Robust Generative Defense Against Adversarial Attacks in Intelligent Modulation Recognition

Zhenju Zhang, Linru Ma, Mingqian Liu, *Member, IEEE*, Yunfei Chen, *Senior Member, IEEE*, Nan Zhao, *Senior Member, IEEE*, and Arumugam Nallanathan, *Fellow, IEEE*

***Abstract*—Deep neural network (DNN) greatly improves the efficiency of modulation recognition in wireless communication, but it also suffers from attacks. Generative artificial intelligence (GAI) possesses powerful data generation capabilities, which can be used to defend against attacks in modulation recognition. In practical scenarios, black box attack can be implemented without information on the model. This is a great security threat. The existing defense methods are difficult to improve the robustness of the model while ensuring the recognition accuracy of the original signals. Therefore, this paper uses GAI to propose an adversarial decoupled defense method to protect modulation recognition. Firstly, for weak adversarial perturbations, the empirical mode decomposition (EMD) is used to highlight the high-frequency features in the signal, and the adversary detector is designed to detect the suspiciousness. Then, the signal is regenerated based on the generative adversarial network (GAN) to weaken the antagonism in the example. Further, the traditional adversarial training is decoupled into an original branch and an adversarial branch, and the outputs of the two branches are fused according to the suspiciousness. Simulation results show that the proposed defense method has high recognition accuracy for both original examples and adversarial examples even under attacks, and can effectively improve the robustness of the intelligent recognition model.**

***Index Terms*—Adversarial attack, adversarial defense, generative artificial intelligence, intelligent modulation recognition, generative adversarial network.**

I. Introduction

COMMUNICATION devices in wireless networks produce more and more data, which increases the difficulty of communication data processing. Artificial intelligence (AI) gives wireless communication systems the ability to automatically process communication data, serving as a crucial tool for intelligent allocation of communication and network resources, greatly enhancing the efficiency of data processing and communication [1], [2], [3], [4], [5]. In wireless communication networks, automatic modulation recognition (AMR) is a key technology in cognitive radio and non-cooperative communication, which provides important information for subsequent steps such as demodulation. However, the traditional AMR relies on manually extracted features or prior knowledge of the channel, and the recognition efficiency needs to be improved.

In recent years, many researchers have developed deep learning (DL) methods for recognition, using deep neural network (DNN) to extract deep features of signals and classify them, which has greatly improved the recognition speed and accuracy [6], [7]. In communication scenarios such as cognitive radio [8], edge computing [9], interference recognition [10] and radio monitoring [11], intelligent recognition models based on DL have shown great advantages. For the AMR system with complex and variable channels, DL can automatically extract the features from the signal and accurately identify the modulation mode of the signal by using its powerful nonlinear mapping ability [12], [13], [14], [15]. To cope with the diversity and dynamic changes in signal data distribution in actual wireless communication environments, Zhang *et al.* introduced unknown categories into the source classifier, achieving feature separation for both known and unknown modulation categories, thereby enhancing the accuracy of modulation recognition in practical communication environments [16].

However, the intelligent model based on DL has been proved to be vulnerable to adversarial attacks [17], [18]. Attackers can use the broadcast nature of the wireless channel to inject carefully designed small adversarial perturbations into the receiver, which seriously affects the reliability of the intelligent recognition model. Depending on the attack stage, common attacks include poisoning attacks that contaminate training data during the training phase and evasion attacks that mislead model prediction during the prediction phase. Since the poisoning attack needs to understand the training set information of the target model, which is usually difficult due to data privacy, the evasion attack is more threatening to the intelligent modulation recognition model.

According to the understanding of the target model, common evasion attacks usually include white-box attacks against known models and black-box attacks against unknown models. Lin *et al.* introduced the fast gradient sign method (FGSM), the basic iterative method (BIM) and the momentum iterative method (MIM) from the field of image processing to the

This work was supported by the National Natural Science Foundation of China under Grant U2441250, 62231027 and 62071364, Natural Science Basic Research Program of Shaanxi under Grant 2024JC-JCQN-63, the Guangxi Key Research and Development Program under Grant 2022AB46002 and Innovation Capability Support Program of Shaanxi under Grant 2024RS-CXTD-01. *(Corresponding author: Linru Ma.)*

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).

L. Ma is with the Institute of Systems Engineering, AMS, Beijing 100071, China (e-mail: malinru@163.com).

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).

A. Nallanathan is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London and also with the Department of Electronic Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 17104, Korea (mailto:a.nallanathan@qmul.ac.uk).

field of communications, proving the vulnerability of the DL-based modulation recognition model [19]. Liu *et al.* proposed a dynamic iterative method (DIM) to attack the white-box modulation classifier, which improved the success rate of attack [20]. Ke *et al.* limited the adversarial perturbation to a narrow frequency band, so that the filter could not filter it out, and covertly attacked the intelligent modulation recognition model [21]. In practice, attackers do not know the structure and parameters of the model, and often use the migration of adversarial examples to perform black-box attacks. Hu *et al.* proposed a substitute meta-learning black-box attack method, which combines meta-learning with the training of surrogate models to improve training efficiency and attack performance [22]. Dong *et al.* generated adversarial examples on the reconstructed surrogate model and transferred them directly to the unreachable black-box model [23].

To reduce the huge security risks brought by adversarial attacks to intelligent recognition models, researchers have carried out works on detecting adversarial examples [24], [25], [26]. In order to protect the intelligent modulation recognition model from attacks, Xu *et al.* realized the detection of adversarial signals by fusing multiple features of radio signals [27]. However, in many application scenarios, it is not enough to detect whether the input is an adversarial example, and it is necessary to use an adversarial defense method to identify the real category of the input. Zhang *et al.* improved the robustness of the AMC model by using homomorphic filtering to attenuate the high-frequency perturbation in the signal [28]. Chen *et al.* used the principle of distillation learning to extract multiple knowledge through adversarial training (AT) and normal training, which improved the robustness of the AMR model to attacks [29]. AT is a simple and effective defense method that uses adversarial knowledge to train a classifier to reduce the vulnerability of the classifier [30], [31]. It has strong robustness to specific attacks, but it is difficult to adapt to new types of attacks that are more aggressive. Therefore, Kim *et al.* proposed a Gaussian smoothing (GS) method, which uses Gaussian noise to enhance the training data and improves the robustness of the modulation classifier to unknown attacks [32]. Although GS weakens the antagonism of the examples, it reduces the recognition accuracy of the classifier to the original examples.

With the rapid development of generative artificial intelligence (GAI), existing work has used autoencoder (AE) and generative adversarial network (GAN) to eliminate adversarial features in input data. AE can learn the representation of data through its encoder and decoder, while GAN can learn the distribution of data through its generator and discriminator, which can achieve more advanced data generation tasks. Sahay *et al.* developed a denoising autoencoder (DAE), which learns the mapping relationship between potential adversarial examples and corresponding original examples, and improves the reliability of power allocation model based on DL [33]. Zhou *et al.* proposed a siamese neuron network based on GAN, which can correctly identify the modulation type of the signal when the intelligent recognition model is attacked [34]. Dong *et al.* constructed a defensive end-to-end communication system based on GAN through triple-training, and used GAN to enhance the robustness of the communication system [35]. Traditional GAN uses random noise as input, which has the problem that the training process is easy to fall into mode collapse [36]. Different from the traditional GAN, Wang *et al.* used the adversarial signal as the input of the generator, which weakened the influence of the adversarial perturbation [37].

