Adversarial Threats to Automatic Modulation Open Set Recognition in Wireless Networks

†Yandie Yang, †Sicheng Zhang, †Kuixian Li, ‡Qiao Tian, †Yun Lin*

†,* College of Information and Communication Engineering, Harbin Engineering University, Harbin, China
‡ College of Computer Science and Technology, Harbin Engineering University, Harbin, China
E-mail: {yangyandie, zhangsc, likuixian}@ieee.org, tianqheu@163.com, linyun@ieee.org

Abstract—Automatic Modulation Open Set Recognition (AMOSR) is a crucial technological approach for cognitive radio communications, wireless spectrum management, and interference monitoring within wireless networks. Numerous studies have shown that AMR is highly susceptible to minimal perturbations carefully designed by malicious attackers, leading to misclassification of signals. However, the adversarial security issue of AMOSR has not yet been explored. This paper adopts the perspective of attackers and proposes an Open Set Adversarial Attack (OSAttack), aiming at investigating the adversarial vulnerabilities of various AMOSR methods. Initially, an adversarial threat model for AMOSR scenarios is established. Subsequently, by analyzing the decision criteria of both discriminative and generative open set recognition, OSFGSM and OSPGD are proposed to reduce the performance of AMOSR. Finally, the influence of OSAttack on AMOSR is evaluated utilizing a range of qualitative and quantitative indicators. The results indicate that despite the increased resistance of AMOSR models to conventional interference signals, they remain vulnerable to attacks by adversarial examples.

Index Terms—Open set recognition, adversarial attack, automatic modulation classification, wireless network security

I. INTRODUCTION

Amidst the swift progress of wireless technology and the contraction of available spectrum resources, the complexity of spectrum sharing, interference diagnostics, and management within wireless networks has escalated significantly [1]. Automatic Modulation Classification (AMC) [2], [3] is a crucial step in demodulating signals with unknown modulation types, essential for frequency monitoring and security maintenance in complex wireless networks. Deep learning (DL) has become a widely used method in AMC [4], [5], due to its unique advantages such as autonomous analysis and nonlinear fitting. Traditional DL-based AMC methods rely on closed set assumptions and fail to account for unknown interference signals intentionally emitted by non-cooperative users, thus do not meet the practical needs of real-world environments. Consequently, Open Set Recognition (OSR) [6], which can identify unknown signal types has begun to receive widespread attention.

Given the limitations of deep neural network (DNN) inevitably has a closed set nature due to its use of the traditional softmax cross-entropy loss function during training. Bendale et al. [7] introduce a pioneering solution for open set deep learning architectures by substituting the softmax with an OpenMax that is fine-tuned using the Weibull distribution. Wen et al. [8] present the CenterLoss function as

an innovative supervisory signal for recognition tasks. It is designed to ascertain the feature centers of each class and imposes penalties on the divergence between features and their respective class centers. Kong et al. [9] synthesizes unknown class samples using Generative Adversarial Networks (GAN) to aid network training, offering explicit probability estimates for the generated unknown class samples. Sun et al. [10] presents an OSR algorithm based on Conditional Gaussian Distribution Learning (CGDL), which utilizes a probabilistic ladder architecture to preserve the representation of input information in intermediate layers and detects unknown samples by approximating latent features to different Gaussian models.

Lately, it has been uncovered by researchers that DNNs are susceptible to being fooled by adversarial examples to produce erroneous outputs. The concept of adversarial examples is first introduced by Szegedy et al. [11] and it is found that adversarial examples can be transferred between different models. Lin et al. [12] introduce carefully designed adversarial examples into modulation signals under different environments, validating the potential adversarial threats in AMC. To mitigate the frequency leakage and glitch caused by high frequency components in adversarial perturbations, Zhang et al. [13] devised the spectrum-focused frequency adversarial attack algorithm, enhancing the stealthiness of the attacks. Considering the openness of the electromagnetic space, Qi et al. [14] developed a black box adversarial attack of detection-tolerant that reduces the cost of attacks while increasing the success rate. In response to the adversarial threats, researchers have also proposed corresponding defense methods [15]–[17].

