high recognition speed and accuracy [9]–[14]. However, due to the broadcast nature of wireless communication, the communication information may be eavesdropped [15]. For example, eavesdroppers may reconstruct the signal characteristics of the transmitter by exploiting the physical-layer feature parameters, thereby launching malicious attacks or interference, which seriously threatens the reliability of wireless communication systems. For physical layer security, Xie *et al.* employed physical-layer authentication schemes based on phase noise and tags, achieving defense against spoofing attacks, detection of impersonation attacks, and multi-user classification, thereby enhancing authentication performance and the reliability of the physical layer [16], [17]. Moreover, Nan *et al.* pointed out that adversarial waveform design is also extremely important for physical-layer security [18]. Compared with traditional encryption methods, adversarial waveforms directly increase the unrecognizability of signals at the physical layer, making it impossible for eavesdroppers to correctly identify legitimate signals, thereby protecting communication systems from eavesdropping and interference.

The eavesdropper obtains the modulation mode of the wireless signal through automatic modulation classification (AMC) using deep learning, and then generates interference to the legitimate receiver. In order to adapt to the diverse modulation schemes and complex channel conditions in wireless communication systems, the eavesdropper designs and applies more complex networks to improve the recognition performance. However, this leads to vulnerabilities in these networks within high-dimensional spaces. Therefore, it is possible to deceive the eavesdropper's intelligent recognition model according to its high-dimensional characteristics.

This work was supported by the National Natural Science Foundation of China under Grant U2441250 and 62231027, Natural Science Basic Research Program of Shaanxi under Grant 2024JC-JCQN-63, the Guangxi Key Research and Development Program under Grant 2022AB46002, the Innovation Capability Support Program of Shaanxi under Grant 2024RS-CXTD-01, the Fundamental Research Funds for the Central Universities and the Innovation Fund of Xidian University under Grant YJSJ25007. *(Corresponding author: Mingqian Liu.)*

Z. Zhang and M. Liu are with the State Key Laboratory of Integrated Service Networks, Xidian University, Shaanxi, Xi'an 710071, China (e-mail: zhenjuzhang@stu.xidian.edu.cn; mqliu@mail.xidian.edu.cn).

Y. Chen is with the Department of Engineering, University of Durham, South Road, Durham, UK, DH1 3LE (e-mail: yunfei.chen@durham.ac.uk).

N. Zhao is the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: zhaonan@dlut.edu.cn).

J. Tang is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: eejtang@scut.edu.cn).

K. K. Wong is affiliated with the Department of Electronic and Electrical Engineering, University College London, Torrington Place, WC1E 7JE, United Kingdom and he is also affiliated with Yonsei Frontier Lab, Yonsei University, Seoul, Korea. (e-mail: kai-kit.wong@ucl.ac.uk).

G. K. Karagiannidis is with Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece. (e-mail: geokarag@auth.gr).

In recent years, many researchers have discovered the vulnerability of deep learning models in the wireless domain, providing adversarial schemes for attacking eavesdroppers to protect communication systems from eavesdropping [19]–[21]. If the eavesdropper's intelligent recognition model is a white box, the communication transmitter can use the parameters of the model and the loss gradient to create an adversarial example for the model, and achieve self-protection by attacking the eavesdropper's recognition model. Sadeghi *et al.* generated white-box adversarial examples[1] against the DL-based radio signal classifier, significantly reducing the performance of the classifier with minimal perturbation power [22]. Lin *et al.* applied four gradient attack methods to intelligent modulation detection, which significantly reduced the accuracy of the DL-based modulation detection model [23]. Liu *et al.* used the feature layer and decision layer of the model to design small perturbations of wireless signals, which further improved the effect of gradient attacks [24]. However, these works did not consider channel effects such as multipath fading between the attacker and the receiver. This will affect the antagonism of the designed perturbation waveform by changing the direction and magnitude of the perturbation, resulting in the failure of the attack. Therefore, Kim *et al.* considered the channel effect and attacked the wireless signal classifier based on DL, which destroyed the channel perception accuracy of the classifier [25]. In order to evaluate the impact of adversarial attacks, many researchers have analysed the threat of adversarial examples to the robustness of the target model from different adversarial metrics [26], [27].

In practice, the eavesdropper's modulation recognition model is a black box, and its internal network parameter information is often hidden and cannot be obtained. Most of the adversarial examples generated by the traditional white-box attack method are for a specific source-target model, and their attack capabilities are highly dependent on the model. This makes it difficult to migrate to the black-box target model to implement the attack [28], [29]. For the black-box model, the commonly used attack methods mainly include substitution attack and migration attack. The substitution attack mimics the classification boundary of the target model by constructing a shadow model, and fine-tunes the classification boundary of the substitution model by querying the model [30]–[33]. However, since the output of the target receiver cannot be obtained, the black-box attack based on the substitution model is not practical in wireless communication. Ensemble attack is a commonly used transmission-based attack method. By fusing the output of each network in the ensemble model, it generates strong transferable adversarial examples that have a good attack effect on the target black-box model [34]–[36]. However, existing ensemble attack methods have not fully exploited the differences in input predictions between different networks in the ensemble model, making it difficult for the resulting adversarial examples to successfully attack target models with classification boundaries far away from the original example points.

To counter these attack methods, many adversarial defence methods have been proposed against various attacks. Zhang *et al.* studied a defence mechanism based on training time and running time for white-box attacks, which improved the robustness of DL-based modulation classifiers against adversarial attacks [37]. Nesti *et al.* proposed a method called defence perturbation to effectively detect robust adversarial examples [38] by detecting adversarial examples through transformations. Yang *et al.* used the SecureSense framework to process the input, which is generally robust to common attacks and can reduce the negative impact of adversarial attacks [39]. Therefore, in wireless communication, eavesdroppers can also adopt some defensive strategies to reduce the impact of adversarial attacks. However, there is a contradiction between the effectiveness of a defence and its generalisability. Although the defence designed for a specific attack is very efficient, its generalisability is not strong and it is difficult to defend effectively against other types of attacks [40]. The defensive methods that generally have certain defensive effects for different attacks have strong generalisability, but their defensive effects are not pronounced. Therefore, it is necessary to find new ways to increase the transferability of attacks to disable the eavesdropper's modulation detection model.

In this paper, we consider the influence of the channel environment on adversarial attacks and propose a black-box attack method based on the attention mechanism with strong transferability. The main contributions of this paper are summarised as follows:

- We propose a new attack method based on the attention mechanism. By using the prediction of different networks in the ensemble model to calculate the adversarial attention factor to adjust the movement of examples in the classification difference region, the transferability of adversarial examples to the target black-box model is improved.
- We analyse the negative impact of the channel environment on the adversarial nature of the designed perturbation waveform, and use the channel information to improve the adaptability of the adversarial waveform to the channel.
- We derive the bound of the adversarial risk increment caused by the adversarial attack on the target model, and evaluate the attack performance by testing the proximity of the adversarial loss to the bound.
- We use the proposed attack algorithm to design adversarial perturbation waveforms and transfer them to the eavesdropper's modulation recognition model to implement the self-protection attack, significantly reducing the eavesdropper's modulation recognition accuracy to offer maximum protection for the legitimate communication.

The rest of the paper is organized as follows: Section II introduces the system model of self-protective attack against the eavesdropper in wireless communication system. Section III proposes an attention-based ensemble attack method, which enhances the transferability of adversarial examples by assigning attention to different networks. Section IV considers the influence of the perturbation channel, and uses the channel

[1] The adversarial example refers to the superposition of the designed adversarial waveform onto the original clean signal.

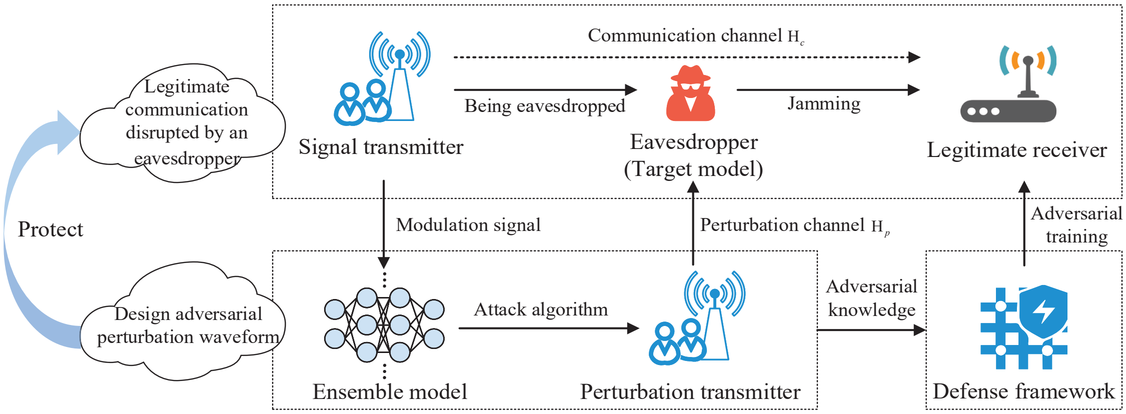


Fig. 1. Wireless communication system model with jamming and protection.

information to further enhance the adaptability of the designed perturbation waveform to the channel. Section V theoretically derives the boundary of the adversarial risk increment of the target model after being attacked to measure the performance of the generated adversarial examples. Section VI shows the performance of the proposed attack method through simulation. Section VII concludes the paper.

