Signal Augmentation Method based on Mixing and Adversarial Training for Better Robustness and Generalization

Li Zhang, Gang Zhou, Gangyin Sun, and Chaopeng Wu

***Abstract*—More and more deep learning methods have been applied to wireless communication systems. However, the collection of authentic signal data poses challenges. Moreover, due to the vulnerability of neural networks, adversarial attacks seriously threaten the security of communication systems based on deep learning models. Traditional signal augmentation methods expand the dataset through transformations such as rotation and flip, but these methods improve the adversarial robustness of the model little. However, common methods to improve adversarial robustness such as adversarial training not only have a high computational overhead but also potentially lead to a decrease in accuracy on clean samples. In this work, we propose a signal augmentation method called adversarial and mixed-based signal augmentation (AMSA). The method can improve the adversarial robustness of the model while expanding the dataset and does not compromise the generalization ability. It combines adversarial training with data mixing and then interpolates selected pairs of samples to form new samples in an expanded dataset consisting of original and adversarial samples thus generating more diverse data. We conduct experiments on the RML2016.10a and RML2018.01a datasets using automatic modulation recognition (AMR) models based on convolutional neural networks (CNN), long short-term memory (LSTM), convolutional long short-term deep neural networks (CLDNN), and transformer. And compare the performance in scenarios with different numbers of samples. The results show that AMSA allows the model to achieve comparable or even better adversarial robustness than using adversarial training, and reduces the degradation of the model's generalization performance on clean data.**

***Index Terms*—Adversarial training, automatic modulation recognition, data augmentation, mixing signals, robustness.**

I. Introduction

THE modern communication environment exhibits a trend of wide spectrum, abundant quantity, and high accuracy in the recognition of radio signals. The modulation modes of signals become more complex and diverse to meet the demands of increasingly complex communication scenarios, which makes feature extraction and recognition based on prior knowledge face many difficulties. Traditional signal processing methods make it difficult to process a large number of signals quickly and efficiently. Therefore, more and more research [1]–[3] began to use deep learning algorithms to identify the modulation types, making signal recognition tend to be automated. It avoids the requirement of artificial feature extraction based on experience and prior knowledge and improves the recognition ability of complex radio signals.

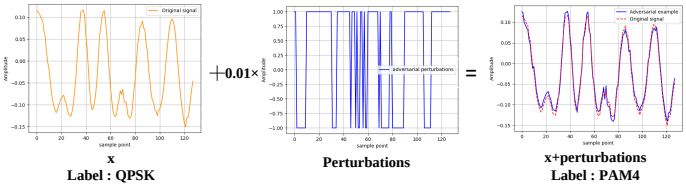


Fig. 1. Example of signal adversarial attack.

Firstly, automatic modulation recognition (AMR) based on deep learning has the problem of insufficient training samples. The performance of deep learning models is inseparable from the support of large volumes of data. The current standard data sets for radio signals are the RML2016 [4] and RML2018 series, which are obtained by simulating signals and channels through GNU Radio, and HisarMod2019.1 [5], which are obtained using MATLAB [6].However, to develop deep learning-based AMR models that can effectively operate under real channel conditions, it is imperative to have access to data sets comprising real-life scenarios. It is difficult and costly to collect and label a large amount of real-world radio signals, resulting in a scarcity of authentic signal datasets.

One way to solve the problem of insufficient samples is to design models that give better performance even with fewer training samples. However, this design is often not universal and requires a separate design for each model. A simple but effective way is to utilize data augmentation to reduce the dependence on training data. The training set is artificially expanded by data augmentation and the dataset can be used by multiple models. Traditional signal augmentation methods generate new data by transformations such as rotation, cropping, mixing, adding Gaussian noise, etc. to the signal, which can improve the generalization of the model.

In addition to the challenge of insufficient data, AMR models based on deep learning also have the vulnerability of neural networks. In the realm of image recognition, it was initially proposed that neural network predictions can be manipulated by applying imperceptible perturbation [7]. In the domain of signal recognition, related studies [8], [9] have demonstrated the feasibility of generating adversarial samples for deep learning-based AMR models by introducing disruptive noise to the waveform, resulting in errors. Adversarial

Manuscript received May 5, 2024; revised September 30, 2024; approved for publication by Hur, Junbeom Division 3 Editor, November 7, 2024.

The authors are with the Information Engineering University, Zhengzhou 450000, China. email: lzhang00@126.com, gzhougzhou@126.com, sungangyin2000@163.com, and wcpeng2023@163.com.