Although adversarial attacks and defenses are commonly studied in AMC, they have not been extensively explored in AMOSR. Rozsa et al. [18] design closed set adversarial attacks against deep features and find that OpenMax is as vulnerable by closed set adversarial perturbation as softmax. To explore the impact of open set adversarial attacks on AMOSR, this paper proposes Open Set Adversarial Attacks (OSAttack), including Open Set Fast Gradient Attack Methods (OSFGSM) and Open Set Projected Gradient Descent (OSPGD), aiming to draw attention to the adversarial security challenges in OSR. The main contributions of this paper are as follows:

- Building on research in adversarial attacks and AMOSR, we establish an adversarial threat model for AMOSR.
- Regarding the discrimination criteria for discriminative

and generative models, we propose the attack algorithm of OSFGSM and OSPGD.

- We carry out comprehensive experiments to evaluate the performance of AMOSR before and after OSAttack, discovering that AMOSR also exhibits adversarial vulnerabilities.

The remainder of this paper is organized as follows: Section II provides an overview of adversarial attacks and the adversarial threat model of AMOSR. Section III details the OSAttack algorithm and evaluation indicators. Section IV discusses the datasets used and experimental validation of OSAttack on effectiveness. Section V summarizes the findings and suggests directions for future research.

II. Background and Threat model

A. Background of Adversarial Attack

Adversarial examples x_{adv} are deliberately crafted inputs that aim to fool models by introducing specific perturbations. FGSM [19] as a classical gradient-based algorithm of adversarial examples generation, which modifies each point in the input sample by an equal amount of increase or decrease with the gradient direction of the loss function for multi classification model f_θ

$$x_{adv} = x + \varepsilon \cdot \text{sign}\big(\nabla_x J(\theta, x, y)\big), \quad (1)$$

where ε is the maximum perturbation limit, θ represents the parameters of f_θ, y is the label related to input x, $\nabla_x J(\cdot)$ is the gradient of loss function $J(\cdot)$ with respect x, and $\text{sign}(\cdot)$ represents the sign function. PGD [20] iteratively generates adversarial examples to improve upon FGSM

$$x_{adv}^{i+1} = \prod_{x+\delta} \Big(x_{adv}^i + \alpha \text{sign}\big(\nabla_x J(\theta, x, y)\big)\Big), \quad (2)$$

where i denotes iteration index, δ is adversarial perturbation, α represents the step size of perturbation, and the projection Π assigns its input to the nearest element within the perturbed input set $x + \delta$.

B. Adversarial Threat Model of AMOSR

The goal of AMOSR is to determine whether x belongs to unknown class and to correctly classify signals of known classes. Then, the goal of adversarial attacks against AMOSR is to induce misclassification of unknown signals as known classes by applying subtle perturbations, thereby evading electromagnetic spectrum monitoring and interference signal demodulation.

Let the training dataset be $\mathcal{D}_c = \{(x_i, y_i)\}_{i=1}^m$ in which $x_i \in \mathbb{R}^N$ and $y_i \in \{1, 2, ..., K\}$, and define the potential unknown dataset as $\mathcal{D}_u = \{(x_u, y_u) \mid y_u \in \{K+1\}\}$. During the testing phase of AMOSR, defining the test dataset $\mathcal{D}_o = \{(x_o, y_o) \mid y_o \in \{1, 2, ..., K+1\}\}$. The open space risk is the ratio of the amount of x_u recognized as closed sets to the amount of x_o recognized as closed sets

$$R_o(\mathcal{D}_u, \mathcal{D}_o; f_\theta) = \frac{\int_{\mathcal{D}_u} f_\theta(x) dx}{\int_{\mathcal{D}_o} f_\theta(x) dx}. \quad (3)$$

It can be observed that an increased misclassification of more x_u as known signals will raise R_o, therefore, the objective of adversarial attacks can be expressed as

$$\max_\delta R_o(D_{u,adv}, D_{o,adv}; f_\theta),$$
$$\text{s.t. } f_\theta(x_{u,adv}) \neq K+1, \quad (4)$$
$$x_{u,adv} = x_u + \delta \in \mathcal{D}_{u,adv},$$

where $\mathcal{D}_{u,adv}$ is the unknown dataset after adding adversarial examples, $\mathcal{D}_{o,adv}$ is the test dataset that includes adversarial unknown data. we aim increase R_o of target AMOSR model by adding minor perturbations to the received unknown signal.