Therefore, traditional defense methods face challenges such as weak detection capabilities for subtle perturbations and difficulty in balancing the recognition effects between original clean examples and adversarial examples. In order to accurately detect the adversarial nature of input signals and identify their true modulation categories, we utilize empirical mode decomposition (EMD) and GAN for detection and filtering, respectively, and enhance the model's robustness through decoupled adversarial training. The main contributions of this paper are summarized as follows:

- We propose an EMD-based intelligent adversary detection method, which uses an adversary detector to detect weak adversary perturbations after enhancing the high-frequency features in the examples with EMD.
- We use GAN to regenerate the input signal and randomly shape the reconstructed perturbation, which is beneficial to eliminate the hidden antagonism.
- We propose an adversarial decoupled defense method, which decomposes the traditional adversarial training and uses the adversary detection results to identify the modulation category of the signal to improve the robustness of the model.
- We use the proposed defense method to deal with advanced adversarial attacks, which significantly improves the defense ability of the intelligent recognition model against attacks while ensuring the recognition accuracy of the original examples.

The rest of this paper is organized as follows: Section II introduces the intelligent modulation recognition, adversarial attacks and defense in wireless communication systems. Section III proposes an intelligent adversary detection method based on EMD, which detects the adversary by enhancing the high frequency characteristics of the input signal. Section IV designs an adversarial decoupling defense based on GAN to improve the robustness of the model to attacks while ensuring the original accuracy. Section V shows the effectiveness of the proposed defense method through simulation. Finally, Section VI summarizes the paper.

II. System Model

A. Communication Model

In wireless communication, due to the broadcast nature of the channel, the wireless signal transmitted by the transmitter will be eavesdropped. If the eavesdropper uses the signal and the attack algorithm to attack the intelligent modulation recognition model, it will pose a great threat to the security and reliability of the wireless communication system. Therefore, it is necessary to develop some defense frameworks to resist adversarial attacks. For the intelligent modulation recognition task, the system model of adversarial attack and defense is shown in Fig. 1.

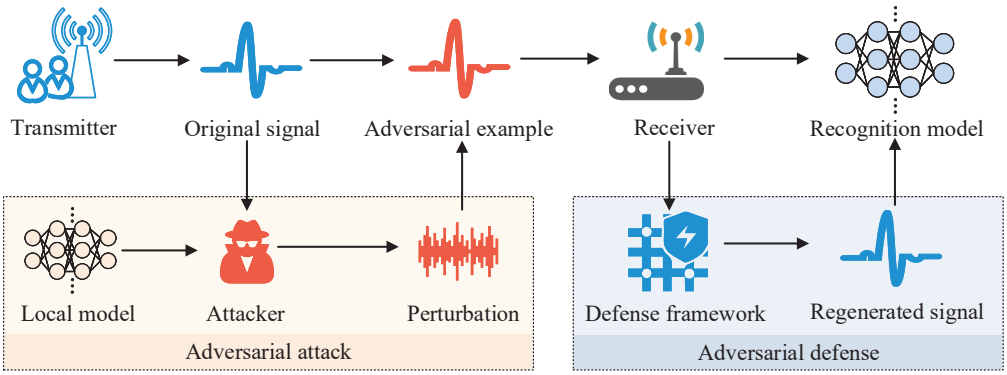


Fig. 1. Adversarial attack and defense system model for intelligent modulation recognition in wireless communication networks.

In Fig. 1, the signal transmitter transmits a wireless signal to the legitimate receiver. In this process, the attacker eavesdrops on the signal and uses its local model and adversarial attack algorithm to generate a perturbation. At the receiving end, the perturbation is superimposed with the original signal to generate an adversarial example, which misleads the receiver's intelligent recognition model to misidentify the modulation. In order to reduce the threat of attack, the adversarial defense strategy can be used to process the signal received by the receiver. If the input example is detected to be an adversarial example, it is discarded or filtered before its modulation mode is identified.

B. Intelligent Modulation Recognition Model

The signal received at the receiver can be expressed as

$$x(t) = s(t) * h(t) + n(t), \tag{1}$$

where $s(t)$ is the modulated signal transmitted by the transmitter, $h(t)$ is the channel response, and $n(t)$ is the additive white Gaussian noise (AWGN).

For IQ modulation, (1) can be expressed as

$$x(t) = I(t) + jQ(t). \tag{2}$$

The in-phase component $I(t)$ and the quadrature component $Q(t)$ are sampled as the input of the recognition model.

Intelligent modulation recognition models usually require excellent classification ability to perform modulation recognition tasks well, and have natural robustness to noise. We use ResNet as the modulation recognition model, which can avoid the disappearance of the gradient and is easier to deepen. It has strong representation ability and is very suitable for complex classification tasks such as modulation recognition [38], [39]. After training, the prediction results of the model for the test examples can be expressed as

$$\arg\max_{k} \{f(y_k|x)\}, \quad k = 1, 2, 3, \cdots, K, \tag{3}$$

where $f(\cdot)$ represents the predicted probability distribution of the model, y_k is the one-hot encoding of the real label, and K is the number of modulation categories.

C. Adversarial Attack Model

In multi-classification tasks such as modulation recognition, the cross-entropy loss function can be used to characterize the difference between the prediction probability of the recognition model and the real label. The cross-entropy loss of the model for the input signal can be expressed as

$$\mathcal{L}(x, y) = -\sum_{k=1}^{K} y_k(x) \log(f_k(x)). \tag{4}$$

In the black box scenario, attackers use attack algorithms such as FGSM, BIM, MIM and DIM to design adversarial perturbations, and use their transferability to mislead the intelligent modulation recognition model. FGSM is a one-step attack, which can quickly attack the model by adding a fixed perturbation size ε to generate adversarial examples in the direction of increasing the model loss [40]. BIM divides ε into multiple parts and iteratively generates perturbations, which can enhance the concealment of the attacks [41]. MIM introduces momentum into BIM, resulting in an adversarial perturbation with stronger attack performance and transferability [42]. DIM improves perturbation size on the basis of MIM, so that it can be adaptively adjusted according to the loss gradient during the iteration process, avoiding the redundancy of the perturbation, and the perturbation generated can be expressed as [20]

$$\eta_{n+1} = \alpha_n \text{sign}\left(\mu g_n + \frac{\nabla_{x_n^*}\mathcal{L}(x_n^*, y)}{\left\|\nabla_{x_n^*}\mathcal{L}(x_n^*, y)\right\|_1}\right), \tag{5}$$

where μ is the momentum decay factor, g_n is the momentum accumulation, x_n^* is the adversarial example generated by η_n after the n-th iteration and $x_n^* = x_{n-1}^* + \eta_n$, $\nabla_{x_n^*}\mathcal{L}$ is the loss gradient of the model, and

$$\alpha_n = \left|2\nabla_{x_n^*}\mathcal{L} - \nabla_{x_{n-1}^*}\mathcal{L}\right| \cdot \left\|2\nabla_{x_n^*}\mathcal{L} - \nabla_{x_{n-1}^*}\mathcal{L}\right\|_1^{-1}. \tag{6}$$

Ensemble attack is a mainstream black-box attack, which generates strong transferable adversarial examples by fusing the output of different networks in the ensemble model. Dong *et al.* pointed out that integrating logit of different networks can effectively improve the black-box attack ability [42].

Therefore, in this paper, we combine FGSM, BIM, MIM, DIM and ensemble attack to describe the attacker's attack behavior.

D. Adversarial Defense Model

1) Adversary Detection: Adversary detection is used to detect adversarial examples and reject them into the model when the recognition model is attacked. Using the original examples and adversarial examples to directly train and test the neural network is a simple binary classification problem, which can detect the adversarial of the examples to a certain extent. For example, Aigrain *et al.* constructed an anomaly detector Inspection-Net, using the internal representation of the model to detect adversarial examples [43]. In general, adversary detection needs to obtain the features of the input through the detection algorithm, and compare them with the decision threshold to determine whether it is an adversarial example, which can be expressed as

$$\mathcal{B}(x) \underset{ori}{\overset{adv}{\gtrless}} \tau, \tag{7}$$

where $\mathcal{B}(x)$ is the output of the binary adversary detector, and τ is the decision threshold.