G. Zhou is the corresponding author.

Digital Object Identifier: 10.23919/JCN.2024.000067

1229-2370/24/$10.00

jamming, as depicted in Fig. 1, which requires less power and offers greater covertness compared to traditional noise jamming, has emerged as one of the methods employed for signal interference.

When performing AMR, the model is required to have robustness against adversarial attacks. Traditional signal augmentation methods provide less improvement in the model's defensiveness against adversarial attacks. For defense against adversarial attacks, the most commonly used effective method is adversarial training [10]–[12], which learns the features of adversarial attacks by adding adversarial samples to the training set. However, since the adversarial samples are concentrated near the decision boundary, resulting in differences in the distribution of the samples, researches [10], [11] have shown that adversarial training may lead to reduced recognition rates for clean samples.

To address the above problems, this paper proposes a data augmentation method that enables the model to have sufficient adversarial robustness and reduced degradation of generalization performance on clean samples. Mixed data augmentation not only enhances the diversity of data but also adjusts the distribution of samples by generating data samples that incorporate different information. Therefore, we consider the method of combining adversarial training and data mixing, using adversarial training to enhance adversarial robustness, and using data mixing to improve the generalization while adjusting the data distribution to improve the shortcomings of adversarial training. The method mainly has two difficult problems, one is that the signal has characteristics different from image and text. It's necessary to design a way suitable for signal data mixing. The second is how to combine adversarial training with data mixing to alleviate the generalization performance degradation brought about by the adversarial training in some scenarios, and to achieve the simultaneous improvement of robustness and generalization.

This paper proposes the adversarial and mixed-based signal augmentation (AMSA) method. Firstly, adversarial attacks are employed to generate corresponding adversarial examples, maintaining the information of original labels through the creation of mixed labels, while introducing information predicted by the model. Subsequently, the original signals are interpolated with their adversarial examples and signals of the same class, rather than randomly mixed across different types. Finally, we obtain a data set composed of the original samples, adversarial samples, and interpolated mixed samples. Our contributions are as follows:

- We propose a simple and effective data augmentation method, AMSA, which improves adversarial training and data mixing to adapt to one-dimensional signals, aiming to improve the model's adversarial robustness while alleviating the decrease in generalization performance caused by adversarial training.
- Aiming at the threat of white-box attacks faced by deep learning models, the AMSA proposed in this paper is evaluated under different signal-to-noise ratios (SNR) on multiple AMR models based on convolutional neural networks (CNN), long short-term memory (LSTM), convolutional long short-term deep neural networks (CLDNN), and Transformer, respectively.
- For the scenario of insufficient training data, the situation is simulated using 10% of the training dataset, and the performance of AMSA and traditional data enhancement methods are comprehensively compared.

II. Related Works

A. AMR Model based on Deep Learning

Traditional methods for modulation recognition include likelihood-based and feature-based approaches. However, these methods face challenges in adapting to the increasingly complex electromagnetic environment. Consequently, there is a growing trend towards the utilization of deep learning models for AMR tasks. Deep learning-based AMR methods are primarily implemented through CNN [1] and recurrent neural networks (RNN). CNNs learn signal spatial relationships through the local perception of convolutional kernels, and O'Shea [2] *et al*. combined CNNs with the fully connected network to achieve radio data feature extraction and data classification. RNNs, such as LSTM, control information flow through memory units to extract temporal features of signals. Rajendran [3] *et al*. transformed the IQ signals into phase and amplitude for modulated classification, and the temporal characteristics of radio data can be adequately captured using the LSTM structure. Additionally, CLDNN [13] have been shown to produce promising results by combining the advantages of CNN and RNN. However, CNN and RNN can only capture local signal features without realizing global information perception and interaction. In contrast, the Transformer includes a multi-head self-attention mechanism, which can extract global information features. Consequently, several studies [14] have proposed models based on Transformers. We select representative AMR models based on CNN [2], LSTM [3], CLDNN [13], and Transformer [15] as target.

B. Data Augmentation Method

Data augmentation is a method of artificially generating new data from existing data and thus increasing the amount of data, including adding different perturbations to the data or using neural networks to learn the original data distribution and thus generate new data.

The basic data augmentation methods mainly involve transforming the data. For images, the most common way is to change the shape and pixels of the image [16], such as rotating, cropping, masking, etc. For radio signals, these types of methods are also very effective. The work in [17] investigated the effects of rotation, flipping, and Gaussian noise augmentation methods on LSTM-based AMR models. It was found that such methods of generating new data from simple transformations provide a degree of defense against various types of damage. These methods are uncomplicated and user-friendly, but they typically do not significantly modify the original dataset and have a limited impact on the generalization of the augmented model, especially in terms of robustness against attacks.