III. Attack Methods and Measurement Indicator

This section describes the proposed OSAttack methods and attack performance measures. OSFGSM and OSPGD are developed for the discriminative and generative models.

A. OSAttack of Discriminative Model

1) **OpenMax and CenterLoss:** Discriminative models usually classify by increasing the difference between closed and open sets and setting thresholds τ. We analyze the discriminative rules of OpenMax, CenterLoss, and CGDL, and propose corresponding methods for generating adversarial examples.
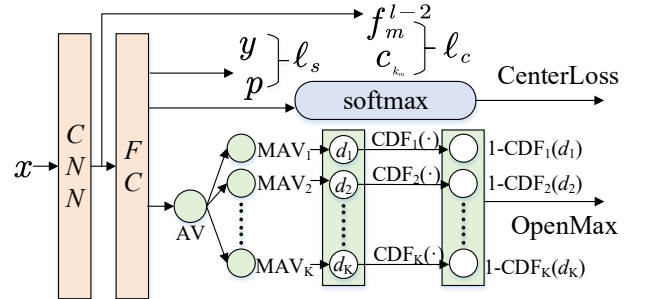

Fig. 1. The OSAttack framework diagram for OpenMax and CenterLoss.

As illustrated in Fig. 1, OpenMax method initially trains through DNN with by minimizing ℓ_s

$$\ell_s = CE(p, y), \quad (5)$$

where p is the logits, $CE(\cdot, \cdot)$ represents cross-entropy loss function. It computes Mean Activation Vectors (MAVs) for each class from correctly recognized samples, and fits a Weibull distribution for each class using the distance between these samples and their MAVs.

During the testing phase, the algorithm calculates the distance d between the activation vector (AV) and the MAV of each class, obtaining probabilities $CDF(d)$ through the $CDF(\cdot)$ function. After adjusting weights with $1 - CDF(d)$, it uses softmax to compute the new class probabilities $P(y = k|x)$. Define $y^* = \text{argmax} P(y = k|x)$, represents the predicted label. when $y^* = K+1$ or $P(y = y^*|x) < \tau$, x will be classified as belonging to an unknown class.

The CenterLoss method adds a loss function ℓ_c on the basis of ℓ_s

$$\ell_c = \frac{1}{2}\sum_{m=1}^{M}\|f_m^{l-2} - c_{k_m}\|_2^2, \qquad (6)$$

where f_m^{l-2} represents the features before the fully connected layer, c_{k_m} is the center for the deep feature corresponding to the k-th class of input samples. This approach reduces intra-class distances while enlarging inter-class distances, effectively distinguishing between classes. Finally the predicted probability is output by softmax, when $P(y = y^*|x) < \tau$, x is determined by CenterLoss to come from an unknown class.

To sum up, OpenMax recalibrates AVs based on Weibull distribution fitting and CenterLoss utilizes two loss functions training approach, thereby minimizing the open space for each known class and enhancing the probability distinction between open set and closed set classes. Accordingly, our attack aims to increase the probability that open set samples are misclassified as the most likely closed set classes, rendering open and closed sets more indistinguishable. Therefore, the adversarial examples generated by the OSFGSM and OSPGD algorithms are respectively denoted as (7) and (8):

$$x_{u,adv} = x_u - \varepsilon\mathrm{sign}(\nabla_{x_u}CE(p_u, y^*)), \qquad (7)$$

$$x_{u,adv}^{i+1} = \prod_{x+\delta}\left(x_{u,adv}^i - \alpha\mathrm{sign}(\nabla_{x_u}CE(p_u, y^*))\right). \qquad (8)$$

2) **CGDL:** The framework diagram of the CGDL method is shown in Fig. 2, which utilizes a variational autoencoder to construct reconstructed features, and classifies them by enforcing various Gaussian models approximated through different latent features.
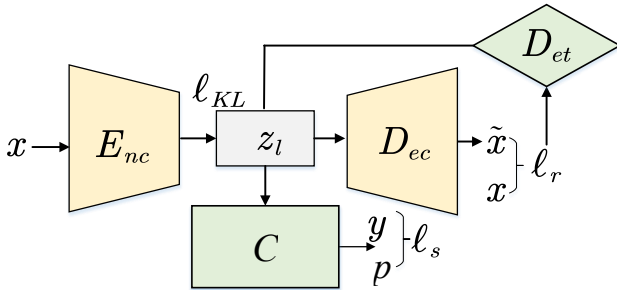

Fig. 2. The OSAttack framework diagram for CGDL.