2) Adversarial Defense: In general, the input is discarded when it is detected as an adversarial example, which causes the loss of communication information. Therefore, it is necessary to further study adversarial defense methods such as adversarial training (AT), Gaussian smoothing (GS) and GAN, to restore their original modulation information by learning or filtering adversarial perturbations. AT makes the model robust to attack by adding adversarial examples to the training process of the model [44]. The process of AT can be expressed as

$$\min_{\theta} \max_{\|x^*-x\|_p \le \varepsilon} \mathbb{E}_{(x,y)} \left[\mathcal{L}\left(f\left(x^*,\theta\right),y\right)\right], \tag{8}$$

where $\mathbb{E}$ represents the expectation, θ is the model parameter and is continuously updated during the training process. The internal maximization problem in (8) is used to find the perturbation ε that maximizes the loss of the model under the p-norm constraint, and the external minimization problem is used to update θ to make the model robust to the perturbation.

GS uses Gaussian noise to enhance the training set of the model and improve the robustness of the modulation recognition model to perturbations that may exist in multiple directions [32]. By adjusting the standard deviation of the noise and the number of noise samples added, an example in the enhanced training set can be expressed as

$$\bar{x}_i = \left\{x_i + n_1, x_i + n_2, \cdots, x_i + n_s\right\}, \tag{9}$$

where n represents the Gaussian noise with mean zero and standard deviation σ.

GAN can weaken the adversarial perturbation contained in the signal by reconstructing the signal [37]. GAN consists of generator G and discriminator D. G generates an output $G(x^*)$ similar to the real data distribution through the feature mapping of the adversarial example x^*, and D is used to distinguish the original example x and the generated example $G(x^*)$.During the training process, G and D constantly complement and optimize each other as

$$\begin{aligned}\min_{G}\max_{D} \quad & \mathbb{E}_{x\sim p_{data}(x)}\left[\log\left(D\left(x\right)\right)\right] \\ & + \mathbb{E}_{x^*\sim p_{data}(x^*)}\left[\log\left(1-D\left(G\left(x^*\right)\right)\right)\right].\end{aligned} \tag{10}$$

In (10), the first term represents the probability that the real example is judged as real data, and the second term represents the probability that the generated $G(x^*)$ is judged as false data. Finally, G generates the example that is close to the distribution of the original example, and corrects the offset of the data distribution caused by the perturbation in x^*.

III. EMD Based Intelligent Adversary Detection

Adversary detection is used to detect whether the input is adversarial. Due to the high-dimensional characteristics of DNN, the small differences in the input can be amplified in the process of propagation between network layers, and even mislead the output results of the network. This is the reason why the adversarial examples are aggressive. Therefore, the output of the last feature extraction layer of the intelligent recognition model can be used to distinguish the original example and the adversarial example. Using the logical output of the recognition model to train a simple binary detector, good detection results can be achieved without changing the structure of the model [43]. The output of the detector is usually a probability indicating the suspiciousness of the example, which can be regarded as the threat level of the example to the intelligent recognition model. However, when the adversarial perturbation power is very small, the detection of this method is not ideal. In order to solve this problem, this paper proposes an empirical mode decomposition based adversary detection (EMD-AD) method to improve the accuracy of detection for low perturbation. The process of the proposed method is shown in Fig. 2.

Compared with the original examples, adversarial examples tend to have higher power at high frequencies [45], [46]. Therefore, the high frequency characteristics can be used to detect the suspiciousness of examples. EMD is an adaptive time-frequency signal processing method, which decomposes the signal according to the time scale characteristics of the data itself, without pre-setting any basis functions. It can separate the signal into intrinsic mode function (IMF) and residual function with different frequencies, which is especially suitable for the analysis and processing of nonlinear and non-stationary signals.

Firstly, the upper envelope $e_{\max}(t)$ and the lower envelope $e_{\min}(t)$ are obtained by connecting the local maximum point and the local minimum point of the signal respectively through the cubic spline curve, and the average value of the two envelopes is calculated as

$$m_1(t) = \frac{e_{\max}(t) + e_{\min}(t)}{2}. \tag{11}$$

Then, using the difference between the original signal and (11), the intermediate signal can be expressed as

$$C_{1,1}(t) = x(t) - m_1(t). \tag{12}$$

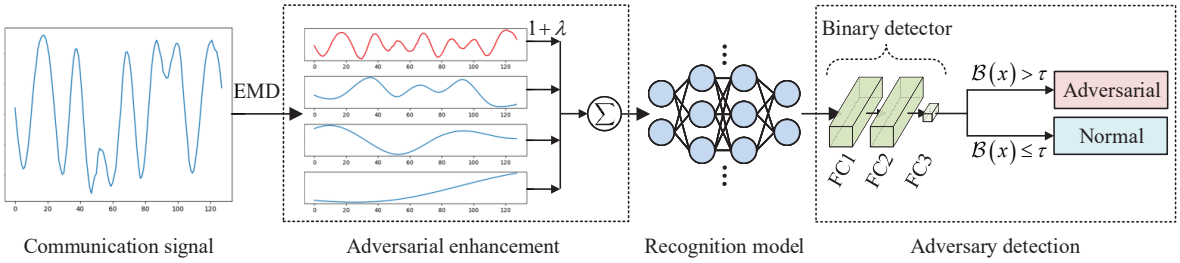


Fig. 2. Illustration of adversary detection process based on empirical mode decomposition.

If the number of extreme points of $C_{1,1}(t)$ is equal to or at most one different from the number of zero-crossing points, and the upper and lower envelopes at any time are locally symmetric with respect to the time axis, then $IMF_1 = C_{1,1}(t)$ and the residual component $r_1(t) = x(t) - IMF_1$. Otherwise, continue to decompose the signal. After several iterations, when the residual component $r_n(t) = r_{n-1}(t) - IMF_n$ is monotonous, the decomposition process ends and the residual component is $R(t) = r_n(t)$. At this time, the signal is decomposed into multiple IMF components and a residual component, which can be expressed as

$$x(t) = \sum_{i=1}^{n} IMF_i + R(t). \tag{13}$$

In (13), IMF_1 has the highest frequency. Therefore, the main information of adversarial perturbation is hidden in IMF_1. When the perturbation is large, the power in the IMF_1 of the adversarial example is significantly larger than that of the original example, which can easily be detected. However, when the perturbation is small, the gap is small, making it difficult for the neural network to learn adversarial features. Therefore, this paper enhances the antagonism hidden in the example by increasing the power of IMF_1, so as to provide better feature differences for the network to facilitate detection. After the high frequency characteristic is enhanced, the signal can be expressed as

$$\hat{x}(t) = \lambda \cdot IMF_1 + \sum_{i=1}^{n} IMF_i + R(t), \tag{14}$$

where λ denotes the high-frequency feature enhancement factor and $\lambda \geq 0$.

After the high-frequency feature enhancement, the example is used in the recognition model to obtain its logits output $f_l(\hat{x})$, which is input into the binary detector to obtain the prediction probability $\mathcal{B}(f_l(\hat{x}))$. Then, it is decided whether the example is adversarial according to (7). The detector used in this paper has simple structure and consists of only three fully connected layers, and dropout is used to prevent overfitting, as shown in Fig. 2.

Therefore, the proposed EMD-AD first enhances the high-frequency feature in the signal based on EMD, highlighting the weak adversarial information hidden in the signal. Then, the enhanced signal is input to the recognition model to obtain its logits output. Finally, the adversary detector is used to evaluate the suspiciousness of the signal and determine whether it is an adversarial example.

IV. GAN-Based Adversarial Decoupled Defense

The adversary detector only detects if the received signal is an adversarial example and discards it if it is. However, in many cases, when the detector finds that the example is adversarial, it is also necessary to accurately identify the modulation mode of the example. Therefore, in this section, we propose an adversarial decoupled defense (ADD) method based on GAN to weaken the adversarial perturbation of coupling in the example and decouple the traditional adversarial training process.