II. System Model

A. Communication Model

In the wireless communication system, a trained DL-based classifier at the receiver can quickly and accurately identify the category of modulation signals transmitted by the legitimate transmitter, and it can adapt well to a multi-protocol environment. However, legitimate communications can be eavesdropped and jammed. The eavesdropper employs an intelligent recognition model to identify the modulation types of the communication signals and then generates jamming signals based on the recognition results and a jamming algorithm to interfere with the legitimate receiver. It is evident that the eavesdropper's ability to intelligently recognize modulation types is a fundamental prerequisite for implementing intelligent jamming, directly affecting whether effective jamming signals can be generated to disrupt legitimate communications. Therefore, it is necessary to develop countermeasures to disrupt the eavesdropper's intelligent recognition model in order to prevent the eavesdropper from obtaining accurate modulation information. In this paper, we consider a wireless communication system consisting of a signal transmitter, a legitimate receiver, an eavesdropper and a perturbation transmitter, as shown in Fig. 1.

Fig. 1 is a system model for the design and use of adversarial waveforms for wireless transceivers. When wireless communication is performed between the transmitter and the receiver through the communication channel H_c, the signal is intercepted by the eavesdropper. Through the DL-based modulation recognition model, the eavesdropper can quickly obtain information such as the modulation category of the signal and further perform jamming on the receiver. Take the eavesdropper's DL-based modulation detection model as the target model. To protect legitimate communication, an adversarial perturbation waveform can be designed to destroy the target model. In practice, the target model is usually a black-box model, and the ensemble attack algorithm can be used to generate adversarial waveforms corresponding to the modulation signal. The perturbation transmitter sends an adversarial waveform, which is superimposed on the modulation signal at the eavesdropper after passing through the perturbation channel H_p, to deceive the target model and realise the self-protection attack on the eavesdropper. At the same time, the legitimate receiver generates adversarial examples using the same attack algorithm as the transmitter and incorporates them into the training set to train the model [41]. This enhances the model's robustness against such adversarial examples. Meanwhile, since the eavesdropper is unaware of the attack method employed between the cooperative transmitter and receiver, it cannot effectively counteract this type of attack, resulting in a significant degradation of its recognition performance.

In practical deployment, whether the perturbation transmitter needs to be integrated into the signal transmitter depends on the specific scenario requirements. When deployment is difficult or resources are limited, the perturbation waveform generator can be integrated into the signal transmitter. The generated perturbation is then superimposed on the signal and transmitted together, arriving at the receiver via the communication channel. When greater flexibility of the perturbation is required, the signal transmitter and the perturbation transmitter are separated and transmit independently. The two signals arrive at the receiver via the communication channel and the perturbation channel, respectively.

B. DL-Based Modulation Recognition Model

RADIOML2016.10B is an open-source modulation signal data set, which is often used to test the performance of modulation recognition models [42]. It has additive white Gaussian noise (AWGN), multipath fading, sampling rate offset and center frequency offset to simulate the real wireless communication environment, which is suitable for testing the influence of the adversarial attack algorithm on the modulation recognition model in this paper. The data set contains ten modulation modes, and the signal-to-noise ratio (SNR) under each modulation mode is evenly distributed in the interval [-20 dB, 18 dB] with an interval of 2 dB. These modulation modes include eight digital modulations such as 8PSK, QPSK, BPSK, GFSK, CPFSK, 4PAM, 16QAM and 64QAM, and two analog modulations such as WBFM and AM-DSB. In each modulation mode, the number of signal examples at each SNR is about 6000, including a total of 1.2 million signal examples. Each signal example has 128 pairs of I/Q sample points, which can be expressed as

$$S\left(t\right)=I\cos\left(2\pi ft\right)+Q\sin\left(2\pi ft\right), \tag{1}$$

where f is the carrier frequency.

In the modulation recognition task, DNN has the advantages of high recognition speed and high recognition accuracy. ResNet is a deep residual network that solves the problem of gradient disappearance and loss of representational capability in deep neural networks by introducing residual connections so that the network can be trained and optimised more easily [43]. The network has been shown to work well for modulation recognition tasks [44]. In this paper we choose ResNet as the modulation recognition model used by the eavesdropper. This model is a black box and can only be used to test the effect of the proposed attack algorithm. In addition, when implementing an ensemble attack, different networks must be selected to form an ensemble model. The difference between the networks helps to increase the portability of the attacks. O'Shea *et al.* used CNN, VGG, ResNet and other networks to perform the modulation detection task and showed the detection performance of different networks [45]. In this paper we use VTCNN, Inception and VGG to form an ensemble model and improve the transferability of the adversarial examples generated.

C. Adversarial Attack Model

The DL-based target modulation recognition model uses a deep neural network to predict the modulation mode of the input signal x, and its prediction loss can be expressed as:

$$\mathcal{L}\left(x,y\right)=-\sum_{k=1}^{K}y_k\left(x\right)\log\left(f_k\left(x\right)\right), \tag{2}$$

where K is the number of modulation modes, and y is the true label of the signal. $\mathcal{L}$ represents the difference between the predicted result $f\left(x\right)$ of the target model and the true label.

To disrupt the target model, the adversarial perturbation waveform is designed to increase the prediction loss of the target model. Since the direction of the loss gradient $\nabla\mathcal{L}\left(x,y\right)$ indicates the direction of increased model loss, adding a perturbation value ε to the original signal in this direction will deceive the target model into making incorrect predictions, that is, $f\left(x\right)\neq y$, which has been proven in [23]. At this point, the adversarial perturbation waveform can be expressed as:

$$\eta=\varepsilon\text{sign}\left(\nabla_x\mathcal{L}\left(x,y\right)\right). \tag{3}$$

In this paper, we consider gradient-based attacks, including iterative attacks such as the fast gradient sign method (FGSM), the basic iterative method (BIM) and the momentum iterative method (MIM), and ensemble attacks. These attack algorithms are usually subject to norm constraints, which reduce the perceptibility of the perturbation by constraining the power of the perturbation waveform. The l_p-norm constraint of the adversarial perturbation waveform is

$$\left\|\eta\right\|_p=\left(\sum_{i=1}^{n}\left|\eta_i\right|^p\right)^{\frac{1}{p}}. \tag{4}$$

When $p=0$, it represents the number of nonzero perturbation points. When $p=2$, it represents the Euclidean distance between the example before and after perturbation. When $p=\infty$, it represents the maximum perturbation value among all sampled points.

1) Iterative Attack: FGSM uses the loss gradient of the target model to determine the direction of the perturbation, and generates an adversarial example by directly adding the maximum perturbation constraint value ε in this direction. It can be expressed as

$$x^*=x+\varepsilon\text{sign}\left(\nabla_x\mathcal{L}\left(x,y\right)\right), \tag{5}$$

where x and y denote the original input and its true label, respectively, and $\nabla_x\mathcal{L}\left(x,y\right)$ denotes the loss gradient of the target model.

BIM evenly divides the perturbation level of FGSM into N segments for iteration, and continuously iterates to update the perturbation waveform. The adversarial example generated by the $(n+1)$-th iteration can be represented as

$$x_{n+1}^*=x_n^*+\alpha\text{sign}\left(\nabla_{x_n^*}\mathcal{L}\left(x_n^*,y\right)\right), \tag{6}$$

where $\alpha=\varepsilon/N$ denotes the step size of each iteration.

MIM uses the cumulant of the loss gradient instead of the loss gradient used to determine the adversarial direction in BIM, which is expressed as

$$\text{sign}\left(g_{n+1}\right)=\text{sign}\left(\mu g_n+\frac{\nabla_{x_n^*}\mathcal{L}\left(x_n^*,y\right)}{\left\|\nabla_{x_n^*}\mathcal{L}\left(x_n^*,y\right)\right\|_1}\right), \tag{7}$$

where μ denotes the momentum decay factor, g_n denotes the gradient accumulation of the n-th iteration and $g_0=0$.

2) Ensemble Attack: Ensemble attack refers to a method of generating adversarial examples using an ensemble model composed of M networks[2], which can effectively enhance the transferability of adversarial examples for attacking unknown models. In the ensemble model, the ensemble loss is obtained by combining the predictions, losses, or logits of different networks. Adversarial examples x^* are then generated using the gradient of the loss. Therefore, the goal of the ensemble attack is to deceive all networks in the ensemble model, which can be expressed as

$$\arg\max\left\{y_p^m\left(x\right)\right\}\neq y_t,\ \ \forall m\in\left\{1,2,\cdots,M\right\} \tag{8}$$

where $y_p^m\left(x\right)$ is the prediction probability vector of the m-th network, and y_t is the true class of the signal.

[2] These networks are composed of a large number of interconnected neurons, collectively forming the ensemble model, which is capable of fully extracting both deep and shallow features from the data.

According to the different output forms of the network, ensemble attacks include prediction, loss and logit. The prediction-based ensemble attack generates an adversarial example by merging the prediction probability vectors of each network in the ensemble model, which will generally be adversarial to these networks. The loss of the ensemble model is

$$\mathcal{L}(x,y) = -y \log \left(\sum_{m=1}^{M} \omega y_p^m(x) \right), \tag{9}$$

where ω is the fusion weight and $\omega = 1/M$.

The loss-based ensemble attack generates an adversarial example by fusing the predicted losses of each network. The fused loss can be expressed as

$$\mathcal{L}(x,y) = \sum_{m=1}^{M} \omega \mathcal{L}^m(x,y), \tag{10}$$

where $\mathcal{L}^m(\cdot)$ is the prediction loss of the m-th network in the ensemble model.

The logit-based ensemble attack fuses the logits of the networks in the ensemble model before normalization using the activation function softmax, preserving more primitive network output information. The loss of the ensemble model is

$$\mathcal{L}(x,y) = -y \log \left(\text{softmax} \left(\sum_{m=1}^{M} \omega l^m(x) \right) \right), \tag{11}$$

where $l^m(x)$ is the logit of the m-th network for the input in the ensemble model. Dong *et al.* proved that the adversarial examples generated by (11) have better transferability than (9) and (10) [46]. Therefore, this paper uses (11) to study attack methods that improve the transferability of adversarial examples.