The encoder E_{nc} and decoder Dec extract abstract latent features z_l by using a probabilistic ladder architecture, which in consequence achieves the desired reconstructed signal $\tilde{x}$

$$z_l = E_{nc}(x), \qquad (9)$$

$$\tilde{x} = D_{ec}(z_l). \qquad (10)$$

Then, z_l is used as input to the classifier C to get the predicted labels of the known samples

$$y^* = \mathrm{argmax}(C(z_l)). \qquad (11)$$

The training objective of CGDL is to minimize the sum of reconstruction loss ℓ_r, KL divergence ℓ_{KL} and classification loss ℓ_s. Specifically, the mean square error $MSE(x, \tilde{x})$ is used to calculate ℓ_r, the $CE(p, y)$ is used to calculate ℓ_s and ℓ_{KL} is calculated in the latent space and middle layers. In the testing phase, the unknown detector D_{et} will determine whether a sample is unknown by modeling the z_l and ℓ_r.

By combining ℓ_r and z_l for double discrimination, the CGDL method effectively reduces the open set risk of the input signal. In this paper, ℓ_r and ℓ_s constitute the constrain of the OSAttack used to generate the adversarial example, aiming to decrease the reconstruction error of open set signals by reducing ℓ_r and to increase the probability that open set signals are classified as closed set signals by encoder E_{nc} and classifier C through the reduction of ℓ_s, minimizing the difference between the z_l of open and closed set. Therefore, the loss function of OSAttack generating adversarial example for CGDL model is

$$\ell_{CGDL} = \frac{1}{2}CE(p_u, y^*) + \frac{1}{2}MSE(x_u, \tilde{x}_u), \qquad (12)$$

consequently, the OSFGSM and OSPGD algorithms for generating adversarial examples can be represented respectively as in (13) and (14)

$$x_{u,adv} = x_u - \varepsilon\mathrm{sign}(\nabla_{x_u}\ell_{CGDL}), \qquad (13)$$

$$x_{u,adv}^{i+1} = \prod_{x+\delta}\left(x_{u,adv}^i - \alpha\mathrm{sign}(\nabla_{x_u}\ell_{CGDL})\right). \qquad (14)$$
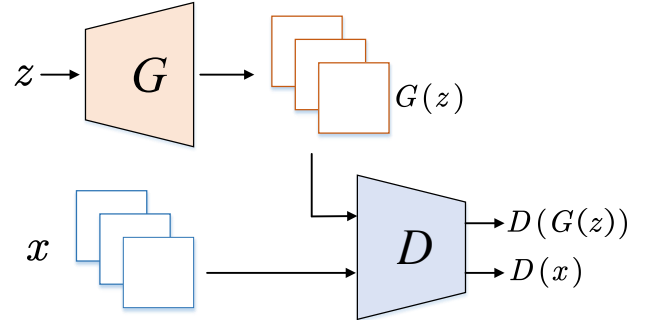
B. OSAttack of Generative Model

Fig. 3. The OSAttack framework diagram for OpenGAN.

In generative models, it is common for GANs to be employed to generate unknown samples so that DNNs can be exposed to a multitude of unknown samples. In Fig. 3, OpenGAN is a method that utilizes GANs to enhance OSR performance, where the discriminator D aims to distinguish between real and fake samples, while the generator G is used to produce fake samples.

First, D is optimized to achieve good discriminative performance

$$\max_{D}\mathbb{E}_{x\sim\mathcal{D}_{tr}}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))], \qquad (15)$$

where z as random noise taken from a Gaussian distribution $p(z)$, and $G(z)$ represents the fake samples produced by the generator.