A. GAN-based Example Regeneration

The structure of GAN used in this paper is shown in Table I. To train the generator G, the original example and its adversarial example are inputs. During the training process, the waveform reconstructed by G not only needs to approximate the original signal waveform in the time domain, but also needs to be able to deceive the discriminator D to make an error prediction. Therefore, the loss function of G can be defined as

$$\begin{aligned}\mathcal{L}_G(x, x^*) =& \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left\|x_{i,j} - G_j\left(x_i^*\right)\right\|_2^2 \\ &+ \frac{\beta}{M}\sum_{i=1}^{M}\log\left(1 - D\left(G\left(x_i^*\right)\right)\right),\end{aligned} \tag{15}$$

where M is the number of examples, N is the length of examples, and β is the discriminant loss coefficient.

To train discriminator D, it is necessary to increase the prediction probability of the discriminator for the original examples while reducing the prediction probability of the discriminator for the reconstructed examples. The loss of D can be defined as

$$\begin{aligned}\mathcal{L}_D(x, x^*) &= \frac{1}{M}\sum_{i=1}^{M}\left(\log\left(D\left(G\left(x_i^*\right)\right)\right) + \log\left(1 - D\left(x_i\right)\right)\right) \\ &= \frac{1}{M}\sum_{i=1}^{M}\log\left(D\left(G\left(x_i^*\right)\right)\left(1 - D\left(x_i\right)\right)\right).\end{aligned} \tag{16}$$

The training process continuously updates the network parameters by minimizing $\mathcal{L}_G(x, x^*)$ and $\mathcal{L}_D(x, x^*)$, and

TABLE I
NETWORK STRUCTURE OF GAN

Generator			Discriminator		
Layer	**Kernel**	**Output shape**	**Layer**	**Kernel**	**Output shape**
Input	(batch,128,2)	(batch,128,2)	Input	(batch,128,2)	(batch,128,2)
Conv1D	(64,3)	(batch, 128, 64)	Conv1D	(4,5)	(batch, 128, 4)
MaxPooling1D	2	(batch, 64, 64)	MaxPooling1D	2	(batch, 64, 4)
Conv1D	(64,3)	(batch, 64, 64)	BN+Relu	-	(batch, 64, 4)
MaxPooling1D	2	(batch, 32, 64)	Conv1D	(4,3)	(batch, 64, 4)
Conv1D	(64,3)	(batch, 32, 64)	MaxPooling1D	2	(batch, 32, 4)
UpSampling1D	2	(batch, 64, 64)	BN+Relu	-	(batch, 32, 4)
Conv1D	(64,3)	(batch, 64, 64)	Flatten+Dropout	0.5	(batch, 128)
UpSampling1D	2	(batch, 128, 64)	BN+Relu	-	(batch, 128)
Conv1D	(2,3)	(batch, 128, 2)	Dense	1	(batch, 1)

enhances the reconstruction ability of G and the discrimination ability of D. After the training, the test example is fed into G, and G reconstructs a modulated signal waveform that is similar to the original signal and can mislead D to identify it as true.

During the test, the test example x_t may be either original or adversarial. The difference of examples before and after GAN is recorded as reconstruction perturbation

$$\Delta x_t = x_t - G\left(x_t\right). \tag{17}$$

We replace Δx_t with a random noise Δn with the same power, and add it to the reconstructed example $G\left(x_t\right)$ to obtain the regenerated example

$$x_r = G\left(x_t\right) + \Delta n, \tag{18}$$

and

$$\Delta n = \phi \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\Delta x_{t,i}\right)^2}, \tag{19}$$

where ϕ denotes a random signal following a standard normal distribution and has a length of N.

When the input is an adversarial example, Δx_t contains adversarial information. If it is replaced with random noise, this part of adversarial information will be destroyed. When the input is the original example, Δx_t contains only random noise, and replacing it with random noise has little effect on signal and model prediction. Compared with the traditional method of adding random noise directly to the input example, the proposed method can adaptively adjust the noise power according to the size of the reconstructed perturbation, and avoid the influence of unreasonable noise power setting on the modulation characteristics of the example.

B. Adversarial Training Decoupling

After regenerating the example, the adversarial information contained in the adversarial example is greatly reduced. The traditional AT adds adversarial examples to the training set to train the recognition model with the original examples during the training process, which can lead to overfitting and reduce the recognition accuracy of the model for the original examples. To solve this problem, we decouple the adversarial training into the original branch and the adversarial branch. The original branch only uses the original examples to train the recognition network for identifying the original examples. The adversarial branch only uses the adversarial examples to train the network with the same structure as the original branch, which is used to identify unknown adversarial examples. Compared with the untreated adversarial examples, the signal features in the examples regenerated by GAN will be more obvious. Therefore, we use the regenerated examples to train the adversarial network.

In this paper, we use the projection gradient descent method (PGD) to generate adversarial examples to train the adversarial network. PGD randomly initializes examples before iteration, which enhances the antagonism of examples in multiple directions. It can be expressed as

$$x_{n+1} = \prod_{x+S}\left(x_n^* + \alpha \text{sign}\left(\nabla_{x_n^*}\mathcal{L}\left(f\left(x_n^*\right), y\right)\right)\right), \tag{20}$$

where S denotes the introduced random perturbation. In order to further improve the generalization of the adversarial branch for unknown attacks, we use the idea of Gaussian smoothing in [32] to enhance the data of adversarial examples.

Before defense, we use the adversary detector to detect the suspiciousness of the signal after enhancing IMF_1, and then input the suspiciousness and the signal before feature enhancement into the defense model. Since the suspiciousness of the adversary detector indicates the probability that the input is an adversarial example, it can be used as the fusion coefficient of the original branch and the adversarial branch. Then the final output of the modulation recognition model is

$$f\left(x\right) = \left(1-\mu\right) f_{ori}\left(x\right) + \mu f_{adv}\left(x_r\right), \tag{21}$$

where f_{ori} and f_{adv} represent the original network and the adversarial network respectively, and $\mu = \mathcal{B}\left(f_l\left(\hat{x}\right)\right)$ is the suspiciousness of the example. If $\mu = 0$, it indicates that the input example is not threatened and can be completely treated as the original example. The recognition accuracy of the model is completely consistent with that of the original

Algorithm 1 Adversarial Decoupled Defense

Input: Test example set $\mathcal{X}$ and label set $\mathcal{Y}$; generator G; adversary detector $\mathcal{B}$; original recognition network f_{ori} and adversarial recognition network f_{adv};

Input: High frequency feature enhancement factor λ.

Output: modulation mode set $\mathcal{M}$.

1: **for** $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$ **do**

2: Use EMD to enhance the high-frequency feature of x_i. The enhanced signal

$$\hat{x}_i = \lambda \cdot IMF_1 + \sum_{i=1}^{n} IMF_i + R;$$

3: Use $\mathcal{B}$ to detect suspiciousness

$$\mu = \mathcal{B}\left(f_{ori,l}\left(\hat{x}_i\right)\right);$$

4: Based on G, regenerate the example

$$x_r = G\left(x_i\right) + \Delta n;$$

5: Fuse the outputs of networks to obtain the prediction

$$f\left(x_i\right) = \left(1-\mu\right) f_{ori}\left(x_i\right) + \mu f_{adv}\left(x_r\right);$$

6: Output the modulation mode

$$\mathcal{M}_i = \arg\max_k \left\{f\left(y_{i,k}|x_i\right)\right\}, \quad k = 1, 2, 3, \cdots, K;$$

7: **end for**

8: **return** $\mathcal{M}$.

recognition model, which avoids the decline of the recognition accuracy of the original example. If $0 < \mu < 1$, it shows that the detector can not fully determine the nature of the input example. At this time, it is necessary to combine the results of the original branch and the adversarial branch to identify. If $\mu = 1$, it shows that the input example is adversarial, and the recognition result of the model is the same as that of the adversarial branch.