III. Attention-Based Ensemble Adversarial Waveform Design

In the scenario where the perturbation generator is integrated into the signal transmitter, we employ an integrated adversarial attack algorithm to generate adversarial waveforms. These waveforms are designed to deceive the eavesdropper's recognition model based on DL. Additionally, we design an attention mechanism to enhance the effectiveness of the adversarial waveforms.

Traditional ensemble attacks achieve ensemble loss by averaging the outputs of different networks to generate adversarial examples. However, this fusion approach treats all networks in the ensemble model equally, making it difficult to fully exploit the structural differences between networks to improve the transferability of adversarial examples. For the same input, different trained networks will produce similar classification results, but there will still be a difference between the predicted probabilities. This means that for the same modulation classification task, there is always a region of classification difference between the classification boundaries of the networks. Therefore, we assign an attention factor[3] to the output of each network in the fusion process during the iterative process, adjusting the examples to cross the classification difference region.

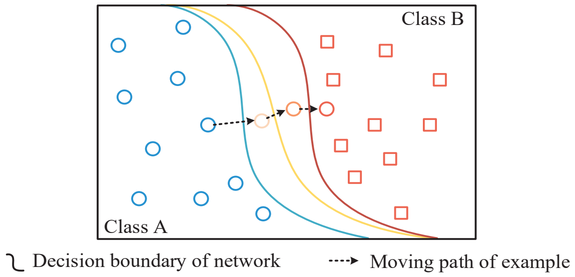


Fig. 2. The movement process of an adversarial example near the classification boundaries of Network 1 (left), Network 2 (middle) and Network 3 (right).

When adversarial examples are generated iteratively in the ensemble model, if the example can cross the classification boundary of a network, it indicates that the example can fool the network, and the attention to this network should be reduced. If the example cannot cross the classification boundary of a network, it indicates that the example is less adversarial to the network. At this point, the attention to the network should be increased in order to use the network more in the next iteration to guide the movement of adversarial examples near the classification boundary. Near the classification boundaries of different networks, the process of an example passing through the classification difference region is simply represented as Fig. 2.

In Fig. 2, the region between the classification boundaries of each net is the classification difference region. In the process of moving the example to gain adversarial power, if the example first passes the classification boundary of network 1, but is still in the classification difference region between network 1 and network 2, network 1 and network 3, the adversarial power of the example is still not enough to fool network 2 and network 3. Therefore, more attention should be paid to these networks. At the same time, the attention is continuously updated during each iteration, and the example is adjusted to pass through all the classification difference regions that can deceive all the networks in the ensemble model, thereby improving the transferability of the adversarial example to the target black-box model.

Non-target attacks maximise the loss between the prediction and the real label by designing adversarial examples to misclassify the model and reduce its reliability. In this paper, we adjust the network attention according to the prediction results of each network and calculate the attention factor used to obtain the logits and prediction loss of the ensemble model by fusion, as shown in Fig. 3.

By comparing the prediction and real labels of each network in the ensemble model for the same input, the attention accumulation is constructed by using the performance of these networks. In the $(n+1)$-th iteration, the attention accumulation

[3] The attention factor is the weight used to fuse the prediction outputs of the networks for the same example, and its value is within the interval $[0, 1]$.

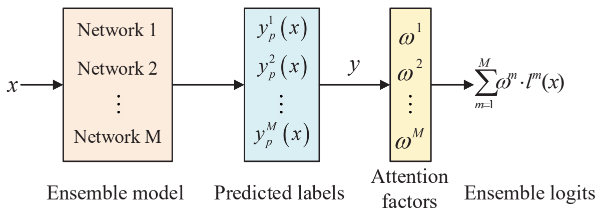


Fig. 3. Adjust the attention to fuse different network outputs to obtain the logits of the ensemble model.

of the m-th network can be expressed as

$$a_{n+1}^m = \begin{cases} a_n^m + 1, & y_p^m(x_n^*) = y, \\ \max(a_n^m - 1, 0), & y_p^m(x_n^*) \neq y, \end{cases} \tag{12}$$

where $y_p^m(\cdot)$ represents the prediction label of the m-th network. We set the initial attention accumulation $a_0^m = 0$. Accordingly, the attention factor of the m-th network can be expressed as

$$\omega_{n+1}^m = \frac{a_{n+1}^m + 1}{\sum\limits_{m=1}^{M} a_{n+1}^m + M}, \tag{13}$$

and the initial weight is set to $\omega_0^m = 1/M$. The purpose of initializing $a_0^m = 0$ and $\omega_0^m = 1/M$ is to ensure unbiased attention to different networks. This allows us to initially roughly select the networks in the ensemble model that require focused attention, that is, the networks that are difficult to deceive. Subsequently, we fine-tune the attention to different networks through continuous accumulation of attention. If these two initial values were biased towards a particular network, that network would be primarily used to generate adversarial waveforms from the outset. This would result in a slow convergence of the entire iterative process and would also reduce the transferability of the adversarial waveforms.

After fusing the logits of each network with the attention factor, the loss function of the ensemble model is proposed as

$$\mathcal{L}(x_n^*, y) = -y \log\left(\text{softmax}\left(\sum_{m=1}^{M} \omega_{n+1}^m l^m(x_n^*)\right)\right). \tag{14}$$

Then, the adversarial waveform is generated by using the loss gradient of the ensemble model as

$$\eta_{n+1} = \alpha \text{sign}\left(\mu g_n + \frac{\nabla_{x_n^*} \mathcal{L}(x_n^*, y)}{\left\|\nabla_{x_n^*} \mathcal{L}(x_n^*, y)\right\|_1}\right), \tag{15}$$

and the adversarial example of this iteration is updated by

$$x_{n+1}^* = \text{Clip}_x^{\varepsilon}\left\{x_n^* + \eta_{n+1}\right\}, \tag{16}$$

where $\text{Clip}_x^{\varepsilon}\{\cdot\}$ is used to clip the amplitude of the example to constrain the maximum power of the perturbation and ensure the concealment of the attack.

Similarly, the targeted attack minimizes the loss between the model's prediction and the targeted label in (14) to induce the model to identify the example as a specified modulation category. It is similar to the implementation of the non-targeted attack, so we will not repeat it here. It should be noted that for the specified label y_s, we need to change the attention accumulation in (12) to

$$a_{n+1}^m = \begin{cases} a_n^m + 1, & y_p^m(x_n^*) \neq y_s, \\ \max(a_n^m - 1, 0), & y_p^m(x_n^*) = y_s, \end{cases} \tag{17}$$

and get the targeted adversarial example

$$x_{n+1}^* = \text{Clip}_x^{\varepsilon}\left\{x_n^* - \eta_{n+1}\right\}. \tag{18}$$

In this way, if the adversarial example can mislead the network into misidentifying as a particular modulation category, the attention to the network will be reduced, and vice versa.

We call the proposed attention-based iterative method AIM. The attention mechanism proposed in this method can be combined with many traditional iterative attacks to improve the transferability. AIM is suitable for the scenario where the perturbation transmitter and the modulation signal transmitter are integrated. The perturbation is then superimposed on the modulation signal and transmitted together to the eavesdropper. However, in many cases, it is necessary to have two separate transmitters in order to react flexibly to the eavesdropper. For example, by comparing whether the signal modulation categories received by the receiver before and after the perturbation are consistent, one can determine whether the eavesdropper is present. Therefore, it is necessary to design an adversarial attack method that is suitable for transmission in a perturbation channel when the channel is considered separately.

IV. Channel and Attention Based Transferable Adversarial Waveform Design

In the scenario where the perturbation generator is separated from the signal transmitter, we design and optimize the perturbation waveform through channel compensation, aligning its direction with the ideal perturbation. This approach enhances the stability of the perturbation waveform in the perturbation channel.

When designing the adversarial waveform, the negative impact of the channel on the waveform cannot be ignored [25]. In this section, we improve the attention-based attack method proposed in the previous section based on the perturbation channel, so that the generated perturbation can be transmitted separately from the communication transmitter, increasing the flexibility of the self-protection attack.

After the perturbation passes through the perturbation channel, the direction and magnitude of the perturbation reaching the eavesdropper's receiver will change, which greatly reduces the attack. It is therefore necessary to consider the channel effect when designing a perturbation. From the generation process of the adversarial example, it can be seen that the perturbation direction is the sensitivity direction that makes the model most prone to error. Before and after the perturbation channel, the change in perturbation size should be minimised while ensuring that the direction of the perturbation remains unchanged to maintain the adversarial nature.