To produce more realistic samples, it is necessary to optimize the G so that the samples it generates can effectively deceive the D

$$\min_{G} - \mathbb{E}_{z \sim p(z)}[\log D(G(z))]. \tag{16}$$

It is the overall training goal of OpenGAN to optimize both loss functions, so that the G learns to generate more and more real samples, and the D is able to distinguish between real and fake samples more effectively. Thus, this paper attacks the output of D within this structure, increasing the error with which the D incorrectly identifies open set samples

$$\ell_{OpenGAN} = BCE\left(D(x), y_g\right), \tag{17}$$

where $y_{\mathrm{g}} = D(x) > \tau_{\mathrm{g}}$, τ_{g} representing the probability threshold, which is constant close to 1. corresponding the OSFGSM and OSPGD method for generating adversarial examples can be expressed as (18) and (19)

$$x_{u,adv} = x_u + \varepsilon \mathrm{sign}(\nabla_{x_u}\ell_{OpenGAN}), \tag{18}$$

$$x_{u,adv}^{i+1} = \prod_{x+\delta}\left(x_{u,adv}^{i} + \alpha \mathrm{sign}(\nabla_{x_u}\ell_{OpenGAN})\right). \tag{19}$$

C. Evaluation of Indicators

Considering previous studies have widely proven that adversarial attacks pose a significant threat to CSR, this paper will not discuss it in detail but focuses primarily on the ability of AMOSR to recognize unknown class. In order to analyze the threat of adversarial attacks against AMOSR, we choose the accuracy of unknown samples (AUS) [21] as the evaluation indicator

$$\mathrm{AUS} = \frac{\mathrm{TU}}{\mathrm{TU} + \mathrm{FU}}, \tag{20}$$

where TU and FU are the number of correctly and incorrectly recognized unknown class samples, respectively.

In addition, we evaluate the performance of AMOSR using the Area Under the Receiver Operating Characteristic curve (AUROC). AUROC is the area under the ROC curve, which consists of True Positive Rate (TPR) and False Positive Rate (FPR).

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{21}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \tag{22}$$

where TP is the amount of true positives, FN is the amount of false negatives, FP is the amount of false positives, and TN is the amount of true negatives. The value of AUROC ranges from 0 to 1, which assesses the ability of model to distinguish between two classes by comparing TPR and FPR at various thresholds. This metric does not rely on specific classification thresholds, making it particularly suitable for evaluating model performance across different thresholds.

IV. EXPERIMENTS AND DISCUSSIONS

In this section, We describe a detailed introduction to the datasets used and conduct extensive experiments to validate the OSFGSM and OSPGD algorithms.
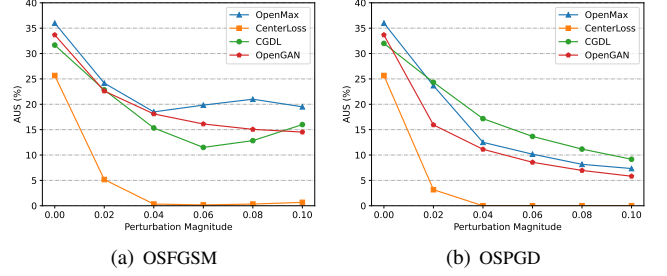

(a) OSFGSM   (b) OSPGD

Fig. 4. The AUS of different AMOSR models under OSFGSM and OSPGD attacks as the perturbation changes.
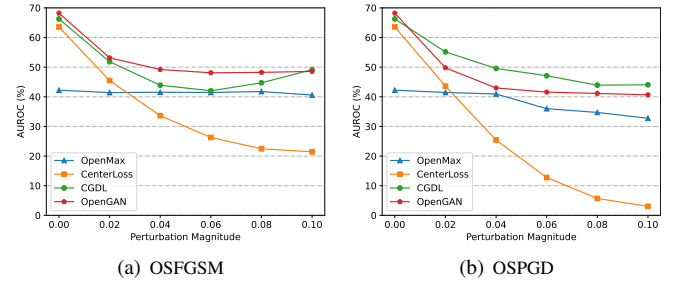

(a) OSFGSM   (b) OSPGD

Fig. 5. The AUROC of different AMOSR models under OSFGSM and OSPGD attacks as the perturbation changes.