In this section, we use GAN to regenerate the input signal, which greatly weakens the antagonism in the example, and fuses the decoupled original branch and the adversarial branch according to the suspiciousness to accurately identify the modulation mode of the example. The steps of the proposed ADD are shown in Algorithm 1.

V. Simulation Results and Discussion

In this section, we verify the effectiveness of the proposed defense algorithm against black-box attacks through simulation. According to the purpose of the attack, we assume that the attacker only focuses on the original examples that are recognized correctly by the intelligent recognition model and generates corresponding adversarial examples to mislead the model, and test the recognition accuracy of the model to these examples after defense. When training the intelligent recognition model, we set the batch and epoch to 1024 and 100, respectively, and set the initial value of the learning rate to 0.001 and automatically update it.

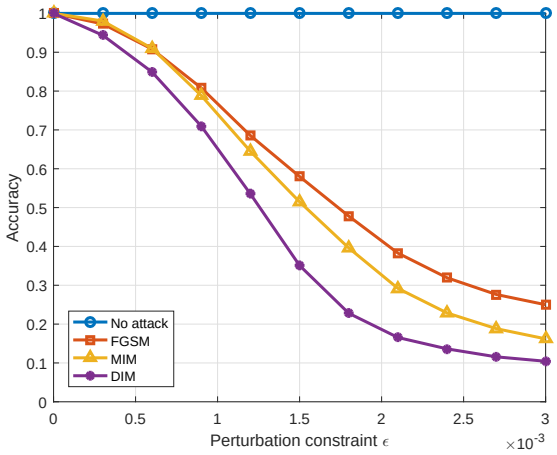


Fig. 3. Recognition accuracy of the model after being attacked.

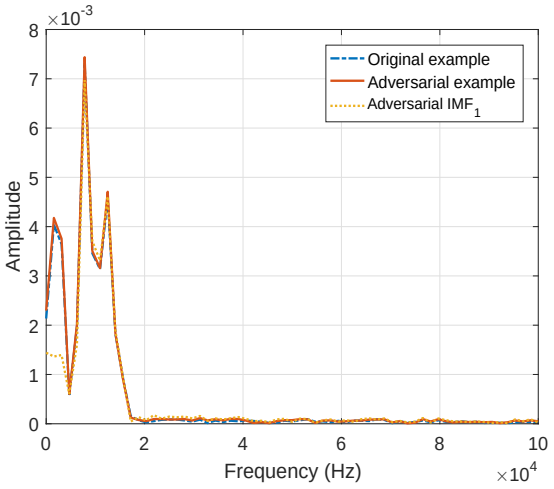
In this paper, we use the modulation signal data set RML2016.10b to verify the threat of the attacker to the AMC model and the performance of the defense method [47]. The data set contains eight digital signals generated in simulated harsh propagation environments: 8 phase shift keying (8PSK), quadrature phase shift keying (QPSK), binary phase shift keying (BPSK), Gaussian frequency shift keying (GFSK), continuous phase frequency shift keying (CPFSK), pulse amplitude modulation 4 (PAM4), quadrature amplitude modulation 16 (QAM16) and quadrature amplitude modulation 64 (QAM64) and two analog signals: double sideband amplitude modulation (AM-DSB) and wide band frequency modulation (WBFM). These signals are affected by AWGN, multipath fading, sampling rate offset and center frequency offset. Each signal sample consists of an in-phase component and an quadrature component. We randomly select 80% of the examples for training the recognition model, and the rest for testing.

A. Adversarial Attacks

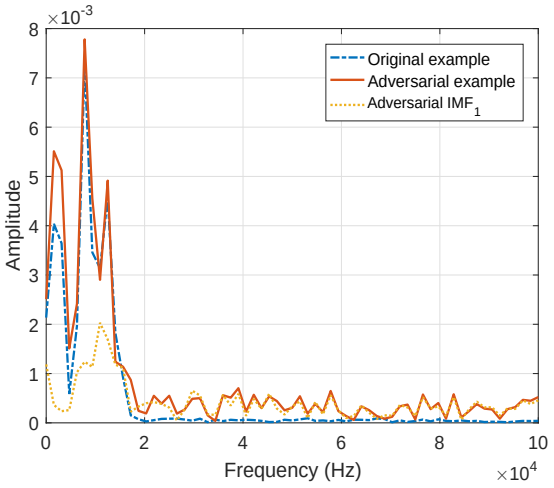
Before testing the effect of the defense algorithm, we first simulate the attacker's attack behavior and test the performance of the ResNet recognition model after being attacked by FGSM, MIM and DIM. When attacking, in order to ensure the concealment of the attack, the attacker usually uses the infinite norm to constrain the power of the perturbation. The perturbation-to-noise ratio (PNR) is often used to measure the invisibility of an attack, defined as [48]

$$\mathrm{PNR}\left[\mathrm{dB}\right] = \frac{\mathbb{E}\left[\left\|\varepsilon\right\|_2^2\right]}{\mathbb{E}\left[\left\|x\right\|_2^2\right]}\left[\mathrm{dB}\right] + \mathrm{SNR}\left[\mathrm{dB}\right]. \tag{22}$$

In this paper, we study the recognition performance of the model at SNR $=$ 10 dB. PNR is the ratio of perturbation power to noise power, which is used to measure the level of perturbation relative to background noise. When PNR $<$ 0 dB, it means that the perturbation power is lower than the noise power, the concealment of the perturbation is strong and the attack trace is difficult to be detected. At this time, the



(a) $\varepsilon = 0.0003$



(b) $\varepsilon = 0.0030$

Fig. 4. Spectrum of BPSK original signal and its adversarial example.

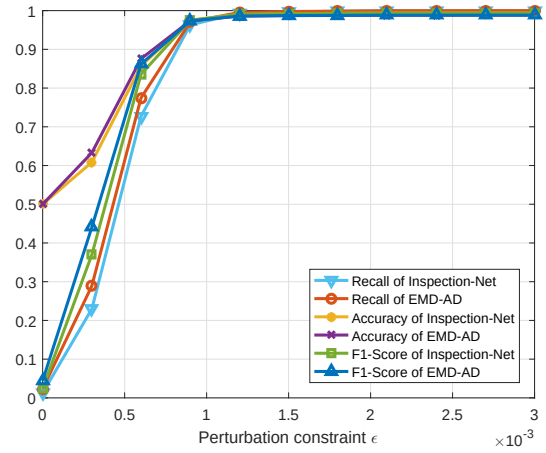


Fig. 5. Detection performance of the adversary detector.

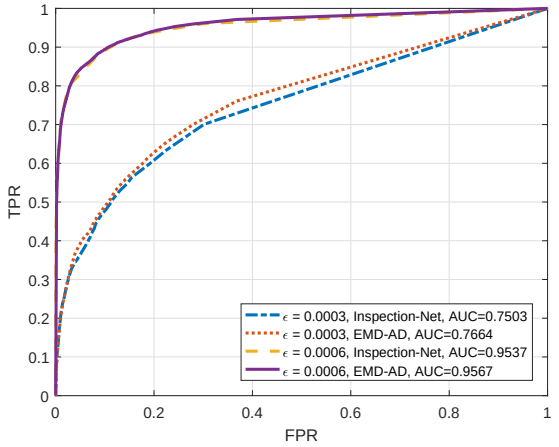


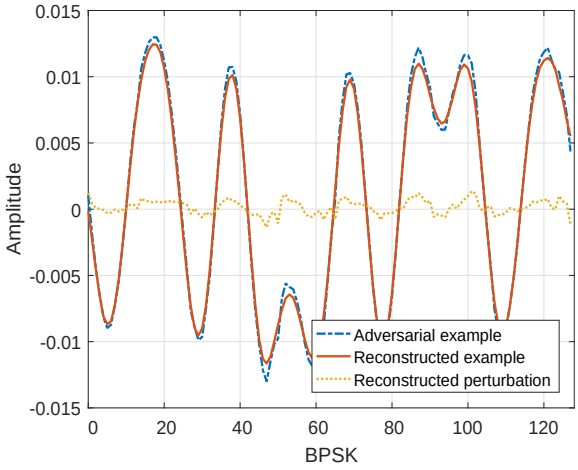
Fig. 6. ROC curves of adversary detectors.