The channel between the signal transmitter and the receiver is denoted as $\text{H}_c = \text{diag}\left\{h_{c,1}, \cdots, h_{c,t}\right\}$ [47]. The perturbation channel between the perturbation transmitter and the eavesdropper is denoted as $\mathrm{H}_p = \operatorname{diag}\{h_{p,1}, \cdots, h_{p,t}\}$, which can be expressed as [48]

$$h_{p,i} = \sqrt{K\left(\frac{d_0}{d}\right)^{\gamma}\psi h_{ray,i}}, \tag{19}$$

where t is the dimension of the perturbation, K is a constant, d and d_0 are the distance between the perturbation transmitter and the eavesdropper and the reference distance, respectively, γ is the path loss index, ψ denotes shadow effect, and $h_{ray,i}$ denotes Rayleigh fading. After transmitting the designed adversarial waveform η to the eavesdropper, the signal received by the eavesdropper is the superposition of the communication signal, the perturbation signal and the complex Gaussian noise n, which can be expressed as

$$x_r = \mathrm{H}_c x + \mathrm{H}_p \eta + n. \tag{20}$$

Without considering the perturbation channel, the perturbation directly designed for the target model is denoted as the ideal disturbance η^{NoCh}, which has the greatest threat to the target model. Assuming that the channel information is known, the difference between the two can be minimized under the l_∞-norm constraint, which is expressed as

$$\begin{aligned} &\min_{\eta} \ \left\|\mathrm{H}_p\eta - \eta^{NoCh}\right\|_2^2, \\ &\text{s.t.} \ \ \|\eta\|_\infty - \varepsilon \le 0. \end{aligned} \tag{21}$$

Since the l_∞-norm constraint inequality is difficult to derive directly, it can be transformed into the l_2-norm form. According to the definition of norm, $\|\eta\|_\infty = \max\{|\eta_1|, \cdots, |\eta_t|\} \le \varepsilon$, so $\|\eta\|_2^2 = |\eta_1|^2 + \cdots + |\eta_t|^2 \le t[\max\{|\eta_1|, \cdots, |\eta_t|\}]^2 \le t\varepsilon^2$. Therefore, the constraint condition in (21) can be written as $\|\eta\|_2^2 - t\varepsilon^2 \le 0$. Since the l_2-norm is a derivable convex function, (21) is a convex optimization problem, and the Lagrangian function can be constructed as

$$L(\eta, \lambda) = \left\|\mathrm{H}_p\eta - \eta^{NoCh}\right\|_2^2 - \lambda\left(\|\eta\|_2^2 - t\varepsilon^2\right), \tag{22}$$

where λ is the Lagrange multiplier. According to (22), the Karush-Kuhn-Tucker (KKT) condition of the optimization problem is

$$\text{KKT conditions} \begin{cases} 2\mathrm{H}_p^*\left(\mathrm{H}_p\eta - \eta^{NoCh}\right) + 2\lambda\eta = 0, \\ \lambda\left(\|\eta\|_2^2 - t\varepsilon^2\right) = 0, \\ \lambda \ge 0, \\ \|\eta\|_2^2 - t\varepsilon^2 \le 0. \end{cases} \tag{23}$$

In order to study the influence of the channel on the perturbation, we obtain the relationship between the generated perturbation and the ideal perturbation according to the first equation in (23), which is expressed as

$$\left(\mathrm{H}_p^*\mathrm{H}_p + \lambda\mathrm{I}\right)\eta = \mathrm{H}_p^*\eta^{NoCh}, \tag{24}$$

where I denotes the unit diagonal matrix with dimension $t \times t$. Suppose $h_{p,i} \ne 0$ for $\forall i \in \{1, 2, \cdots, t\}$, then $h_{p,i}^* h_{p,i} = \|h_{p,i}\|_2^2 > 0$. In addition, since $\lambda \ge 0$, then $h_{p,i}^* h_{p,i} + \lambda > 0$, so the diagonal matrix $\mathrm{H}_p^*\mathrm{H}_p + \lambda\mathrm{I}$ is invertible, then

$$\eta = \left(\mathrm{H}_p^*\mathrm{H}_p + \lambda\mathrm{I}\right)^{-1}\mathrm{H}_p^*\eta^{NoCh}. \tag{25}$$

So $\eta_i = K_1 h_{p,i}^* \eta^{NoCh}$, where K_1 is a real constant and $K_1 = \left(h_{p,i}^* h_{p,i} + \lambda\right)^{-1} > 0$. Therefore, according to the optimization problem (21), by using the channel conjugate multiplied by the perturbation, the difference between the perturbation and the ideal perturbation after passing through the channel can be minimized. At this time, the perturbation arriving at the receiver through the channel is expressed as

$$\eta_{r,i} = h_{p,i}\eta_i = K_1\|h_{p,i}\|_2^2\eta^{NoCh}, \tag{26}$$

and denote it as $K_2\eta^{NoCh}$, where K_2 is a real constant and $K_2 > 0$. Therefore, the perturbation generated after this treatment is consistent with the adversarial direction of the ideal perturbation, and only the amplitude changes.

In order to further adjust the amplitude of the perturbation according to the channel to offset the influence of the channel on the perturbation, we use the channel information to adjust the global l_∞-norm constraint when iteratively generating adversarial examples. When adjusting the power of the final generated perturbation, it can be adjusted in terms of both the number of iterations and the perturbation constraint in a single iteration. However, increasing the perturbation constraint in a single iteration will result in the generated perturbation having many burrs and not being smooth compared to the clean example, which can be easily detected by the eavesdropper. Therefore, we still keep the perturbation constraint of each iteration as ε/N, and use the channel information to adjust the number of iterations to

$$N_p = \left\langle \frac{tN}{\operatorname{Tr}\left(\mathrm{H}_p\mathrm{H}_p^*\right)} \right\rangle, \tag{27}$$

where $\operatorname{Tr}(\cdot)$ denotes the trace of the matrix, and $\langle\cdot\rangle$ denotes the ceiling function. At this time, the final iteratively generated adversarial waveform is $\eta_r = \mathrm{H}_p\sum_{n=1}^{N_p}\eta_n$ after passing through the perturbation channel.

The above method of using the perturbation channel can be combined with FGSM, BIM, MIM and AIM to improve their adaptability in the perturbation channel. For example, after combining the scheme proposed in this section with AIM, we call the perturbation channel and attention-based iterative attack method PC-AIM. The non-targeted attack of PC-AIM is summarised in Algorithm 1, and the process of the targeted attack is similar.

V. Adversarial Risk Increment Bound

In this section, we analyze the adversarial risk generated by the adversarial attack on the target model, and derive the incremental upper bound of the risk to measure the adversarial quality of the perturbation that eventually reaches the eavesdropper receiver.

[26] investigated the prediction error of linear regression under adversarial attacks and derived the bounds of this error. In this paper, we extend it to the modulation recognition task using the Taylor series. Goodfellow *et al.* pointed out that the generation of an adversarial example is directly related to the high feature dimension and linear nature of the target model [49]. Therefore, in the process of iteratively generating the adversarial example x^*, the first-order Taylor expansion

Algorithm 1 PC-AIM non-targeted attack

```
Input: A clean example $x$ and true label $y$;
Input: The perturbation constraint $\varepsilon$; number of networks $M$ in the ensemble model; momentum decay factor $\mu$; original iterations $N$ and perturbation channel $\mathrm{H}_p$.
Output: An adversarial example $x^*$ with $\|x^* - x\|_\infty \le \varepsilon$.
$x_0^* = x$; $g_0 = 0$; $\omega_0^m = 1/M$; $\alpha = \varepsilon/N$;
Calculate the number of iterations $N_p$ for the perturbation channel according to (27);
for $n = 0$ to $N_p - 1$ do
    Input ${x_n}^*$ and output the logits of the $m$th network $l^m(x_n^*)$;
    Update the attention factor $\omega_{n+1}^m$ for different networks by (12) and (13);
    Obtain the logits of the ensemble model by
        $l^{ens}(x_n^*) = \sum_{m=1}^{M} \omega_{n+1}^m l^m(x_n^*);$
    Calculate the loss $\mathcal{L}(x_n^*, y)$ by (14) and loss gradient $\nabla_{x_n^*}\mathcal{L}(x_n^*, y)$ of the ensemble model;
    Update the gradient accumulation by
        $g_{n+1} = \mu g_n + \frac{\nabla_{x_n^*}\mathcal{L}(x_n^*, y)}{\|\nabla_{x_n^*}\mathcal{L}(x_n^*, y)\|_1};$
    Update the adversarial example by
        $x_{n+1}^* = \mathrm{Clip}_x^{\varepsilon}\{x_n^* + \alpha\,\mathrm{sign}(g_{n+1})\};$
end for
return $x^* = x_{N_p}^*$.
```

is used to repeatedly approximate the prediction of the DNN classifier with a linear function, and the output result of the network near the clean sample x_0 can be expressed as

$$\begin{aligned} f(x) &\approx f(x_0) + \nabla f(x_0)^{\mathrm{T}}(x - x_0) \\ &= \nabla f(x_0)^{\mathrm{T}} x + \left(f(x_0) - \nabla f(x_0)^{\mathrm{T}} x_0\right). \end{aligned} \tag{28}$$

Denoting the second term in (28) as b, the output of the network for the adversarial example with a small perturbation superimposed near x_0 can be approximated as $f(x^*) \approx \nabla f(x_0)^{\mathrm{T}} x^* + b$.

In order to measure the error of the target model's predictions with respect to the true labels instead of updating the model parameters, we use the Mean Square Error (MSE) to measure the adversarial risk of the trained model. After the model is attacked, the adversarial risk for an adversarial example generated under the l_∞-norm constraint $\|x^* - x_0\|_\infty \le \varepsilon$ can be expressed as

$$\mathcal{R}_\infty^{adv}(x^*) = \mathbb{E}_{x_0,y_0}\left[\left(y_0 - \nabla f(x_0)^{\mathrm{T}} x^* - b\right)^2\right]. \tag{29}$$

According to the perturbation $\eta = x^* - x_0$, (29) is expanded as

$$\begin{aligned} \mathcal{R}_\infty^{adv}(x^*) &= \mathbb{E}_{x_0,y_0}\left[\left(y_0 - \nabla f(x_0)^{\mathrm{T}} x_0 - b - \nabla f(x_0)^{\mathrm{T}}\eta\right)^2\right] \\ &= \mathbb{E}_{x_0,y_0}\left[\left(y_0 - \nabla f(x_0)^{\mathrm{T}} x_0 - b\right)^2\right] \\ &\quad + \mathbb{E}_{x_0,y_0}\Big[\left(\nabla f(x_0)^{\mathrm{T}}\eta\right)^2 \\ &\qquad -2\left(y_0 - \nabla f(x_0)^{\mathrm{T}} x_0 - b\right)\left(\nabla f(x_0)^{\mathrm{T}}\eta\right)\Big] \\ &= \mathcal{R}(x_0) + \mathbb{E}_{x_0,y_0}\left[r^2 - 2e_0 r\right], \end{aligned} \tag{30}$$

where $\mathcal{R}(x_0) = \mathbb{E}_{x_0,y_0}\left[\left(y_0 - \nabla f(x_0)^{\mathrm{T}} x_0 - b\right)^2\right]$ denotes the model's MSE for the clean example, $r = \nabla f(x_0)^{\mathrm{T}}\eta$ denotes the risk term associated with the perturbation, and $e_0 = y_0 - \nabla f(x_0)^{\mathrm{T}} x_0 - b$ denotes the difference between the model's predicted probability and the true label of the clean example with $e_0 \ge 0$.