A. Dataset

To investigate adversarial examples in wireless communication and their impact on AMOSR systems, this paper utilized the open-source simulated dataset RADIOML2016.10A developed by DEEPSIG. This publicly available dataset contains eight digitally modulated signals and three analog signals with varying signal-to-noise ratios(SNR). In our AMOSR task, eight categories were designated as known classes (BPSK, GFSK, CPFSK, PAM4, QAM64, WBFM, AM-DSB, AM-SSB) and three as unknown classes (8PSK, QPSK, QAM16). The dataset consists of 220,000 samples across 20 SNR levels (from -20dB to 18dB) with 2,000 samples for each signal type. In this work, it selects 18dB signals for experiments and divides the dataset into training, validation and test sets in the ratio of 6:2:2. For each signal vector, it consists of an in-phase component and an quadrature component, each of which has a length of 128.

B. Analysis of AUS

We explore the trends of the AUS of various OSR algorithms under different perturbation magnitudes after being attacked by OSFGSM and OSPGD in Fig. 4(a) and Fig. 4(b). It can be observed that the AUS of four methods decreases significantly after the attack, with the CenterLoss algorithm experiencing the most pronounced decrease. In the case of
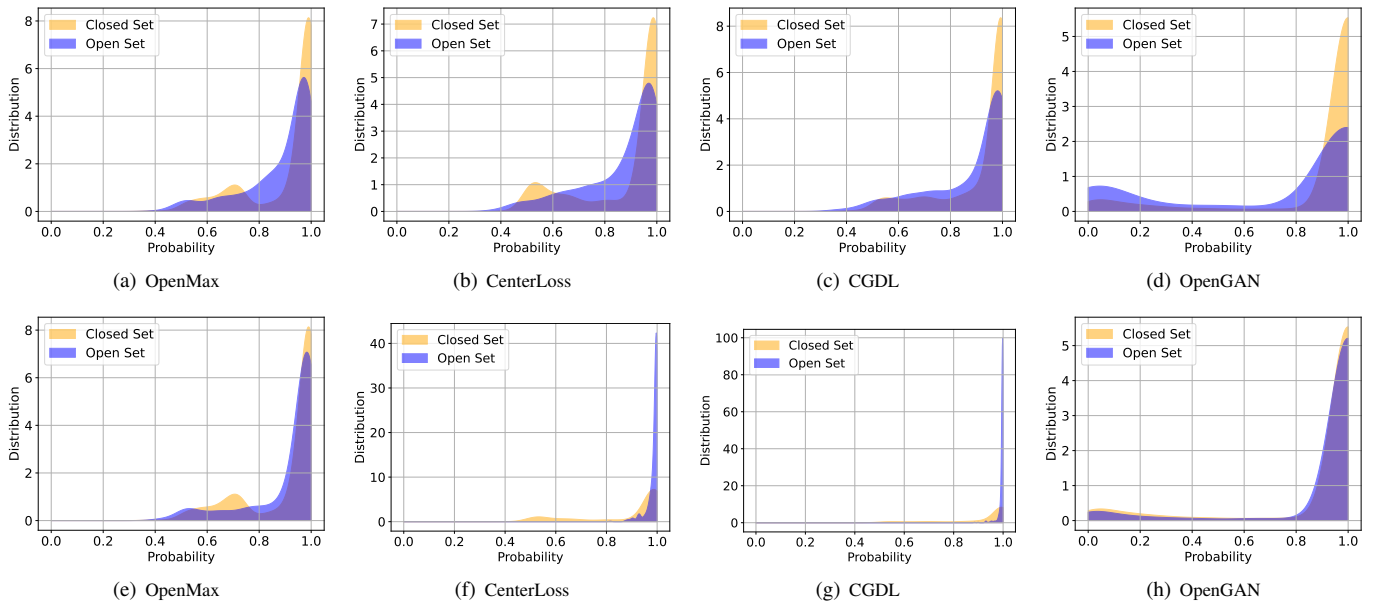
Fig. 6. Comparison of probability distribution in between closed set and open set.

TABLE I
COMPARISON OF AUS ACROSS DIFFERENT OSR METHODS UNDER VARIOUS OSATTACK.

