Multi-Model Output Fusion-based Average Gradient Mapping Adversarial Attack

Sicheng Zhang[1], Mengchao Wang[1] Zhida Bao[1], Yandie Yang[1], and Qiao Tian[2]

[1] College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China
{2015080325, wangmengchao, baozhida, yyd_20000614}@hrbeu.edu.cn,
[2] College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
tianqiao@hrbeu.edu.cn
Corresponding Author: Qiao Tian

Abstract. To address the issue of weak transferability of adversarial attacks based on a single surrogate model, this paper proposes a Multi-Model Output Fusion-based Average Gradient Mapping adversarial attack method (MMOF-AGM). By weighting and combining the outputs of multiple models, it alleviates the issue of poor generalization that arises from over-reliance on the gradient of a single model. The average gradient mapping strategy preserves the relative magnitudes of gradients both before and after mapping, ensuring consistent updates. Experimental results demonstrate that the proposed approach outperforms traditional gradient-based attacks in both white-box and black-box settings, improving the effectiveness of adversarial examples across different radar signal modulation classification models.

Keywords: Radar signal modulation classification, adversarial example attack, multi-model output fusion, average gradient mapping

1 Introduction

Recently, the rapid advancement of Deep Learning (DL) technologies has been notable, with significant progress made in applying these techniques to radar signal modulation classification. DL-based models are capable of automatically extracting features from vast amounts of complex radar signal data, significantly improving the accuracy and efficiency of signal classification. Initially, researchers innovatively combined time-frequency transforms with deep learning architectures to develop intelligent recognition models based on two-dimensional time-frequency spectrograms [1]. To better meet the demands of real-time processing, many researchers, inspired by automatic modulation classification methods in the communication field [2], began using the in-phase and quadrature (IQ) components of radar signals directly as inputs to DL-based models for pulse modulation type classification [3].

However, existing research has shown that deep learning models are vulnerable to attacks involving specially crafted adversarial examples [4]- [9], and DL-based radar signal classification models face the same security risks. This has led to the gradual development of related attack research. Wang et al. [10] observed that radar signal classification models are typically based on time-frequency images and proposed a cross-modal attack method. This method constructs a surrogate model that incorporates time-frequency analysis, data quantization, and classifiers, successfully launching adversarial attacks on time-frequency-based radar signal classification models. However, in real-world adversarial attack scenarios, attackers cannot directly access the target model's knowledge. Instead, they must construct a local model, generate adversarial examples targeting the local surrogate model, and further exploit the transferability of these adversarial examples to attack the target model. Since attacks based on a single surrogate model often rely too heavily on the gradient of that model, leading to poor performance across different models and limited generalization, this paper proposes a Multi-Model Output Fusion-based Average Gradient Mapping adversarial attack method (MMOF-AGM) adversarial attack method. This method integrates the outputs of multiple models and applies average gradient mapping to process the gradients, improving the alignment of the gradient update direction and thereby enhancing the generalization capability of adversarial attacks from both the surrogate model and gradient correction perspectives.

2 Related Works

2.1 DL-Based Radar Signal Classification

Initially, researchers concentrated on leveraging both time-domain and frequency-domain features of radar signals. They employed time-frequency analysis techniques to convert these signals into image-like representations, which were subsequently classified using DL models. Xiao et al. [11] combined the texture features of time-frequency images with deep features extracted by deep neural networks to compensate for the limitations of deep characteristics used to represent image information, thereby enabling effective recognition of radar signals under low signal-to-noise ratios (SNR). Additionally, Huynh-The et al. [12] developed a network consisting of three primary complex modules and incorporated a skip connection mechanism to facilitate multi-scale feature fusion, thereby enabling the effective classification of 13 types of radar signals under low SNR.

Subsequently, research has increasingly focused on directly utilizing the IQ data of radar signals as input to DL-models for end-to-end recognition. Wei et al. [13] introduced an innovative network that combines shallow convolutional neural networks, long short-term memory networks, and deep neural networks to classify radar signals across an SNR range from -14 to 20dB. Furthermore, Zhang et al. [14] proposed a unified recognition network that incorporates a local feature extraction module in conjunction with a global similarity mining module. This network takes IQ-format signal data as input to classify pulse modulation

types in radar signals under low SNR, requiring fewer computational resources compared to time-frequency image-based methods.