Under the l_∞-norm constraint, let $\mathcal{M}(r) = r^2 - 2e_0 r$. By Hölder's inequality we have $|r| = \left|\nabla f(x_0)^{\mathrm{T}}\eta\right| \le \|\eta\|_\infty \|\nabla f(x_0)\|_1 \le \varepsilon\|\nabla f(x_0)\|_1$, i.e., $-\varepsilon\|\nabla f(x_0)\|_1 \le r \le \varepsilon\|\nabla f(x_0)\|_1$. Moreover, $r \le e_0$ can be obtained according to

$$\begin{aligned} e_0 - r &= y_0 - \nabla f(x_0)^{\mathrm{T}} x_0 - b - \nabla f(x_0)^{\mathrm{T}}\eta \\ &= y_0 - \left(\nabla f(x_0)^{\mathrm{T}} x^* + b\right) \ge 0. \end{aligned} \tag{31}$$

In addition, the adversarial risk of the model will increase after being attacked, i.e., $r^2 - 2e_0 r \ge 0$ in (30). Therefore, the risk term $r \le 0$, then

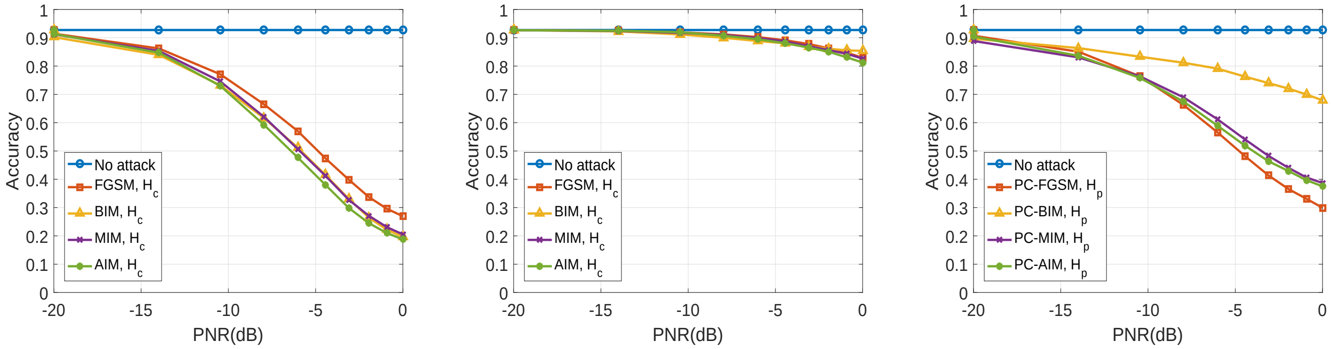
$$\begin{aligned} \mathcal{M}(r) &\le r^2 - 2e_0 r \,\big|_{r=-\varepsilon\|\nabla f(x_0)\|_1} \\ &= \varepsilon^2\|\nabla f(x_0)\|_1^2 + 2\varepsilon\|\nabla f(x_0)\|_1 e_0. \end{aligned} \tag{32}$$

According to (30), the adversarial risk increment of the target model for the adversarial example can be expressed as

$$\begin{aligned} \Delta\mathcal{R} &= \mathcal{R}_\infty^{adv}(x^*) - \mathcal{R}(x_0) \\ &\le \mathbb{E}_{x_0,y_0}\left[\varepsilon^2\|\nabla f(x_0)\|_1^2 + 2\varepsilon\|\nabla f(x_0)\|_1 e_0\right] \\ &= \varepsilon^2\mathbb{E}_{x_0,y_0}\left[\|\nabla f(x_0)\|_1^2\right] + 2\varepsilon\mathbb{E}_{x_0,y_0}\left[\|\nabla f(x_0)\|_1 e_0\right]. \end{aligned} \tag{33}$$

$\Delta\mathcal{R}$ denotes the increment of the model's prediction error after being attacked, and its upper bound can be used as a measure of the effectiveness of the attack algorithm against the target model. The closer to the upper bound, the more aggressive the adversarial example is. From (33), it can be seen that the upper bound of $\Delta\mathcal{R}$ is related to the perturbation constraint ε, the target model f, and the clean example label pair (x_0, y_0).

Further, we analyze the case when the adversarial risk increment reaches the upper bound. It can be seen from (32) that when $r = -\varepsilon\|\nabla f(x_0)\|_1$, $\mathcal{M}(r)$ reaches the maximum value, then $\nabla f(x_0)^{\mathrm{T}}\eta = -\varepsilon\|\nabla f(x_0)\|_1 = -\varepsilon\nabla f(x_0)^{\mathrm{T}}\mathrm{sign}(\nabla f(x_0))$, that is, $\eta = -\varepsilon\mathrm{sign}(\nabla f(x_0))$. It means that a perturbation of size ε is applied in the direction perpendicular to the classification boundary of the model, which can make the prediction probability of the model decrease the fastest. Thus the maximum perturbation level is applied in the ideal perturbation direction. In fact, the existing



(a) (b) (c)

Fig. 4. Modulation recognition accuracy of ResNet after the non-targeted attack. (a) The model is attacked by the perturbation generated at the communication end through the communication channel H_c. (b) The model is attacked by the perturbation generated at the communication end through the perturbation channel H_p. (c) The model is attacked by the perturbation generated at the perturbation end through the perturbation channel H_p.

attack algorithms mainly optimize the direction and size of the perturbation to approximate the ideal adversarial effect as much as possible.

VI. Simulation Results and Analysis

In this section, we investigate the impact of eavesdroppers on legitimate communication systems. We conduct simulation experiments under the system model shown in Fig. 1, employing deep learning models to simulate the threat posed by eavesdroppers to legitimate communications, which has been proven to be feasible [50], [51]. We use the signal examples in RADIOML2016.10B to verify the attack performance of the proposed method for eavesdroppers. These examples include additive white Gaussian noise (AWGN), multipath fading, sampling rate offset and center frequency offset, which effectively simulate the real wireless communication environment. In the following, we verify the attack effect of the designed perturbation waveforms without considering the perturbation channel or with considering the perturbation channel, respectively, and the attack modes include non-targeted attack and targeted attack.

Before the adversarial attack, we utilize VTCNN, Inception, and VGG to construct the ensemble model, and select ResNet as the target model to be attacked. When training the network, 80% of the examples in the dataset are used as the training set, and the remaining examples are used as the test set. We set the batch and epoch to 1024 and 100, respectively. In addition, our experiments were conducted on an NVIDIA GeForce RTX 3080Ti GPU, with models trained using the TensorFlow 2.0 framework. During training, the Adam optimizer was employed, with an initial learning rate set at 0.001 and an automatic update mechanism in place. When generating adversarial waveforms, we set the momentum decay factor and the number of iterations to 1.0 and 10, respectively.

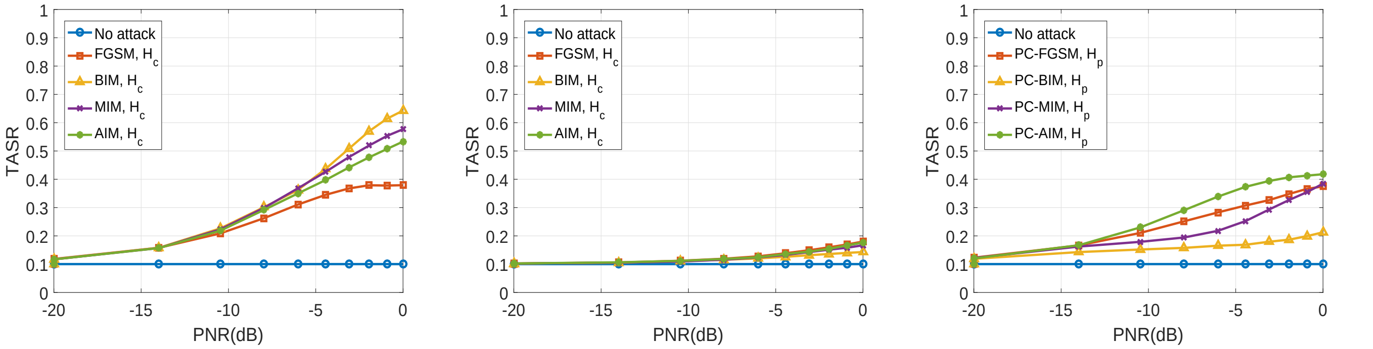
A. Perturbation-to-Noise Ratio

When implementing a self-protective attack on an eavesdropper's modulation recognition model, the concealment of the attack should be ensured so that it is not perceived and defended by the eavesdropper. The perturbation amplitude represents the strength of the attack. If it is too strong, it will reduce the concealment of the attack. Therefore, the attack under the condition that the perturbation perception is invisible is beneficial to covertly destroy the eavesdropper's recognition model. The perturbation-to-noise ratio (PNR) can be used to measure the power of the perturbation relative to the noise, which is defined as [52]

$$\mathrm{PNR}\,[\mathrm{dB}] = \frac{\mathbb{E}\left[\|\varepsilon\|_2^2\right]}{\mathbb{E}\left[\|x\|_2^2\right]}\,[\mathrm{dB}] + \mathrm{SNR}\,[\mathrm{dB}]\,. \tag{34}$$