Mixed sample data augmentation (MSDA) is also a popular basic augmentation method. Mixup [18] is the first proposed

mixed sample data augmentation method which linearly interpolates and mixes two different samples to generate new training samples and corresponding labels. Based on Mixup, CutMix [19] swaps the patches of an image thus mixing the image, and the real labels are also mixed in proportion to the patches. To obtain a more appropriate shape for the patches, Fmix [20] obtains a randomized binary mask by applying a threshold to the low-frequency images sampled from Fourier space. These various simple yet effective methods have been proposed to enhance the training set by mixing images with other images or their variants with different sample mixing and label calculation approaches. By generating data samples that incorporate different information, the distribution of a given training set is extended, which can enhance the generalization performance of the model. Additionally, the decision boundary of the model can be made smoother through interpolation. And it may improve the model's robustness against adversarial attacks to some extent. The work of [21] proposes a data augmentation method for generating samples by mixing multiple signals, effectively improving the classification accuracy in some cases. Still, the technique is only applicable to less noisy samples, and it can be a hindrance at low signal-to-noise ratios.

Advanced data augmentation mainly refers to ways of generating new data using relatively complex networks, such as using generative adversarial networks (GAN) to learn the latent distributions of the original data, thus producing synthetic data similar to the original data. Inspired by the reconstructive and generative capabilities of GAN, many studies have utilized GAN for data augmentation for signal modulation classification [22]–[25]. Tang *et al.* [22] proposed data augmentation using GAN after converting signals into constellation diagrams, but the method requires preprocessing of the signals. In [24], a method of generating augmentation using GAN directly on time domain signals was proposed, obtaining a better enhancement in classification results. However, approaches based on GAN often display a bias towards the already skewed dataset, resulting in a less diverse generated data distribution compared to the training distribution [26]. Moreover, these models require a considerable amount of training data to achieve satisfactory results, which may restrict their efficacy when confronted with inadequate data.

C. Adversarial Training

The results of recent studies indicate that adversarial training is one of the most effective methods for defending against adversarial attacks. Furthermore, the application of this technique has been shown to result in the development of a robust model with relatively interpretable gradients [10]. Adversarial training adds generated adversarial samples to the original dataset and uses the adversarial samples to train the model thereby improving the recognition accuracy of the model. Commonly used methods for generating adversarial samples include the fast gradient sign method (FGSM) [27], project gradient descent (PGD) [10], etc. Szegedy *et al.* [7] introduce adversarial samples into the model and modify their labels to make it more robust in the face of adversarial samples. Goodfellow *et al.* [27] initially proposed adversarial training, in which adversarial samples generated by the FGSM algorithm are used along with clean samples to train the network, thereby enhancing its robustness. In [28], the highest accuracy can be obtained by employing the PGD algorithm to generate adversarial samples for training, but the computational cost of this method is high.

Adversarial training can be viewed as a special type of advanced data augmentation by adding generated adversarial samples to the original dataset. However, adversarial training has its limitations. Firstly, the overhead of adversarial training increases dramatically with the number of samples and the complexity of the attack method. It is impractical to include all unknown attack samples in adversarial training. Secondly, adversarial training can lead to a decrease in the model's recognition accuracy on the original data, especially for larger models, which means that the model needs to balance generalization with adversarial robustness.

D. Mixed and Adversarial Data Augmentation

To improve the generalization and robustness of the model at the same time and solve the limitations of adversarial training, various approaches have been proposed that combine adversarial training with data mixing. In the domain of images, Lamb *et al*. [12] introduced the interpolated adversarial training (IAT), which trains on interpolations of both adversarial and original samples to enhance the robustness of adversarial samples while maintaining a slight decrease in classification accuracy for original samples. Alfred *et al*. [29] proposed mixup-targeted labeling adversarial training (M-TLAT), a data augmentation strategy that combines mixup and targeted labeling adversarial training (TLAT) to improve the robustness against nineteen common damages and five adversarial attacks on images without compromising the accuracy of the original data. Liu *et al*. [30] presented AugRmixAT, a data processing and training approach that incorporates AugMix [31] and adversarial training followed by random data mixing to enhance robustness against common interference destruction, white box attacks, and black box attacks. In the realm of text, Si *et al*. [32] proposed an adversarial defense method called adversarial and mixup data augmentation (AMDA), which expands the sample space through interpolation to bring post-adversarial trained samples closer to the distribution of original data. This alleviates performance degradation issues observed on original samples after adversarial training while providing a solution for enhancing model robustness.