perturbation constraint $\varepsilon < 0.00316$ can be obtained according to (22). Therefore, we choose the perturbation constraint with interval of 0.0003 in [0, 0.0030] to generate adversarial examples. Under different perturbation constraints, three attack algorithms are used to generate adversarial examples in the ensemble model composed of VTCNN, Inception and VGG. The recognition accuracy of the recognition model for these examples is tested, as shown in Fig. 3. It can be seen that the accuracy of the recognition decreases significantly after being attacked, and as the perturbation constraint increases, the accuracy decreases more. Among the attacks tested, DIM has the best attack effect. For example, when $\varepsilon = 0.0015$, the accuracy of the recognition model is reduced by 64.91%, which seriously damages the reliability of the intelligent modulation recognition model.

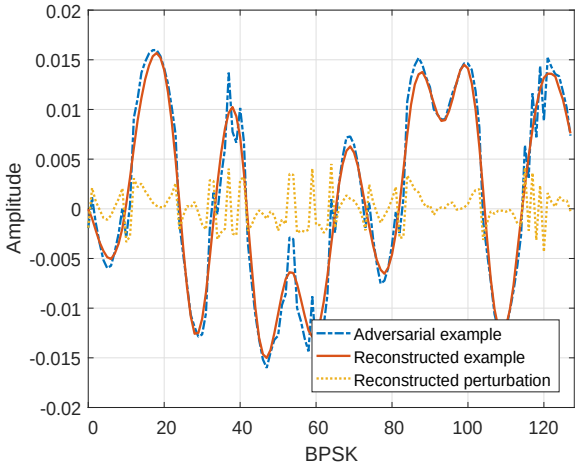
B. Adversarial Example Detection

In this section, we use the most adversarial DIM as the attack method to test the detection performance of Inspection-Net in [43] and the proposed EMD-AD. We first study the contribution of IMF_1 to the adversarial examples under small and large perturbations in the frequency domain to verify the effect of high-frequency feature enhancement on adversarial example detection, as shown in Fig. 4. It can be seen that when the perturbation is small, the characteristics of the clean examples are mainly concentrated in the low frequency, and the spectral curves of the original examples and the adversarial examples are basically coincident, which is not conducive to the detection of adversarial examples. When the perturbation is large, the difference between the two is mainly reflected in the high frequency, which is also the place where the adversarial example plays a role, and the spectrum curve here basically coincides with IMF_1. Therefore, enhancing the high-frequency features of examples based on EMD is helpful to detect the existence of adversarial perturbations.

When detecting, we regard the original example as 'negative' and mark it as 0, and regard the adversarial example as 'positive' and mark it as 1. In order to quantitatively measure the overall detection performance of the detector for the original examples and the adversarial examples, we use Accuracy, Recall, Precision and F1-Score as metrics, which

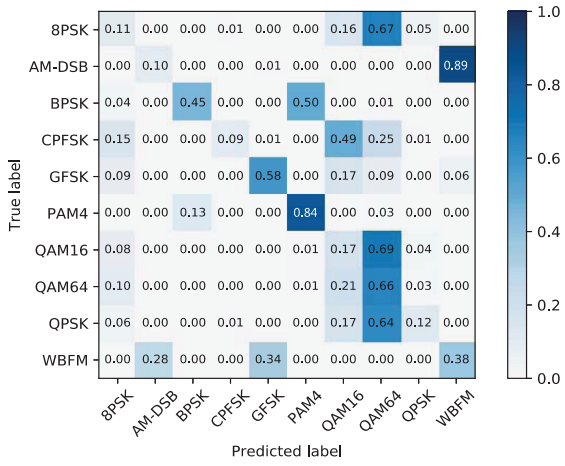


(a) Original Example

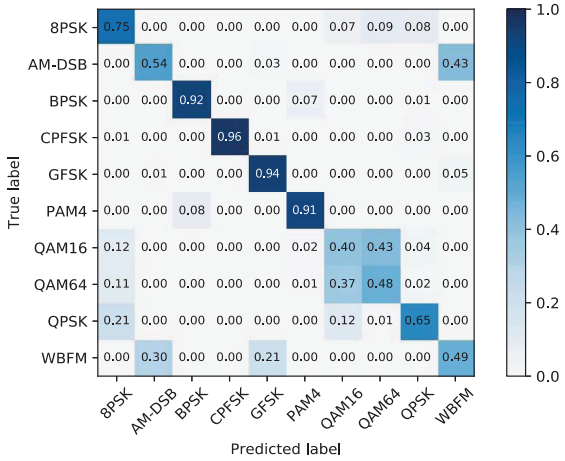


(b) DIM Adversarial Example

Fig. 7. Time domain waveforms of the original example and the adversarial example before and after reconstruction.



(a) DIM Attack



(b) ADD Defense

Fig. 8. Confusion matrix of intelligent modulation recognition model.

are expressed as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{23}$$

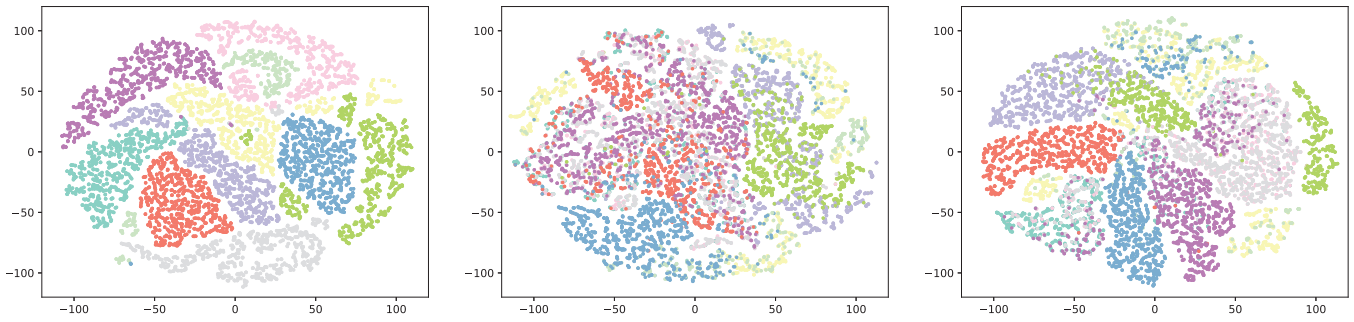
$$Recall = TP/\left(TP + FN\right), \tag{24}$$

$$Precision = TP/\left(TP + FP\right), \tag{25}$$

$$F1\text{-}Score = \frac{2\left(Precision \times Recall\right)}{Precision + Recall}, \tag{26}$$