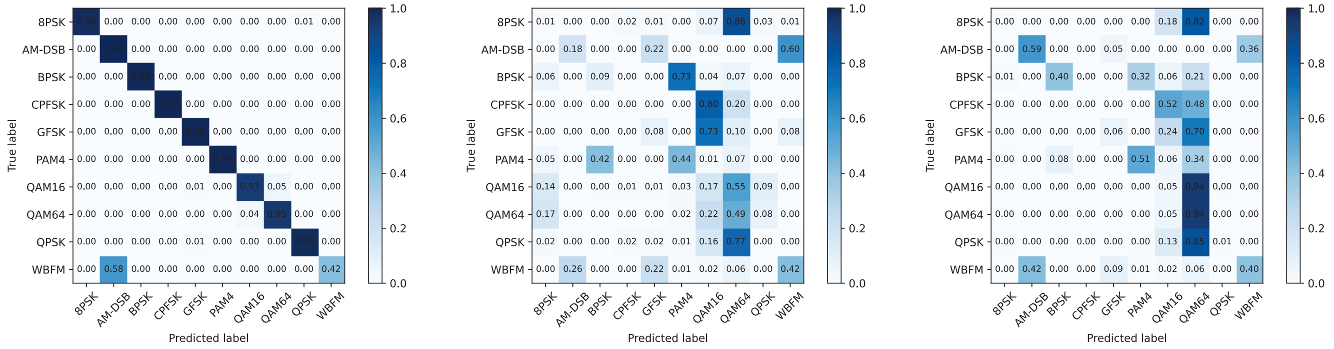
When PNR is less than 0 dB, the adversarial waveform is considered imperceptible [53]. Therefore, we use the PNR distributed in [-20 dB, 0 dB] to constrain the power of the perturbation, thereby generating an imperceptible adversarial waveform to covertly attack the target model. In different channel scenarios, we use the traditional average fusion method and FGSM, BIM, MIM to generate adversarial examples in the ensemble model, and use the proposed attention mechanism and AIM to generate adversarial examples to test the effect of non-targeted attack and targeted attack on the target model. The recognition accuracy of the model before and after being subjected to non-targeted attacks is shown in Fig. 4, and the attack success rate of implementing targeted attacks on the recognition model is shown in Fig. 5.

Fig. 4 shows the effect of the proposed algorithm on the non-targeted attack on the recognition model of the eavesdropper. It can be seen that the perturbations significantly reduce the model accuracy when they are generated at the transmitter and pass through H_c to attack the target model. However, when the perturbations pass through H_p to perform the attack, their threat to the model is significantly reduced. At this point, the perturbations are generated at the perturbation end using the channel-based enhanced adversarial approach in Section IV, and these perturbations again reduce the model accuracy, which indicates that these perturbations retain their adversarial nature after passing through the perturbation channel. The effect of PC-FGSM in Fig. 4(c) is essentially the same as that of FGSM in Fig. 4(a), which is due to the fact that FGSM is a single-step iterative attack whose perturbation constraints based on channel tuning are directly used in their entirety to



(a) (b) (c)

Fig. 5. Success rate of the targeted attack on ResNet, and the specified modulation type is 64QAM. (a) The model is attacked by the perturbation generated at the communication end through the communication channel H_c. (b) The model is attacked by the perturbation generated at the communication end through the perturbation channel H_p. (c) The model is attacked by the perturbation generated at the perturbation end through the perturbation channel H_p.



(a) (b) (c)

Fig. 6. The prediction confusion matrix of ResNet for clean signals and AIM adversarial examples when PNR = 0 dB. (a) No attack. (b) AIM non-targeted attack. (c) AIM targeted attack.

counteract the effects of the perturbation channel. Therefore, the perturbations generated by PC-FGSM retain the original perturbation information after passing through the perturbation channel.

In Fig. 5, we choose 64QAM as the specified modulation type in the targeted attack, and the proportion of all examples recognized as this modulation type by the model of the eavesdropper is denoted as the targeted attack success rate (TASR). It can be seen that although AIM is not the best in Fig. 5(a), in Fig. 5(c) of the scenario where the perturbation channel is considered, PC-AIM makes optimal under different perturbation constraints, which shows the advantage of PC-AIM in the perturbation channel environment. In addition, we record the generation time of an AIM adversarial waveform and a PC-AIM adversarial waveform, which are 0.056 s and 0.098 s, respectively, indicating that the proposed method has good generation efficiency and can meet the real-time requirements in practical applications. In addition, the generation time of adversarial waveforms can be further shortened by selecting the lateral network in the integrated model, which will be explained in section VI.B. Therefore, the proposed adversarial waveform design method is feasible in practical applications.

We use the confusion matrix in Fig. 6 to visually display the modulation recognition results of the target model before and after the AIM attack. In Fig. 6(a), the unattacked recognition model tends to misclassify WBFM as AM-DSB. This is primarily due to the silent periods present when generating these two types of data by sampling simulated audio signals. During the data sampling process, the data samples of WBFM and AM-DSB signals, due to the existence of intermittent silent phases, only retain the carrier feature parameters. This results in significant confusion in the recognition of the modulation types of the two signals [54]. Fig. 6(b) is the prediction matrix of the model after the non-targeted attack. The smaller the probability on the diagonal of the matrix, the greater the prediction error of the model, the better the effect of non-targeted attack. Fig. 6(c) is the prediction matrix of the model after the targeted attack. The greater the probability on the column corresponding to the specified modulation category in the matrix, the better the effect of the target attack. It can be seen that the non-targeted attack greatly reduces the reliability of the target model. It is worth noting that in Fig. 6(c), when the specified category is digital modulation 64QAM, the targeted attack success rates of digital modulation QPSK and 16QAM are as high as 0.85 and 0.94, respectively, while the targeted attack success rates of analog modulation WBFM and AM-DSB are only 0.06 and 0.00. This means that when the modulation type is specified as digital modulation, the

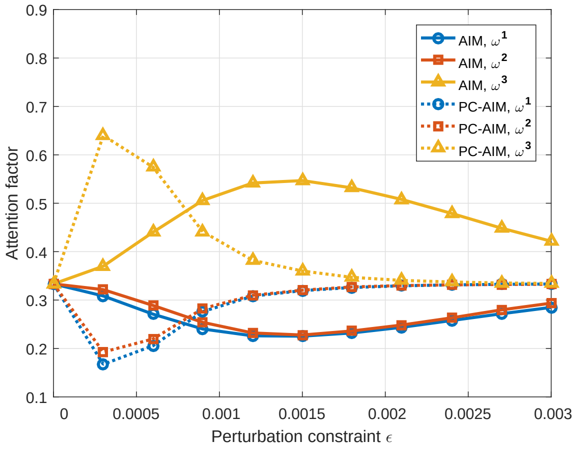


Fig. 7. Attention factor of different networks in the ensemble model.

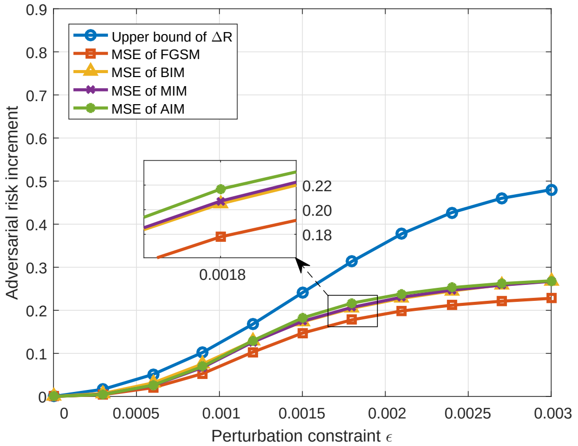
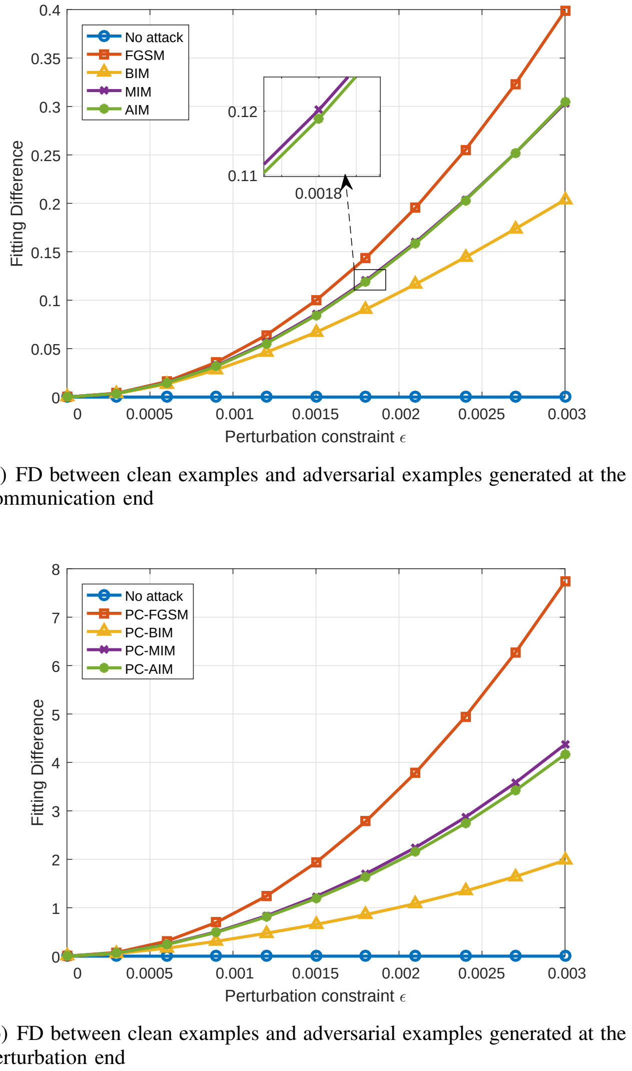


Fig. 8. Adversarial risk increment of the target model after being attacked under different perturbation constraints.