III. The Proposed Method

This paper proposes a method for augmenting radio signal data in order to address the requirement of signal data augmentation for the AMR model based on deep learning and to mitigate the potential threat of adversarial attacks. Firstly, defensiveness against adversarial attacks is achieved by adding adversarial samples generated using PGD to the dataset. And mixed labels are generated corresponding according to the perturbation. To mitigate the decrease in generalization

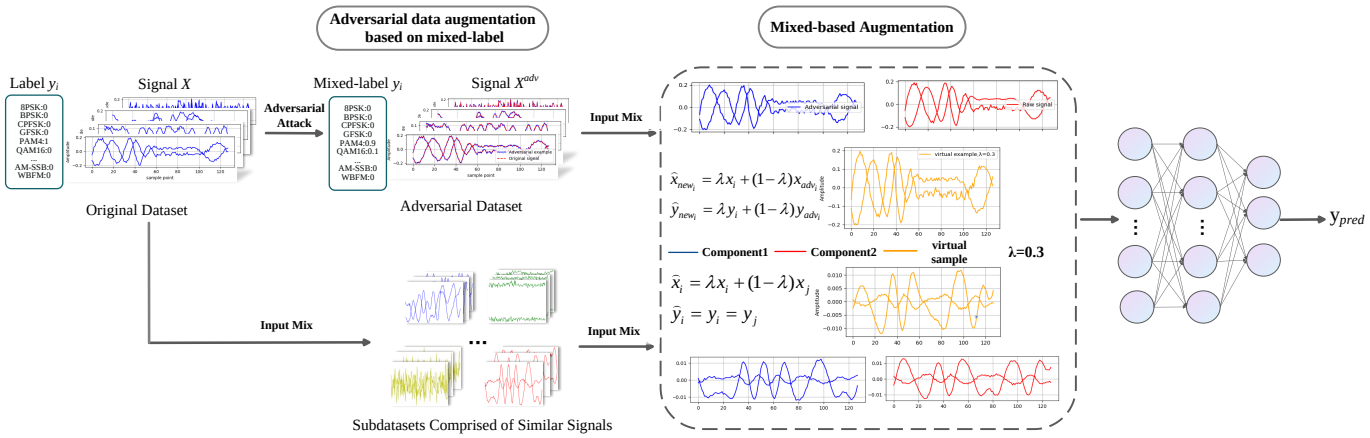


Fig. 2. Overall framework of AMSA model.

performance caused by the distribution difference between adversarial and original samples, interpolation mixing of adversarial and original sample pairs is employed to adjust the sample distribution. Additionally, to further enhance model performance, mixing augmentation is conducted within the same class signals, and the augmented data is jointly trained with the original data. The overall framework of the method is illustrated in Fig. 2.

A. AMR

Symbol sequences $s_0(t)$ are generally not suitable for direct transmission in the channel, so it is necessary to modulate them into signals suitable for channel transmission. Assuming that the modulation function is $M(\cdot)$, the signal $x(t)$ received by the receiving end is represented as:

$$x(t) = h(t)M(s_0(t)) + w, \tag{1}$$

where $h(t)$ represents the influence of the channel on signal transmission, and w represents additive white Gaussian noise (AWGN). The receiver needs to recognize the type of $M(\cdot)$ for demodulation processing, to recover $s_0(t)$. AMR refers to the automated process of identifying $M(\cdot)$.

The signal $x(t)$ is typically composed of in-phase (I) and quadrature (Q) components, and can thus be represented in complex form as:

$$x(t) = x_I(t) + jx_Q(t). \tag{2}$$

After sampling and preprocessing at the receiver end, the discrete signal $x(n) \in R^{2\times l}$ along with its corresponding label y are obtained as inputs to the AMR model, where l represents the number of sample points. Different AMR models extract features from the input time-domain signal to predict the modulation type of $x(t)$. AMSA is an online data augmentation operation performed on the input signals after feeding into the AMR model.

B. Adversarial Augmentation based on Mixed-Label

To defend against unknown adversarial attacks, the model needs to learn the features of the adversarial sample. Given the original training data set $X = \{x_1, x_2, \cdots, x_m\}$ corresponding to label $Y = \{y_1, y_2, \cdots, y_m\}$, each of the signal samples $x \in \mathbb{R}^{2\times l}$ can be iterated to produce corresponding adversarial sample x_{adv}, and m is the number of samples. In the AMSA, PGD, the strongest first-order attack method, is used to generate adversarial samples. By learning the adversarial samples found by PGD, the neural network will have the defensiveness to resist other first-order adversarial attacks.