Model	Clean(%)	OpenMax(%)		CenterLoss(%)		CGDL(%)		OpenGAN(%)	
		OSFGSM	OSPGD	OSFGSM	OSPGD	OSFGSM	OSPGD	OSFGSM	OSPGD
OpenMax	36.0	**20.67**	**10.17**	27.83	22.5	10.17	20.17	33.0	36.0
CenterLoss	25.67	23.83	19.17	**0.17**	**0.0**	15.0	16.83	28.0	26.33
CGDL	33.33	70.0	66.33	65.17	56.0	**11.5**	**13.67**	74.33	73.67
OpenGAN	33.67	31.17	33.5	35.5	39.0	6.0	1.0	**16.89**	**8.57**

the OSFGSM attack, a mere 0.04 perturbation is sufficient to reduce the AUS of CenterLoss to nearly zero. Similarly, in the OSPGD attack, even slight perturbations lead to a sharp decrease in AUS.

Additionally, for the OSFGSM attack, as the intensity of the attack increases, the AUS of the OpenMax and CGDL algorithms initially shows a gradual decline. However, when the perturbation exceeds 0.06, the AUS begins to gradually increase. This phenomenon is not observed in the OSPGD attacks. The difference primarily stems from the single step nature of the OSFGSM attack, which relies on a substantial gradient update to generate adversarial examples. In contrast, PGD finely tunes the perturbations through continuous multi step iterations, effectively preventing the model from correctly recognizing the samples due to excessive perturbation, thereby maintaining the efficacy and stability of the attack.

C. Analysis of AUROC

The change in AUROC with perturbation magnitudes is investigated in Fig. 5, as the perturbation magnitude increases, the AUROC also displays a gradual decline. Specifically, when the perturbation magnitude reaches 0.06, the AUROC for CenterLoss, openGAN, and CGDL all decline by more than 15%, dropping from initially high levels to below 50%. In

contrast, the decline in AUROC for OpenMax following an attack is less pronounced compared to the aforementioned methods, possibly because OpenMax inherently has a lower AUROC. Under the more elaborate OSPGD attack, there is still a substantial drop in performance for OpenMax, which drops by about 10%. The above results show that the OSAttack algorithm based on AMOSR decision criteria proposed in this paper significantly weakens the ability of the model to reject unknown samples, which reveals that adversarial attacks pose a wide range of threats to the AMOSR system and may lead to poorer recognition performance.

D. Analysis of Probability Distribution

In order to show in detail the changes in the output of the AMOSR model after adversarial attacks, we present the maximum predictive probability distributions of OpenMax, CenterLoss, and CGDL, as well as the predictive probability distribution of OpenGAN for real samples. Fig. 6(a)-(d) and Fig. 6(e)-(g) respectively demonstrate the probability distribution before and after OSAttack. In this part of the experiment, it is used the perturbation magnitude of 0.06.

It can be seen that the probability distribution of closed set signals before the model is attacked by OSAttack is mainly concentrated in the higher probability region, while the

distribution of open set signals is more dispersed. However, after the attack, the predicted probability of the AMOSR model for open set data increases, migrates to high probability regions, and overlaps significantly with the distribution of closed set signals, which increases the difficulty of threshold-based classification discrimination. Particularly, CenterLoss and CGDL predict probabilities close to 1 for most open set signals, demonstrating that even high confidence levels are no longer reliable for distinguishing unknown class.

E. Ablation Experiment

We analyze the effectiveness of the adversarial examples generated by OSAttack on the AMOSR model using an adversarial perturbation magnitude of 0.06 against other models in Table I, where the rows represent the models used to generate adversarial perturbations, and the columns indicate the models used for testing AUS. The diagonally bolded data reflects the AUS of a white box attack, while the other data shows the AUS of a black box attack. Results show that there is a significantly more effective white box attack, and adversarial examples generated using CGDL exhibit the best transferability, effectively deceiving other AMOSR models.

The OSAttack algorithm for CGDL generates adversarial examples by integrating ℓ_r with ℓ_s constraints. The ℓ_s modifies the probability distribution of the output layer to make the probabilistic responses of the adversarial examples more similar to those of closed set, effectively disguising them as closed set samples and blurring the distinction between open set and closed set. Meanwhile, ℓ_r optimizes the performance of adversarial examples in the encoder and decoder, making their intermediate feature representations closer to known samples, enhancing their deceptiveness, and improving the transferability and stealth of the attacks. This demonstrates that effective attacks can be achieved even without specific knowledge of the AMOSR models, and it suggests that future efforts could focus on enhancing transferability through increased constraints.