In summary, both radar signal classification methods offer distinct advantages. Time-frequency image-based methods effectively extract both time-domain and frequency-domain features from radar signals, offering DL models richer feature inputs. In contrast, IQ data-based methods leverage the amplitude and phase information from raw radar signal data, making them less resource-intensive.

2.2 Adversarial Attacks in Radar Domain

With the advancement of deep learning technologies, security concerns related to neural network models have gradually emerged. Among these, adversarial attacks, as a major security threat, have garnered widespread attention. Researchers in radar signal classification have also given this issue the necessary attention and invested in related research. Du et al. [15] proposed an attack method that injects specific perturbations into the vulnerable range cells of radar high-resolution range profiles, effectively deceiving radar automatic target recognition models. Furthermore, the researchers employed Generative Adversarial Networks (GANs) to generated adversarial examples, resulting to inaccuracies in Synthetic Aperture Radar (SAR) image target recognition models [16].

While the aforementioned research has contributed to the advancement of adversarial attacks in the radar field, it remains relatively dependent on the output information of a single model, which imposes certain limitations. Therefore, this paper propose to leverage the output information from multiple surrogate models and applies average gradient mapping to enhance the attack performance of adversarial examples across various radar signal classification models.

3 Methodology

3.1 Fusion of Multi-model Output

Since attacks based on a single surrogate model often depend heavily on the characteristics of the selected model, they exhibit weak generalization across different models, thereby limiting the effectiveness and adaptability of the attack. Therefore, this paper proposes a weighted fusion of outputs from multiple models to fully leverage the output information of several models, guiding the generation of adversarial examples. Fig. 1 illustrates the detailed process for computing the fusion of multi-model outputs.

Additionally, this paper considers potential samples in the neighborhood of the current input sample. Specifically, through random sampling, the results from multiple models are combined using weighted fusion to generate more universally applicable perturbations. In the t-th iteration, the output of the n-th input sample is randomly selected from M models within a set containing K models:

$$l_t^n(x_t^n) = \sum_{m=1}^{M} \omega_m l_m(x_t^n), \tag{1}$$

$$\sum_{m=1}^{M} \omega_m = 1, \tag{2}$$

$$x_t^{n+1} = x_t^n + \delta_n, \tag{3}$$

where $l_m(x_t^n)$ represents the output of model m, $l_t^n(x_t^n)$ is the weighted combination of the outputs from multiple models, and ω_m denotes the weights assigned to the output of different models. In Equation (3), the value range of δ_n is $[-(\lambda \cdot \varepsilon)^d, (\lambda \cdot \varepsilon)^d]$, λ is the hyperparameter that controls the sampling range.

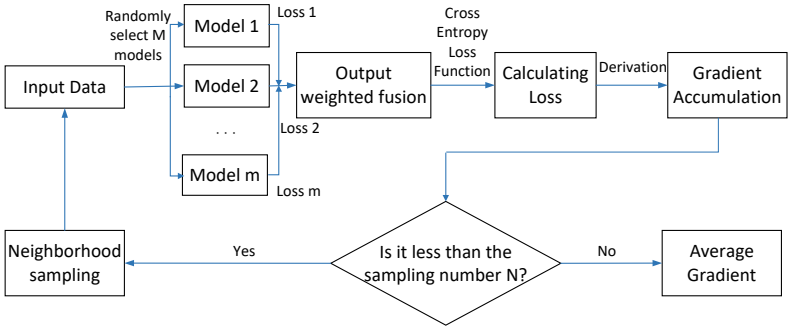

Fig. 1: Multi-model output fusion flow chart.