(a) FD between clean examples and adversarial examples generated at the communication end

(b) FD between clean examples and adversarial examples generated at the perturbation end

Fig. 9. FD between adversarial examples and clean examples.

designed adversarial example is difficult to disguise as analog modulation, because it usually requires greater perturbation power, which is difficult to achieve under the norm constraint used to ensure the invisibility of the attack. Therefore, the targeted attack is applicable to the same modulation mode.

In the following, we focus on the non-targeted attacks and conduct simulations from three aspects: attention factors in the ensemble model, adversarial risk increment of the target model, and waveform correlation between adversarial examples and clean modulation signals to verify the effectiveness of the proposed method.

B. Attention Factor

In order to study the distribution of the classification boundaries of different networks in the ensemble model, we record the attention factors assigned to these networks by the attention-based non-targeted attack under different perturbation constraints, as shown in Fig. 7.

Fig. 7 shows the variation of the attention factors assigned to different networks in the ensemble model with the perturbation constraints when the adversarial waveforms are generated by AIM and PC-AIM, respectively. The larger the attention factor is, the more difficult the network is to be fooled, and the farther the classification boundary is from the original example. When the perturbation constraint is zero, we allocate the attention to each network on average, and set the attention factor to be 0.33. As the perturbation constraint increases, the attention factor approaches 0.33 after adjustment. This is because it can be seen from Fig. 4 that the increase in perturbation power makes the example more adversarial and gradually has the ability to fool all networks in the ensemble model. Therefore, it can be seen from (12) that the attention accumulation of each network tends to be equal. In Fig. 7, the weight of Network 3 is always the largest, indicating that its classification boundary is the farthest from the original example and is at the outermost side of the classification difference region. Similarly, Network 1 is closest to the original example. Therefore, it is possible to directly use Network 1 and Network 3 to form an ensemble model, which will obtain the largest region of classification variability, produce examples with similar transferability as an ensemble model composed of multiple networks, and

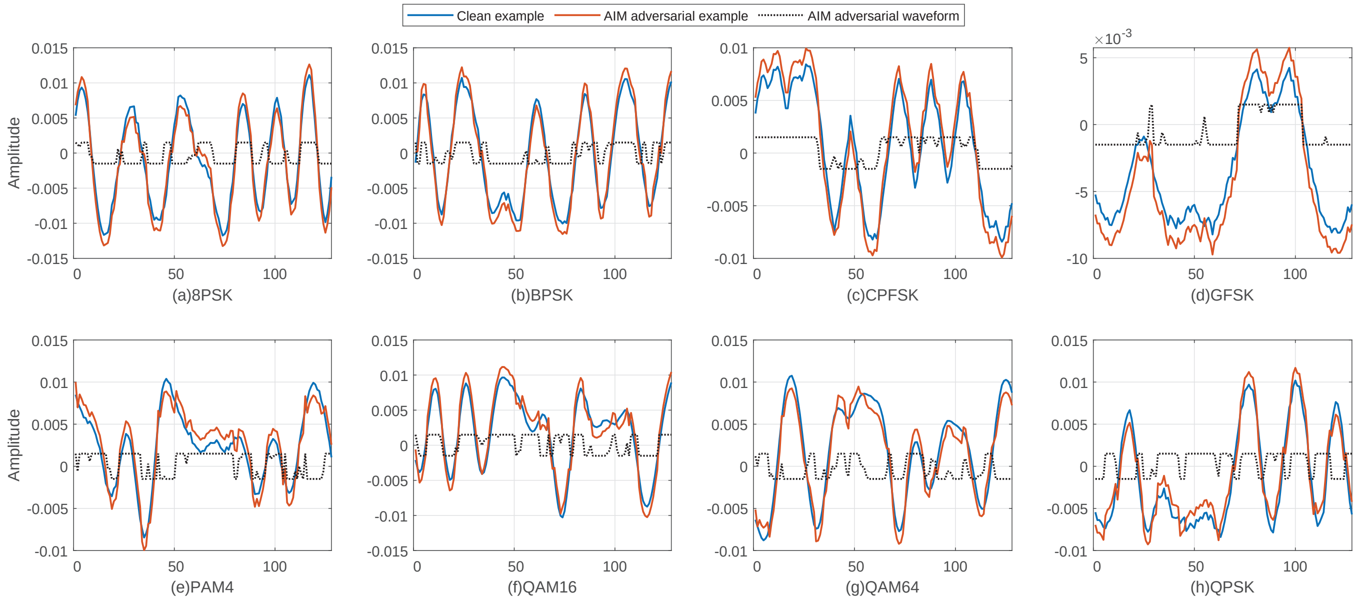


Fig. 10. Time domain waveforms of eight modulation signals before and after AIM attack and their adversarial waveforms.

substantially improve the efficiency of perturbation generation.

C. Adversarial Risk Increment

We use the adversarial risk incremental bound derived in Section V as a benchmark to test the threat of adversarial examples to the target model. The difference between the MSEs of the model for the adversarial examples and the clean examples, respectively, is taken as the actual adversarial risk increment generated by the attack, and the closer it is to the incremental bound, the closer its adversarial nature is to the ideal attack. In order to observe the effect of the attack, we use the sigmoid function to process the risk increment to unify the dimension. The actual adversarial risk increment and the incremental upper bound for different attacks under different perturbation constraints are shown in Fig. 8.

In Fig. 8, the proposed AIM is closer to the upper bound of the risk increment than other attacks, indicating the effectiveness of its adversarial examples. However, as the perturbation constraint increases, the actual risk increment generated by these attacks gradually moves away from the upper bound of the risk increment. This is because the upper bound is generated under ideal conditions, and at the end of the actual iterative attack process, the direction and size of the perturbation are usually difficult to achieve the global optimal point. This allows the adversarial example to deceive the target model, but there is still a deviation from the ideal perturbation. Nevertheless, the upper bound of adversarial risk increment can still be used to measure the quality of adversarial examples. In fact, as the invisibility of attacks tends to deteriorate with the increase of perturbation constraints, we usually focus on generating attacks with small perturbation constraints.

D. Waveform Correlation

In addition to the attack success rate, concealment is another key indicator to measure the attack effect. It is reflected in the correlation between the adversarial example and the clean example, which is directly related to whether the designed perturbation waveform can be detected by the eavesdropper. We use fitting difference (FD) to quantitatively analyze the similarity between adversarial examples and clean examples to compare the concealment effect of different attacks, which is expressed as [55]

$$FD\left(s, s^{*}\right)=\frac{\sum\limits_{i=1}^{N_s}\left(s_i-s_i^{*}\right)^2}{\sum\limits_{i=1}^{N_s}\left(s_i-\bar{s}\right)^2}, \tag{35}$$

where N_s is the length of the signal example, s and $\bar{s}$ are the original signal example and its average value, respectively, that is, $\bar{s}=\sum_{i=1}^{N_s} s_i/N_s$, and s^* is the adversarial example. The smaller the FD, the greater the similarity between the adversarial example and the clean example.

We use AIM and PC-AIM to generate adversarial examples, respectively, and calculate the FD of these adversarial examples, as shown in Fig. 9. From Fig. 9(a), it can be seen that for the adversarial waveform generated at the communication end, FGSM has the largest FD and the worst concealment, which is most easily detected by the eavesdropper. AIM has a slightly smaller FD than MIM, and its concealment is slightly better than that of MIM. BIM has the smallest FD and the best concealment, but it can be seen from Fig. 4 that it has a poorer attack performance than AIM. For the adversarial waveform generated by PC-AIM at the perturbation end, there is a similar analysis.

When SNR $=$ 10 dB and $\varepsilon=0.0018$, we plot the time domain waveforms of AIM perturbations and adversarial examples designed for eight digital modulation signals, as shown in Fig. 10. According to (34), PNR $= -4.89$ dB at this point. It can be seen that since the perturbation amplitude is constrained by the l_∞-norm, the adversarial example does not change much compared with the clean example, but it can be seen from Fig. 4(a) that this can reduce the accuracy of the target model by more than 50%, preventing the eavesdropper from correctly obtaining the modulation information of the communication.

VII. Conclusion

In this paper, we have designed an adversarial waveform for the modulated signal to solve the problem that the legitimate wireless signal transmitter and receiver are vulnerable to eavesdropper monitoring and jamming. First, we designed an attention-based iterative attack scheme. By using the performance of different networks in the ensemble model for the same example, we accumulated and updated the attention for each network, gradually crossing the classification difference regions to improve the transferability of the attack. Then, we further analysed the negative impact of the perturbation channel on the attacks, and used the channel information to compensate for changes in the adversary's power. Finally, we theoretically derived the adversarial risk limit of the target model after the attack and intuitively measured the performance of adversarial examples generated by the attack algorithms. The simulation results show that the method proposed can effectively attack the eavesdropper's illegal modulation recognition model by designing adversarial perturbation waveforms, protecting the security of legitimate communication.

References

[1] B. Li, S. Li, J. Hou, J. Fu, C. Zhao and A. Nallanathan, "A Bayesian approach for adaptively modulated signals recognition in next-generation communications," *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4359-4372, Aug. 2015.

[2] A. O. A. Salam, R. E. Sheriff, Y.-F. Hu, S. R. Al-Araji and K. Mezher, "Automatic modulation classification using interacting multiple model Kalman filter for channel estimation," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8928-8939, Sep. 2019.

[3] H. Zhang, M. Liu, Y. Chen and N. Zhao, "Blockchain and timely auction mechanism-based spectrum management," *Future Gener. Comput. Syst.*, vol. 166, 2025, DOI: 10.1016/j.future.2024.107703.

[4] J. Zhang, W. Lu, C. Xing, N. Zhao, N. Al-Dhahir, G. K. Karagiannidis and X. Yang, "Intelligent integrated sensing and communication: a survey," *Sci. China Inf. Sci.*, vol. 68, no. 3, p. Art. no. 131301, Mar. 2025.