Adversarial samples are generated iteratively, which can be formulated as:

$$\begin{aligned} &x_{adv}^0 = x + \eta, \\ &x_{adv}^{i+1} = Clip_{x,\varepsilon}\{x_{adv}^i + \alpha sign[\nabla_X J(\theta, x_{adv}^i, y)]\}, \\ &i = 1, 2, \cdots, np\_iter, \end{aligned} \tag{3}$$

where x_{adv}^i represents the adversarial sample obtained after the ith iteration. η represents an initial random perturbation, and $Clip_{x,\varepsilon}\{.\}$ restricts the value to a certain neighborhood, with ε denoting the maximum perturbation. α controls the magnitude of the perturbation, and $sign(\cdot)$ represents the sign function. J signifies the loss function, θ denotes the network parameters, and np_iter represents the number of iterations. Then the adversarial data set $X_{adv} = \{x_{adv_1}, x_{adv_2}, \cdots, x_{adv_m}\}$ is obtained.

For the labels of adversarial samples, since the adversarial samples are created by adding an imperceptible perturbation to the original samples, which doesn't change their modulation type, the true label y_{adv_true} should be the same as the original samples.

$$y_{adv\_true} = y \tag{4}$$

However, the objective of adversarial samples is to deceive the model into classifying them as a different category from the true label. They are distributed relatively far from the center of their true category and close to the decision boundary, so they should be assigned lower predictive accuracies, thus necessitating smoother labels. There exists a target label y_{adv_pred} predicted by the model. We propose representing the label of adversarial samples as a mixture of the true label and the predicted label, denoted as:

$$y_{adv} = (1-\varepsilon)y_{adv\_true} + \varepsilon y_{adv\_pred} \tag{5}$$

By giving better-supervised signals to the adversarial samples through mixed labels, the model establishes a relationship between perturbations and adversarial labels, forcing

the model to make more accurate predictions on adversarial samples during training. This enhances the model's robustness against adversarial samples.

C. Mixed-based Signal Augmentation

To bolster the efficacy of adversarial training and increase the diversity of samples, thereby facilitating a more comprehensive understanding of the data distribution, signal mixing is employed on the dataset augmented with adversarial samples. This mixing process encompasses two primary components: firstly, mixing original signals to enhance model generalization, and secondly, conducting interpolation mixing between original signals and adversarial samples to amplify the effectiveness of adversarial training.

In the field of CV, the classical data augmentation method Mixup generates diverse virtual samples by linearly interpolating between randomly selected sample pairs $\{(x_i, y_i), (x_j, y_j)\}$, thereby blending information from different classes. Images often exhibit similarity and comparability during the mixing process, this is not the case for signals of different modulation types. The characteristics of such signals, which may reside in amplitude, frequency, or phase, generally lack comparability. Signal modulation types can be broadly categorized as analog modulation and digital modulation, encompassing various types such as amplitude modulation, frequency modulation, and phase modulation. Directly superimposing samples of different categories may lead to mixed signals that no longer conform to the original modulation type characteristics. This does not comply with the basic assumptions of the data augmentation. Hence, the AMSA method employs the mixing of similar signals. The process of mixing pairs $\{(x_i, y_i), (x_j, y_j)\}$ of the same label is as follows:

$$\begin{aligned} x_i^* &= \lambda x_i + (1-\lambda) x_j, \\ y_i^* &= y_i = y_j, \end{aligned} \tag{6}$$

where λ follows a Beta distribution $\lambda \sim Beta(\alpha, \alpha)$, and α is a hyperparameter.

For a given signal sample (x_i, y_i) paired with its corresponding adversarial sample (x_{adv_i}, y_{adv_i}), the process of interpolating between them to obtain a new sample $(x^*_{new_i}, y^*_{new_i})$ is represented as:

$$\begin{aligned} x^*_{new_i} &= \lambda x_i + (1-\lambda) x_{adv_i}, \\ y^*_{new_i} &= \lambda y_i + (1-\lambda) y_{adv_i}. \end{aligned} \tag{7}$$

The method defends against adversarial attacks by adding adversarial samples to the model. It leverages data mixing to create a multitude of neighboring samples based on the training data, thereby adjusting the distribution of data after adversarial augmentation. The order of data mixing after adversarial training can reduce the overhead of adversarial training compared with the reverse order or direct adversarial training on a large number of samples.