By combining the outputs of multiple models, this approach balances the inherent strengths and weaknesses of each model. Integrating the sensitivities of various models optimizes the generation of perturbations, leading to adversarial examples that demonstrate robust attack performance across different models. This method overcomes the limitations of single-surrogate model attacks and mitigates the issue of weak attack transferability caused by over-reliance on the characteristics of a single model.

3.2 Average Gradient Mapping

Typical gradient-based adversarial attack methods often rely on the use of the sign(·) function. The sign(·) function performs symbolic operations on the gradient, allowing the gradient update process to generate sufficient perturbations, thereby facilitating the rapid convergence adversarial examples toward the target class. Although effective, this method is not necessarily the optimal or near-optimal choice. The perturbation direction generated by the sign(·) function has inherent limitations, particularly during the gradient update process, where it disregards the specific magnitude information of the gradient. This can lead to suboptimal gradient optimization directions when generating adversarial examples.

As shown in Fig. 2, to address the limitations of the sign(·) function in generating adversarial examples, this paper proposes an average gradient mapping

method. This method enhances the transfer of gradient information by adjusting the mapping factor during the gradient update process, thereby refining the direction of gradient updates and improving the attack performance of the generated adversarial examples.

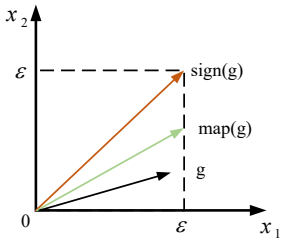

Fig. 2: Gradient direction

According to Equation (1), the output for the i-th input sample at the t-th iteration of the adversarial example generation process is determined by selecting a random subset of m models. The average gradient at this point can be computed as follows:

$$L(x_t^n, y) = -\mathbf{1}_y \cdot \log(soft\max(l_t^n(x_t^n))), \tag{4}$$

$$\overline{g_t} = \frac{1}{N+1} \sum_{n=0}^{N} \nabla L(x_t^n, y), \tag{5}$$

the average gradient is then mapped:

$$map(\overline{g_t}) = \gamma \cdot sign(\overline{g_t}) \cdot sigmoid(f(\log_p |\overline{g_t}|)), \tag{6}$$

$$f(a) = \frac{a - mean(a)}{std(a)}, \tag{7}$$

where $-\mathbf{1}_y$ represents the one-hot encoding form of the label y, and $L(x_t^n, y)$ is the loss calculated using the cross-entropy loss function. $\overline{g_t}$ denotes the average gradient obtained after N iterations of sampling, and the γ in the gradient mapping function map($\overline{g_t}$) is the mapping factor, while the function $f(a)$ standardizes and scales the data.

First, the absolute value of each element in the average gradient is calculated, followed by normalization through standardization and scaling. When applying the logarithmic operation with base 2 in Equation (6), the method exhibits a more pronounced dispersive effect. By mapping intermediate results to a specific function, we achieve smooth processing of the gradient data. The use of the sigmoid($\cdot$) function to smooth the data within the range [0,γ] is particularly effective.

Average gradient mapping enables flexible adjustment of the gradient distribution characteristics, ensuring that the relative magnitudes of the gradient

data are preserved during the smoothing process. This effectively retains the key features of the original gradients, preventing potential distortion or information loss. As a result, the processed average gradient mapping data reduces fluctuations while maintaining critical information from the original gradients.

In summary, by integrating output information from multiple models and fully leveraging gradient information from different models, combined with average gradient mapping to adjust the gradient update direction, the attack effectiveness across various models is significantly enhanced.

4 Experiments

First, we constructed the dataset required for the experiment and developed a baseline model, which served as a benchmark for conducting white-box attack experiments. Subsequently, to further evaluate the proposed method's generalization capability, we performed black-box attack experiments across different target models, offering a comprehensive assessment of the method's efficacy and adaptability in various attack scenarios and across diverse target models.