V. SUMMARY

In this paper, we propose OSAttack attack to evaluate the adversarial vulnerability of AMOSR models. By analyzing the characteristics of AMOSR and constructing an adversarial threat model for AMOSR. We delve into the discriminative mechanisms of four typical OSR models and apply the OSFGSM and OSPGD algorithms to generate adversarial examples at the decision level of these models. Qualitative and quantitative experimental results show that OSAttack is sufficient to significantly weaken the AMOSR performance even with small perturbations. Thus, although AMOSR show certain advantages in dealing with unknown signals, they are still vulnerable to well-designed adversarial signals, a finding that provides an important perspective for the research of future defense strategies. Future work will focus on developing adversarial attacks with feature layer constraints and exploring new attack strategies against specific targets in open sets.

REFERENCES

[1] Z. Wei, F. Liu, C. Masouros, N. Su, and A. P. Petropulu, "Toward multi-functional 6g wireless networks: Integrating sensing, communication, and security," *IEEE Commun. Mag.*, vol. 60, no. 4, pp. 65–71, 2022.

[2] Y. Lin, Y. Tu, Z. Dou, L. Chen, and S. Mao, "Contour stella image and deep learning for signal recognition in the physical layer," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 34–46, 2021.

[3] Y. Tu, Y. Lin, J. Wang, and J.-U. Kim, "Semi-supervised learning with generative adversarial networks on digital signal modulation classification," *Comput. Mater. Continua*, vol. 55, no. 2, pp. 243–254, 2018.

[4] Y. Tu, Y. Lin, C. Hou, and S. Mao, "Complex-valued networks for automatic modulation classification," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10 085–10 089, 2020.

[5] 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, 2020.

[6] T. Li, Z. Wen, Y. Long, Z. Hong, S. Zheng, L. Yu, B. Chen, X. Yang, and L. Shao, "The importance of expert knowledge for automatic modulation open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13 730–13 748, 2023.

[7] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 1563–1572.

[8] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 499–515.

[9] S. Kong and D. Ramanan, "Opengan: Open-set recognition via open data generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 813–822.

[10] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, and G. Peng, "Conditional gaussian distribution learning for open set recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 13 480–13 489.

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

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

[13] S. Zhang, J. Fu, J. Yu, H. Xu, H. Zha, S. Mao, and Y. Lin, "Channel-robust class-universal spectrum-focused frequency adversarial attacks on modulated classification models," *IEEE Trans. on Cogn. Commun. Netw.*, pp. 1–1, 2024, doi: 10.1109/TCCN.2024.3382126.

[14] P. Qi, T. Jiang, L. Wang, X. Yuan, and Z. Li, "Detection tolerant black-box adversarial attack against automatic modulation classification with deep learning," *IEEE Trans. Rel.*, vol. 71, no. 2, pp. 674–686, 2022.

[15] S. Zhang, Y. Yang, Z. Zhou, Z. Sun, and Y. Lin, "DIBAD: A Disentangled Information Bottleneck Adversarial Defense Method using Hilbert-Schmidt Independence Criterion for Spectrum Security," *IEEE Trans. Inf. Forensics Secur.*, pp. 1–1, 2024, doi: 10.1109/TIFS.2024.3372798.

[16] 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, 2022.

[17] S. Zhang, Y. Lin, J. Yu, J. Zhang, Q. Xuan, D. Xu, J. Wang, and M. Wang, "HFAD: Homomorphic filtering adversarial defense against adversarial attacks in automatic modulation classification," *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, 2024, doi: 10.1109/TCCN.2024.3360514.

[18] A. Rozsa, M. Günther, and T. E. Boult, "Adversarial robustness: Softmax versus openmax," *arXiv e-prints*, p. arXiv:1708.01697, Aug. 2018.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv e-prints*, p. arXiv:1412.6572, Dec. 2014.

[20] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv e-prints*, p. arXiv:1706.06083, Jun. 2017.

[21] P. R. Mendes Júnior, R. M. De Souza, R. d. O. Werneck, B. V. Stein, D. V. Pazinato, W. R. De Almeida, O. A. Penatti, R. d. S. Torres, and A. Rocha, "Nearest neighbors distance ratio open-set classifier," *Mach. Learn.*, vol. 106, no. 3, pp. 359–386, 2017.