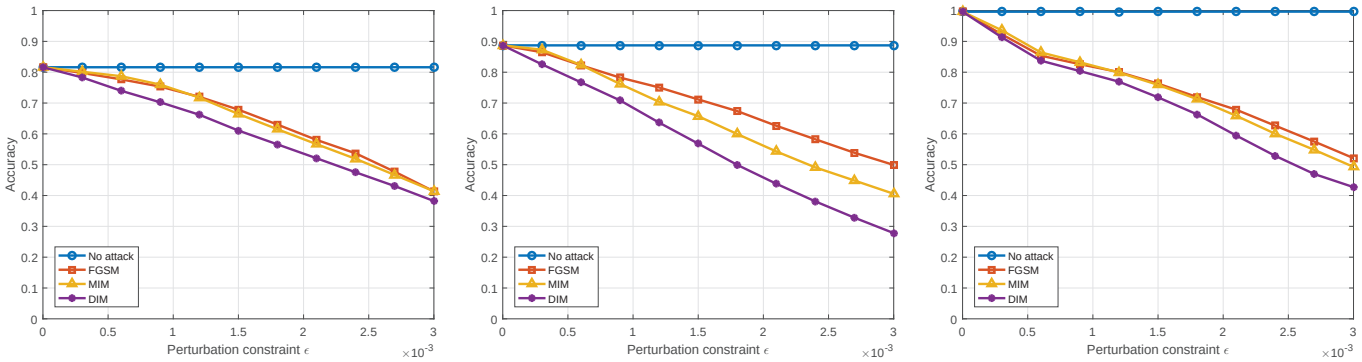
where TP represents the count of adversarial examples that are correctly predicted as adversarial examples, FN represents the count of adversarial examples that are incorrectly predicted as original examples, TN represents the count of original examples that are correctly predicted as original examples, and FP represents the count of original examples that are incorrectly predicted as adversarial examples. The detection results of the adversary detector for adversarial examples under different perturbation constraints are shown in Fig. 5. It can be seen that the overall detection effect of the proposed EMD-AD is better than that of Inspection-Net, especially when the adversarial perturbation is weak. For example, at $\varepsilon = 0.0003$, the Accuracy, Recall and F1-Score of EMD-AD are 2.45%, 5.96% and 7.10% higher than those of Inspection-Net, respectively. This is because when the perturbation is very weak, the adversarial information hidden in the example is difficult to be detected directly by the detector. Since EMD-AD detects the example after enhancing the high-frequency features hidden in the example, it has a better detection effect than Inspection-Net. With the increase of perturbation, the adversarial features in the example are gradually obvious, so that Inspection-Net can accurately detect the adversarial example. As the perturbation continues to increase, the detection accuracy of EMD-AD and Inspection-Net is basically the same, both approaching 100%.

In order to further test the detection ability of the adversary detector under weak perturbations, we generate DIM adversarial examples when the perturbation constraint $\varepsilon = 0.0003$ and $\varepsilon = 0.0006$, and input the original examples and adversarial examples into the detector. We calculate true positive rate



(a) Original examples (b) Adversarial examples (c) Regenerated examples

Fig. 9. Visualization of the features of the examples in the recognition model.



(a) AT Defense (b) GS Defense (c) ADD Defense

Fig. 10. Effect of different defenses on different attacks.

(TPR) and false positive rate (FPR) by

$$TPR = TP/\left(TP + FN\right), \tag{27}$$

$$FPR = FP/\left(TN + FP\right). \tag{28}$$

In the test, for a given threshold τ, when the prediction probability of the detector is greater than the threshold, the example is predicted to be positive. We set 100000 uniformly distributed thresholds τ within [0,1], and the detector will get a TPR and FPR according to each threshold. Then, we draw the Receiver Operating Characteristics (ROC) curve with (FPR, TPR), as shown in Fig. 6, and define the area under the ROC curve as Area Under Curve (AUC). The larger the AUC is, the closer the ROC is to (0,1), the greater the TPR is than the FPR, and the better the overall performance of the detector is. It can be seen from Fig. 6 that under the two weak perturbations, the AUC values of the proposed EMD-AD are higher than those of Inspection-Net, and its detection performance is better.

Time complexity is the growth trend of the running time of the algorithm as the amount of data becomes larger, which can reflect the efficiency of the algorithm. Inspection-Net uses a trained adversary detector to detect examples. Its time complexity is closely related to the test data volume M and the network parameter scale F including the number of network layers and the number of neurons in each layer, which can be expressed as $\mathcal{O}(MF)$. Since the proposed EMD-AD needs to use EMD to enhance signal features before detection, its time complexity is $\mathcal{O}(M) + \mathcal{O}(MF) = \mathcal{O}(MF)$, which is consistent with Inspection-Net. In addition, we record the average detection times for the two methods as 0.0023s and 0.0015s, respectively. Since EMD-AD requires high-frequency feature enhancement for the examples, its detection time is slightly longer than that of Inspection-Net. However, it can more accurately detect the hidden perturbations in the examples, providing crucial adversarial information for subsequent defense.

C. Adversarial Defense

In order to observe the generation ability of generator G in GAN, we draw the time-domain waveforms of a BPSK signal and its DIM adversarial example before and after passing through G, as shown in Fig. 7. It can be seen that when the input is the original example, the example generated by G is very close to the input, and has little effect on the original example. When the input is an adversarial example, the generated example is quite different from the input, and the reconstructed perturbation contains a lot of adversarial information. It can be seen that the trained generator can better regenerate the input example.

In order to intuitively show the defense effect of different modulation modes, we use the confusion matrix to display the recognition results of the model before and after the

defense, and the value on the diagonal represents the prediction accuracy of each modulation. We use the most adversarial DIM to generate adversarial examples under the constraint of $\varepsilon = 0.0015$, and use ADD for defense. The results are shown in Fig. 8. It can be seen that the prediction results of the model after being attacked are very confused, and its reliability drops sharply. After ADD defense, the model can correctly classify most of these adversarial examples, protecting the intelligent recognition model from attack.

Then, we verify the impact of GAN-based example regeneration on the characteristics of adversarial examples. We use t-SNE to visualize the features of the original examples, DIM adversarial examples and regenerated examples in the recognition model, as shown in Fig. 9. It can be seen that the feature distribution of adversarial examples in the model is very chaotic, making it difficult for the model to correctly identify its true category. After regeneration, because some of the adversarial perturbations in the examples are replaced by random noise, the influence of the attack is weakened, and the features become clear.

Finally, we test the defense performance of ADD, AT and GS against attacks. AT uses PGD adversarial examples to expand the training set, and GS uses noise standard deviation $\sigma = 0.003$ and example number $s = 10$. The defense effect is shown in Fig. 10. It can be seen that compared with the accuracy in Fig. 3, the three defense methods can improve the accuracy of the model after the attack. Among them, the defense effect of ADD under different perturbation constraints is better than the other two methods. For example, when $\varepsilon = 0.0015$, ADD improves the accuracy by 10.78% and 15.02% compared with AT and GS, respectively. At the same time, it is worth noting that the recognition accuracy of ADD for clean examples is as high as 99.63%, while that of AT and GS is only 81.61% and 88.71%, respectively. This shows that ADD can improve the robustness of the model to attacks while maintaining the recognition accuracy of the original examples.

VI. Conclusion

This paper has studied the security problem of intelligent modulation recognition model in wireless communication system. In response to the risk of intelligent modulation recognition models being vulnerable to black-box attacks, we combined the powerful data generation capabilities of GAI to design an adversarial decoupled defense method, which effectively enhanced the robustness of the recognition model. Firstly, we designed an adversary detector and improved the detection of subtle adversarial effects based on EMD. Then, we regenerated examples that approximate the true distribution using GAN, weakening the adversarial perturbations in the input signals. Finally, we decoupled the adversarial training into an original branch and an adversarial branch, and fused the outputs of the two branches using the adversary detection results to obtain the modulation type of the signal. The simulation results show that the proposed defense method significantly enhances the robustness of the attack while ensuring the recognition accuracy of the model for original examples, and ensures the safety and reliability of the modulation recognition model in the intelligent communication network.

References

[1] X. Wang, X. Ren, C. Qiu, Y. Cao, T. Taleb and V. C. M. Leung, "Net-in-AI: A computing-power networking framework with adaptability, flexibility, and profitability for ubiquitous AI," *IEEE Netw.*, vol. 35, no. 1, pp. 280-288, Jan./Feb. 2021.

[2] K. B. Letaief, Y. Shi, J. Lu and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5-36, Jan. 2022.