4.1 Dataset

Table 1: Dataset structure

Parameters	Content
Modulation types	Rect, LFM, Barker, Frank, P1, P2, P3, P4, T1, T2, T3, T4
Format of dataset	2×1024; IQ
Range of SNR	[-20,18]dB
Number of dataset	240000
Train: Valid: Test	7:2:1

The dataset consists of 12 types of radar signal waveforms, with the detailed dataset structure presented in Table 1. All signals are stored as IQ pair signals, with each signal sample formatted as 2×1024. The SNR ranges from -20dB to 18dB, with increments of 2dB. At each SNR level, 1,000 signal samples are generated, resulting in a total of 240,000 radar signal samples in the dataset. To more accurately simulate real-world scenarios, all radar signal waveforms are subjected to Gaussian white noise addition, followed by multipath Rayleigh fading. The experimental dataset is then divided into training, validation, and test sets in a 7:2:1 ratio.

Table 2 presents the multipath Rayleigh channel conditions, where U denotes the uniform distribution.

Table 2: Channel Configuration

Item	Rayleigh fading
Path delay	$U(1, 1000)$ ns
Average path gain	$U(-20, 0)$ dB
Maximum doppler shift	$U(10, 1000)$ Hz

4.2 Baseline Model

We developed a baseline model for radar signal classification, and the network architecture is depicted in Fig. 3.

The baseline model is based on a conventional CNN architecture, consisting of 10 convolutional layers, 2 max-pooling layers, and 3 fully connected layers. The convolutional layers automatically extract local features from the input data, learning progressively more complex representations at each layer, which enables the model to effectively capture the time-domain features of the signals. After each convolutional layer, a ReLU activation function is applied to introduce non-linearity, thereby enhancing the model's ability to capture complex patterns. The max-pooling layers are employed for downsampling, reducing the size of the feature maps while preserving important features. This enhances the model's computational efficiency and robustness. By decreasing the number of parameters, the pooling layers also mitigate the risk of overfitting, thereby improving the model's generalization ability.

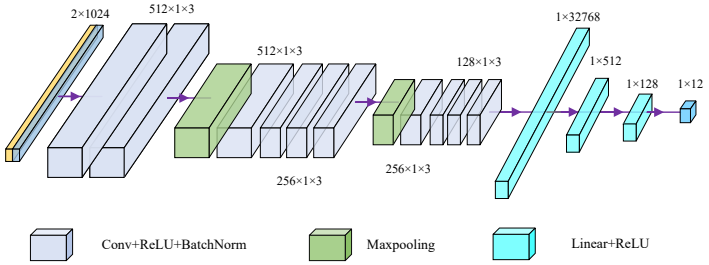

Fig. 3: Network structure diagram of the baseline model

After feature extraction, the model employs 3 fully connected layers to integrate high-level features and make classification decisions. These layers linearly combine the features processed by the convolutional and pooling layers, thus performing the final classification task. This network architecture, through deep convolutional and pooling operations, effectively handles complex signal patterns, demonstrating robust feature extraction and representation capabilities, making it highly effective for radar signal recognition.

In addition to the baseline model, ResNet-18 [17], ResNet-50 [17], VGG-16 [18], and GoogLeNet [19] were also extensively trained on the dataset to assess the attack performance of various adversarial attack methods.

4.3 Classification Performance of Different Models

As shown in Fig. 4, multiple classification models were utilized, and their classification accuracy was evaluated and compared. From Fig. 4, it is evident that as the SNR increase, the accuracy of all models improves steadily.

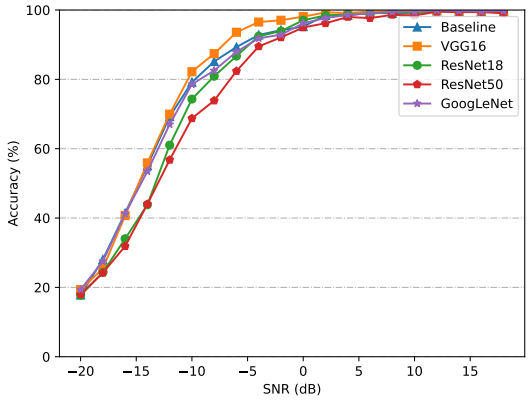

Fig. 4: Recognition accuracy of different models

The experimental results indicate that once the SNR exceeds 0dB, the recognition accuracy of all models gradually stabilizes. The accuracy curves of all models exhibit a similar trend, suggesting that these models yield relatively consistent recognition outcomes on the same test dataset. Although slight numerical variations are observed, the overall trend demonstrates that the selected models perform effectively in radar signal recognition, providing a reliable benchmark for the subsequent evaluation of adversarial attack effectiveness.