[5] J. Wang, G. Gui, T. Ohtsuki, B. Adebisi, H. Gacanin and H. Sari, "Compressive sampled CSI feedback method based on deep learning for FDD massive MIMO systems," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5873-5885, Sep. 2021.

[6] Y. Dong, X. Jiang, H. Zhou, Y. Lin and Q. Shi, "SR2CNN: Zero-shot learning for signal recognition," *IEEE Trans. Signal Process.*, vol. 69, pp. 2316-2329, 2021.

[7] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6227-6240, Sep. 2023.

[8] P. Tang, G. Ding, Y. Xu, Y. Jiao, Y. Song and G. Wei, "Causal learning for robust specific emitter identification over unknown channel statistics," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 5316-5329, 2024.

[9] T. Chen *et al.*, "EMD and VMD empowered deep learning for radio modulation recognition," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 1, pp. 43-57, Feb. 2023.

[10] L. Zhang, X. Yang, H. Liu, H. Zhang and J. Cheng, "Efficient residual shrinkage CNN denoiser design for intelligent signal processing: Modulation recognition, detection, and decoding," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 97-111, Jan. 2022.

[11] W. Zhang, X. Yang, C. Leng, J. Wang and S. Mao, "Modulation recognition of underwater acoustic signals using deep hybrid neural networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 5977-5988, Aug. 2022.

[12] Y. Wang, G. Gui, T. Ohtsuki and F. Adachi, "Multi-task learning for generalized automatic modulation classification under non-Gaussian noise with varying SNR conditions," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3587-3596, Jun. 2021.

[13] Y. Lin, Y. Tu and Z. Dou, "An improved neural network pruning technology for automatic modulation classification in edge devices," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5703-5706, May 2020.

[14] M. Zhang, P. Tang, G. Wei, X. Ni, G. Ding and H. Wang, "Open set domain adaptation for automatic modulation classification in dynamic communication environments," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 3, pp. 852-865, Jun. 2024.

[15] M. Z. Hameed, A. György and D. Gündüz, "The best defense is a good offense: Adversarial attacks to avoid modulation detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 1074-1087, 2021.

[16] N. Xie, W. Xiong, J. Chen, P. Zhang, L. Huang and J. Su, "Multiple phase noises physical-layer authentication," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6196-6211, Sep. 2022.

[17] N. Xie, M. Sha, T. Hu and H. Tan, "Multi-user physical-layer authentication and classification," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6171-6184, Sep. 2023.

[18] G. Nan *et al.*, "Physical-layer adversarial robustness for deep learning-based semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2592-2608, Aug. 2023.

[19] Z. Bao, Y. Lin, S. Zhang, Z. Li and S. Mao, "Threat of adversarial attacks on DL-based IoT device identification," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 9012-9024, 1 Jun., 2022.

[20] Y. Shi and Y. E. Sagduyu, "Membership inference attack and defense for wireless signal classifiers with deep learning," *IEEE Trans. Mob. Comput.*, vol. 22, no. 7, pp. 4032-4043, 1 Jul. 2023.

[21] R. Sahay, M. Zhang, D. J. Love and C. G. Brinton, "Defending adversarial attacks on deep learning-based power allocation in massive MIMO using denoising autoencoders," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 4, pp. 913-926, Aug. 2023.

[22] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 213-216, Feb. 2019.

[23] Y. Lin, H. Zhao, X. Ma, Y. Tu and M. Wang, "Adversarial attacks in modulation recognition with convolutional neural networks," *IEEE Trans. Reliab.*, vol. 70, no. 1, pp. 389-401, Mar. 2021.

[24] M. Liu, Z. Zhang, Y. Chen, J. Ge and N. Zhao, "Adversarial attack and defense on deep learning for air transportation communication jamming," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 973-986, Jan. 2024.

[25] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek and S. Ulukus, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3868-3880, Jun. 2022.

[26] A. H. Ribeiro and T. B. Schön, "Overparameterized linear regression under adversarial attacks," *IEEE Trans. Signal Process.*, vol. 71, pp. 601-614, 2023.

[27] B. Flowers, R. M. Buehrer and W. C. Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1102-1113, 2020.

[28] Y. Shi, Y. Han, Q. Hu, Y. Yang and Q. Tian, "Query-efficient black-box adversarial attack with customized iteration and sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2226-2245, Feb. 2023.

[29] Y. Zhang, X. Tian, Y. Li, X. Wang and D. Tao, "Principal component adversarial example," *IEEE Trans. Image Process.*, vol. 29, pp. 4804-4815, 2020.

[30] R. Zhang, H. Xia, C. Hu, C. Zhang, C. Liu and F. Xiao, "Generating adversarial examples with shadow model," *IEEE Trans. Ind. Inf.*, vol. 18, no. 9, pp. 6283-6289, Sep. 2022.

[31] Y. Dong, S. Cheng, T. Pang, H. Su and J. Zhu, "Query-efficient black-box adversarial attacks guided by a transfer-based prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9536-9548, Dec. 2022.

[32] Z. Luo, S. Zhao, Z. Lu, J. Xu and Y. Sagduyu, "When attackers meet AI: Learning-empowered attacks in cooperative spectrum sensing," *IEEE Trans. Mobile Comput.*, vol. 21, no. 5, pp. 1892-1908, May 2022.

[33] M. Duan, K. Jiao, S. Yu, Z. Yang, B. Xiao and K. Li, “MC-Net: Realistic sample generation for black-box attacks,” *IEEE Trans. Inf. Forensics Secur.*, vol.19, pp. 3008-3022, 2024.

[34] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1-14.

[35] Y. Xiong, J. Lin, M. Zhang, J. E. Hopcroft and K. He, “Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14963-14972.

[36] Z. Che *et al.*, “SMGEA: A new ensemble adversarial attack powered by long-term gradient memories,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1051-1065, Mar. 2022.

[37] L. Zhang, S. Lambotharan, G. Zheng, G. Liao, A. Demontis and F. Roli, “A hybrid training-time and run-time defense against adversarial attacks in modulation classification,” *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1161-1165, Jun. 2022.

[38] F. Nesti, A. Biondi and G. Buttazzo, “Detecting adversarial examples by input transformations, defense perturbations, and voting,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 3, pp. 1329-1341, Mar. 2023.

[39] J. Yang, H. Zou and L. Xie, “SecureSense: Defending adversarial attack for secure device-free human activity recognition,” *IEEE Trans. Mob. Comput.*, vol. 23, no. 1, pp. 823-834, Jan. 2024.

[40] D. Wang, C. Li, S. Wen, S. Nepal and Y. Xiang, “Defending against adversarial attack towards deep neural networks via collaborative multi-task training,” *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 2, pp. 953-965, Mar. 2022.

[41] Z. Chen *et al.*, “Learn to defend: Adversarial multi-distillation for automatic modulation recognition models,” *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 3690-3702, Feb. 2024.

[42] DeepSig, “Deepsig dataset: Radioml 2016.10b,” [Online] Available: https://www.deepsig.io/datasets, 2016.

[43] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.

[44] Z. Yu, J. Tang and Z. Wang, “GCPS: A CNN performance evaluation criterion for radar signal intrapulse modulation recognition,” *IEEE Commun. Lett.*, vol. 25, no. 7, pp. 2290-2294, Jul. 2021.

[45] T. J. O’Shea, T. Roy and T. C. Clancy, “Over-the-air deep learning based radio signal classification,” *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 1, pp. 168-179, Feb. 2018.

[46] Y. Dong *et al.*, “Boosting adversarial attacks with momentum,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185-9193.

[47] T. J. O’Shea and N. West, “Radio machine learning dataset generation with GNU radio,” in *Proc. 6th GNU Radio Conf.*, 2016, pp. 1-6.

[48] S. Sinha and A. Soysal, “Channel aware adversarial attacks are not robust,” in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2023, pp. 1-6.

[49] I. Goodfellow, J. Shlens and C. Szegedy, “Explaining and harnessing adversarial examples,” in *Proc. Int. Conf. Learn. Represent.*, Mar. 2015, pp. 189-199.

[50] P. Qi, Y. Meng, S. Zheng, X. Zhou, N. Cheng and Z. Li, “Adversarial defense embedded waveform design for reliable communication in the physical layer,” *IEEE Internet Things J.*, vol. 11, no. 10, pp. 18136-18153, May 2024.

[51] R. Jiang, W. Rao and S. Chen, “TRANS-G: Transformer generator for modeling and constructing of UAPs against DNN-based modulation classifiers,” *IEEE Trans. Veh. Technol.*, vol. 73, no. 11, pp. 16892-16904, Nov. 2024.

[52] R. Sahay, C. G. Brinton and D. J. Love, “A deep ensemble-based wireless receiver architecture for mitigating adversarial attacks in automatic modulation classification,” *IEEE Trans. Cognit. Commun. Netw.*, vol. 8, no. 1, pp. 71-85, Mar. 2022.

[53] P. F. de Araujo-Filho, G. Kaddoum, M. Chiheb Ben Nasr, H. F. Arcoverde and D. R. Campelo, “Defending wireless receivers against adversarial attacks on modulation classifiers,” *IEEE Internet Things J.*, vol. 10, no. 21, pp. 19153-19162, Nov. 2023.

[54] J. Xu, C. Luo, G. Parr and Y. Luo, “A spatiotemporal multi-channel learning framework for automatic modulation recognition,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1629-1632, Oct. 2020.

[55] H. Zhao, Y. Lin, S. Gao and S. Yu, “Evaluating and improving adversarial attacks on DNN-based modulation recognition,” in *IEEE Glob. Commun. Conf.*, Dec. 2020, pp. 1-5.