[3] C. Huang *et al.*, "Artificial intelligence enabled radio propagation for communications—part I: Channel characterization and antenna-channel optimization," *IEEE Trans. Antennas Propag.*, vol. 70, no. 6, pp. 3939-3954, June 2022.

[4] X. Wang, Y. Zhao, C. Qiu, Q. Hu and V. C. M. Leung, "Socialized learning: A survey of the paradigm shift for edge intelligence in networked systems," *IEEE Commun. Surv. Tutorials*, Oct. 2024, DOI: 10.1109/COMST.2024.3482978.

[5] X. Wang, Y. Zhao, C. Qiu, Z. Liu, J. Nie and V. C. M. Leung, "InFEDge: A blockchain-based incentive mechanism in hierarchical federated learning for end-edge-cloud communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 12, pp. 3325-3342, Dec. 2022.

[6] H. Zha *et al.*, "LT-SEI: Long-tailed specific emitter identification based on decoupled representation learning in low-resource scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 929-943, Jan. 2024.

[7] G. Han, Z. Xu, H. Zhu, Y. Ge and J. Peng, "A two-stage model based on a complex-valued separate residual network for cross-domain IIoT devices identification," *IEEE Trans. Ind. Inf.*, vol. 20, no. 2, pp. 2589-2599, Feb. 2024.

[8] L. Zhang, S. Zheng, K. Qiu, C. Lou and X. Yang, "MASSnet: Deep-learning-based multiple-antenna spectrum sensing for cognitive-radio-enabled internet of things," *IEEE Internet Things J.*, vol. 11, no. 8, pp. 14435-14448, Apr. 2024.

[9] N. P. Shankar, D. Sadhukhan, N. Nayak, T. Tholeti and S. Kalyani, "Binarized ResNet: Enabling robust automatic modulation classification at the resource-constrained edge," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 5, pp. 1913-1927, Oct. 2024.

[10] G. Baldini, F. Bonavitacola and J. -M. Chareau, "Wireless interference identification with convolutional neural networks based on the FPGA implementation of the LTE cell-specific reference signal (CRS)," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 1, pp. 48-63, Feb. 2024.

[11] F. A. Bhatti, M. J. Khan, A. Selim and F. Paisana, "Shared spectrum monitoring using deep learning," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 4, pp. 1171-1185, Dec. 2021.

[12] Q. Xuan *et al.*, "AvgNet: Adaptive visibility graph neural network and its application in modulation classification," *IEEE Trans. Network Sci. Eng.*, vol. 9, no. 3, pp. 1516-1526, 1 May-June 2022.

[13] S. Aer, Z. Wang, K. Wang, X. Zhang and H. Gao, "A super-resolution data processing for automatic modulation classification based on tree compression networks," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1-13, June 2023.

[14] P. Qi, X. Zhou, S. Zheng and Z. Li, "Automatic modulation classification based on deep residual networks with multimodal information," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 1, pp. 21-33, Mar. 2021.

[15] R. Ding, F. Zhou, Q. Wu, C. Dong, Z. Han and O. A. Dobre, "Data and knowledge dual-driven automatic modulation classification for 6G wireless communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 5, pp. 4228-4242, May 2024.

[16] 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, June 2024.

[17] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2014, pp. 1-10.

[18] J. Kang *et al.*, "Adversarial attacks and defenses for semantic communication in vehicular metaverses," *IEEE Wireless Commun.*, vol. 30, no. 4, pp. 48-55, Aug. 2023.

[19] 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.

[20] M. Liu, Z. Zhang, N. Zhao and Y. Chen, "Adversarial attacks on deep neural networks based modulation recognition," in *Proc. IEEE Conf. Comput. Commun. Workshops*, May 2022, pp. 1-6.

[21] D. Ke, X. Wang and Z. Huang, "Frequency-selective adversarial attack against deep learning-based wireless signal classifiers," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 4001-4011, Jan. 2024.

[22] C. Hu, H. -Q. Xu and X. -J. Wu, "Substitute meta-learning for black-box adversarial attack," *IEEE Signal Process Lett.*, vol. 29, pp. 2472-2476, Dec. 2022.
[23] J. Dong, Y. Wang, J. Lai and X. Xie, "Restricted black-box adversarial attack against deepfake face swapping," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 2596-2608, Apr. 2023.
[24] Y. Qing, T. Bai, Z. Liu, P. Moulin and B. Wen, "Detection of adversarial attacks via disentangling natural images and perturbations," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 2814-2825, Jan. 2024.
[25] B. Zhu, C. Dong, Y. Zhang, Y. Mao and S. Zhong, "Toward universal detection of adversarial examples via pseudorandom classifiers," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 1810-1825, Dec. 2024.
[26] B. Liang, H. Li, M. Su, X. Li, W. Shi and X. Wang, "Detecting adversarial image examples in deep neural networks with adaptive noise reduction," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 1, pp. 72-85, 1 Jan.-Feb. 2021.
[27] D. Xu, H. Yang, C. Gu, Z. Chen, Q. Xuan and X. Yang, "Adversarial examples detection of radio signals based on multifeature fusion," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 68, no. 12, pp. 3607-3611, Dec. 2021.
[28] S. Zhang *et al.*, "HFAD: Homomorphic filtering adversarial defense against adversarial attacks in automatic modulation classification," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 3, pp. 880-892, June 2024.
[29] 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.
[30] K. W. McClintick, J. Harer, B. Flowers, W. C. Headley and A. M. Wyglinski, "Countering physical eavesdropper evasion with adversarial training," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1820-1833, Oct. 2022.
[31] 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.
[32] 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, June 2022.
[33] 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.
[34] X. Zhou, P. Qi, W. Zhang, S. Zheng, N. Zhang and Z. Li, "GAN-based siamese neuron network for modulation classification against white-box adversarial attacks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 1, pp. 122-137, Feb. 2024.
[35] Y. Dong, H. Wang and Y. -D. Yao, "A robust adversarial network-based end-to-end communications system with strong generalization ability against adversarial attacks," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 4086-4091.
[36] E. Shtaiwi, A. El Ouadrhiri, M. Moradikia, S. Sultana, A. Abdelhadi and Z. Han, "Mixture GAN for modulation classification resiliency against adversarial attacks," in *Proc. IEEE Glob. Commun. Conf.*, Dec. 2022, pp. 1472-1477.
[37] Z. Wang, W. Liu and H. -M. Wang, "GAN against adversarial attacks in radio signal classification," *IEEE Commun. Lett.*, vol. 26, no. 12, pp. 2851-2854, Dec. 2022.
[38] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, June 2016, pp. 770-778.
[39] 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, July 2021.
[40] I. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, May 2015, pp. 189-199.
[41] A. Kurakin, I. Goodfellow and S. Bengio, "Adversarial examples in the physical world," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2017, pp. 128-141.
[42] Y. Dong *et al.*, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, June 2018, pp. 9185-9193.
[43] J. Aigrain and M. Detyniecki, "Detecting adversarial examples and other misclassifications in neural networks by introspection,", 2019, *arXiv:1905.09186*.
[44] X. Jia, Y. Zhang, B. Wu, J. Wang and X. Cao, "Boosting fast adversarial training with learnable adversarial initialization," *IEEE Trans. Image Process.*, vol. 31, pp. 4417-4430, June 2022.
[45] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, June 2020, pp. 8681-8691.
[46] M. Liu, H. Zhang, Z. Liu and N. Zhao, "Attacking spectrum sensing with adversarial deep learning in cognitive radio-enabled internet of things," *IEEE Trans. Reliab.*, vol. 72, no. 2, pp. 431-444, June 2023.
[47] DeepSig, "Deepsig dataset: Radioml 2016.10b," 2016. [Online]. Available: https://www.deepsig.io/datasets.
[48] 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.