4.4 White-Box Attack Effect

In this experiment, the Fast Gradient Sign Method (FGSM) [20], Projected Gradient Descent (PGD) [21] and Basic Iterative Method (BIM) [22] are used as comparison attack methods. The perturbation intensity is set to 0.05, with a step size of 0.02 for PGD, BIM and MMOF-AGM (K=M=1), and the number of iterations is set to 10. As described previously, when K=M=1, MMOF-AGM reduces to an average gradient attack based on the output of a single model.

As shown in the experimental results in Fig. 5, the FGSM, PGD, BIM and MMOF-AGM (K=M=1) attack methods all result in varying degrees of accuracy degradation. For FGSM attacks, although the attack effect is more pronounced under low SNR conditions, causing a significant accuracy drop, the attack effect gradually diminishes as the SNR increases. As a single-step attack method, the effectiveness of FGSM weakens as the model's resistance to adversarial perturbations increases with higher SNR. In contrast, iterative attack methods like PGD and BIM, can optimize the perturbation more effectively by updating the gradients multiple times, thereby reducing the confidence of the target model and demonstrating stronger attack capabilities. BIM and PGD

have almost the same attack performance because the difference between them lies in whether there is a random initial perturbation when generating adversarial perturbations. When the SNR is below -6 dB, PGD, BIM and MMOF-AGM (K=M=1) can fully deceive the target model under complex noise conditions, achieving efficient attack results. As the SNR increases, and when it exceeds -6 dB, MMOF-AGM (K=M=1) outperforms both PGD and BIM in terms of attack performance, demonstrating its superior attack capability. In summary, under the same conditions, MMOF-AGM (K=M=1) consistently exhibits better attack performance than FGSM, PGD and BIM across various SNR levels.

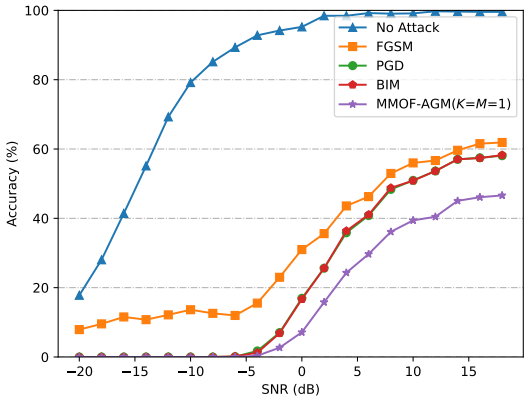

Fig. 5: Baseline model recognition accuracy under different white-box attack algorithms

4.5 Black-Box Attack Effect

Black-box attacks are considered more representative of real-world scenarios compared to white-box attacks, as the attacker lacks knowledge of the target model's structure and parameters, only having access to the input-output relationship. The perturbation size for all four attack methods is set to 0.05, with a step size of 0.02 for PGD, BIM and MMOF-AGM (K=M=1), and the number of iterations is set to 10.

We first investigate the black-box attack performance of MMOF-AGM (K=M=1) across five models. The experimental results are presented in Table 3.

In Table 3, the rows correspond to the models generating adversarial perturbations, while the columns represent the models used to test classification accuracy. The diagonal entries show the accuracy under white-box attacks, whereas the off-diagonal entries indicate the accuracy under black-box attacks. From the data in Table 3, it is evident that the attack performance in the white-box scenario surpasses that in the black-box scenario. This is because, in the white-box setting, the structure and gradients of the target model are accessible, enabling the generation of more effective adversarial examples. In contrast, in the black-box setting, where the relevant information of the target model is unavailable and there are gradient differences between the surrogate and target models, the

adversarial examples produced are less effective compared to those generated through white-box attacks. Although the attack performance of MMOF-AGM (K=M=1) varies across different models, it consistently demonstrates robust performance, further emphasizing the method's ability to generate adversarial perturbations with strong generalization capabilities.

Table 3: Average recognition accuracy of different models under MMOF-AGM(K=M=1) attack

Model	Baseline	VGG16	ResNet18	ResNet50	GoogLeNet
Baseline	16.68%	44.87%	39.36%	34.99%	35.35%
VGG16	49.54%	19.00%	43.98%	45.15%	44.60%
ResNet18	43.99%	41.75%	21.07%	32.94%	36.43%
ResNet50	44.01%	49.08%	38.39%	18.50%	39.69%
GoogLeNet	43.83%	47.14%	41.19%	38.99%	18.01%

By comparing the experimental results shown in Fig. 5 and Fig. 6, it is evident that, in the black-box attack scenario, the performance of all attack methods is inferior to that in the white-box attack scenario. This phenomenon suggests that the absence of internal model information in the black-box environment restricts the effectiveness of the attacks. Specifically, FGSM, PGD and BIM exhibit a similar decline in their attack effectiveness. However, both PGD and BIM continue to outperform FGSM in the black-box setting, as it can adjust the perturbation through multiple updates. Similar to the white-box scenario, BIM and PGD have similar attack performance in the black-box scenario.

Although the MMOF-AGM (K=M=1) algorithm experiences some performance degradation in the black-box environment, it still demonstrates superior attack effectiveness compared to FGSM, PGD and BIM across different SNR conditions. As shown in Fig. 6, the attack performance of MMOF-AGM (K=4, M=2) is the most effective. This can be attributed to its integration of output information from multiple models and the application of the average gradient mapping strategy, which enhances the transferability of adversarial examples in the black-box scenario.

5 Conclusion

This paper introduces an adversarial attack approach based on average gradient mapping, which integrates outputs from multiple models to enhance the attack performance and generalization capability of the generated adversarial examples across various recognition models. By merging the output information from multiple models, this method overcomes the limitations associated with relying on a single model. The average gradient mapping strategy processes the gradient data, ensuring that the gradient update direction of the generated adversarial examples aligns more effectively with the target model's gradient. This addresses a common issue in gradient-based attacks, where the relative magnitude of the

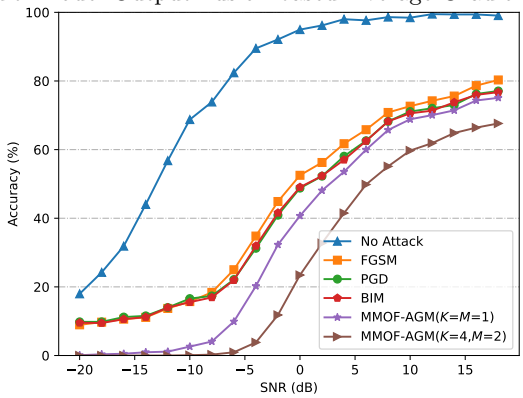

Fig. 6: Recognition accuracy of ResNet50 under different black box attack algorithms

gradient is often overlooked, leading to suboptimal gradient direction fitting. Experimental results show that the proposed method outperforms traditional gradient-based attacks in both white-box and black-box scenarios, confirming the effectiveness and generalizability of the adversarial examples generated. This research significantly enhances the radar system's counter-reconnaissance capability in complex electromagnetic environments, thereby providing robust assurance for the secure, stable, and covert operation of radar detection missions.

6 Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant 62201172.

References

1. Wang C, Wang J, Zhang X.: Automatic radar waveform recognition based on time-frequency analysis and convolutional neural network. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2437-2441. New Orleans (2017).
2. Tu Y, Lin Y, Hou C, Mao S.: Complex-Valued Networks for Automatic Modulation Classification. IEEE Transactions on Vehicular Technology, vol. 69, no. 9, pp. 10085-10089 (2020).
3. Li F, Wang Y, Zhao L, Yang Z.: Radar modulation recognition based on MLP neural network. In: 2019 International Conference on Microwave and Millimeter Wave Technology (ICMMT), pp. 1-3. Guangzhou (2019).
4. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R.: Intriguing properties of neural networks. https://arxiv.org/abs/1312.6199.
5. Lin Y, Zhao H, Ma X, Tu Y, Wang M.: Adversarial Attacks in Modulation Recognition With Convolutional Neural Networks. IEEE Transactions on Reliability, vol. 70, no. 1, pp. 389-401 (2021).
6. Zhao H, Lin Y, Gao S, Yu S.: Evaluating and Improving Adversarial Attacks on DNN-Based Modulation Recognition. In: GLOBECOM 2020 - 2020 IEEE Global Communications Conference, pp. 1-5. Taipei (2020).

7. Lin Y, Zhao H, Tu Y, Mao S, Dou Z.: Threats of Adversarial Attacks in DNN-Based Modulation Recognition. In: IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, pp. 2469-2478. (2020)
8. Bao Z, Lin Y, Zhang S, Li Z, Mao S.: Threat of Adversarial Attacks on DL-Based IoT Device Identification. IEEE Internet of Things Journal, vol. 9, no. 11, pp. 9012-9024 (2022).
9. Zhang S, Fu J, Yu J, Xu H, Zha H, Mao S, Lin Y.: Channel-Robust Class-Universal Spectrum-Focused Frequency Adversarial Attacks on Modulated Classification Models. IEEE Transactions on Cognitive Communications and Networking, vol. 10, no. 4, pp. 1280-1293 (2024).
10. Wang M, Zhang S, Xuan Q, Lin Y.: CMA: A Cross-Modal Attack on Radar Signal Recognition Model Based on Time-Frequency Analysis. In: ICC 2024 - IEEE International Conference on Communications, pp. 2119-2124. Denver (2024).
11. Xiao Y, Liu W, Gao L.: Radar signal recognition based on transfer learning and feature fusion. Mobile Networks and Applications, vol. 25, no. 4, pp. 1563-1571 (2020).
12. Huynh-The T, Doan V-S, Hua C-H, Pham Q-V, Nguyen T-V, Kim D-S.: Accurate LPI Radar Waveform Recognition With CWD-TFA for Deep Convolutional Network. IEEE Wireless Communications Letters, vol. 10, no. 8, pp. 1638-1642 (2021).
13. Wei S, Qu Q, Su H, Wang M, Shi J, Hao X.: Intra-pulse modulation radar signal recognition based on CLDN network. IET Radar, Sonar & Navigation, vol. 14, no. 6, pp. 803-810 (2020).
14. Zhang Z, Zhu M, Li Y, Wang S.: JDMR-Net: Joint Detection and Modulation Recognition Networks for LPI Radar Signals. IEEE Transactions on Aerospace and Electronic Systems, vol. 59, no. 6, pp. 7575-7589 (2023).
15. Du C, Cong Y, Zhang L, Guo D, Wei S.: A Practical Deceptive Jamming Method Based on Vulnerable Location Awareness Adversarial Attack for Radar HRRP Target Recognition. IEEE Transactions on Information Forensics and Security, vol. 17, pp. 2410-2424 (2022).
16. Du C, Zhang L.: Adversarial attack for SAR target recognition based on UNet-generative adversarial network. Remote Sensing, vol. 13, no. 21, pp. 4358 (2021).
17. He K, Zhang X, Ren S, Sun J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778. LAS VEGAS (2016).
18. Liu S, Deng W.: Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 730-734. Kuala Lumpur (2015).
19. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. Boston (2015).
20. Goodfellow I, Shlens J, Szegedy C.: Explaining and harnessing adversarial examples. https://arxiv.org/abs/1412.6572.
21. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A.: Towards deep learning models resistant to adversarial attacks. https://arxiv.org/abs/1706.06083.
22. Kurakin A, Goodfellow I, Bengio S.: Adversarial examples in the physical world. https://arxiv.org/abs/1607.02533.