D. Theoretical Analysis

Adversarial training, one of the most effective adversarial defense methods known, improves the robustness of the model itself by training it on samples. This active defense approach can achieve robustness without detecting the presence of adversarial samples. However, it will make the model have different degrees of performance degradation on clean samples.

According to the deep learning manifold assumption, high-dimensional data is formed by low-dimensional manifold structures embedded in high-dimensional space, and the manifolds can portray the nature of the data. Deep learning models learn a decision boundary plane by training on samples, which can classify low-dimensional data manifolds but may misclassify points near the manifold. Adversarial samples are points that happen to be distributed near the manifold, which is one of the dominant explanations for the genesis of adversarial samples. From the perspective of data manifolds, since the adversarial training produces a large number of adversarial samples concentrated in the decision boundary, and the distribution of the adversarial samples is significantly different from that of the clean samples, it leads to a degradation of the model's performance on the clean samples. From the perspective of feature learning, after adding adversarial samples, the model may overfit the adversarial samples and ignore the features of the clean samples, which leads to a degradation of the generalization performance.

AMSA can optimize the performance of adversarial training using data mixing from both perspectives mentioned above. By interpolating between the adversarial samples and the original samples, on the one hand, it is possible to add more diverse training data to learn smoother decision boundaries; on the other hand, mixing generates new data with the characteristics of the adversarial samples and the characteristics of the clean samples, which balances the performance of the model on different samples.

In addition to this, Mixup generates a large number of samples that are beneficial to the model's generalizability and adversarial robustness. Due to the presence of a complex electromagnetic environment in signal transmission, signal mixing can better simulate signal variations in real-world application scenarios. Diverse data helps to reduce model overfitting and to fill in the blank areas in the training set, allowing the model to learn and predict better in these areas. Moreover, studies have shown that adversarial training requires a higher amount of data, and increasing the training data can obtain a significant robustness improvement. This is the reason we believe AMSA can be effective.

IV. Experiments and Discussions

In this section, we evaluate the generalization and robustness of AMSA in scenarios with different numbers of samples. The main purpose is to prove the following conclusions:

- Compared with traditional augmentation methods, AMSA greatly improves the adversarial robustness of the model.
- AMSA effectively alleviates the degradation of generalization performance caused by adversarial training.
- AMSA achieves its goal in scenarios with both sufficient and insufficient samples.

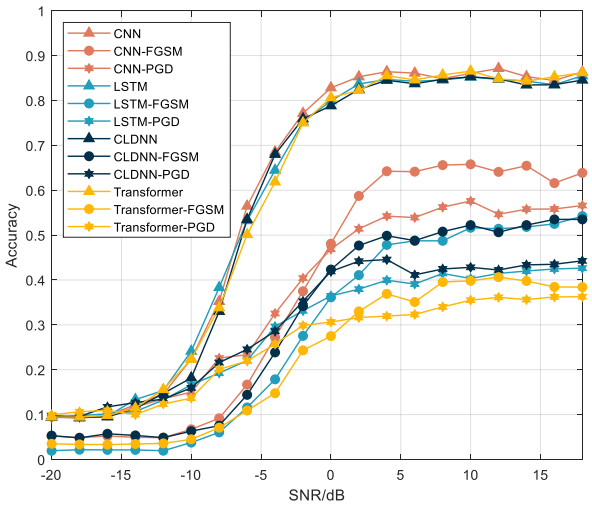


Fig. 3. Recognition accuracy of the target model under FGSM and PGD attacks.

TABLE I
RECOGNITION RATE OF DIFFERENT MODELS USING DIFFERENT DATA AUGMENTATION METHODS UNDER FGSM ATTACK (%).

Model	SNR (dB)	Baseline	+Flip	+Rotation	+AMSA
CNN	-10	6.73	**10.73**	9.27	10.55
	0	48.09	62.05	65.73	**72.64**
	10	65.77	75.55	76.45	**81.50**
	18	63.86	74.36	74.91	**81.77**
LSTM	-10	3.77	10.41	8.88	**12.00**
	0	36.14	67.86	71.95	**81.14**
	10	51.64	76.59	78.91	**84.32**
	18	54.23	75.77	77.14	**85.23**
CLDNN	-10	6.32	7.64	7.73	**13.05**
	0	42.32	55.68	60.86	**71.64**
	10	52.23	63.91	67.32	**79.36**
	18	53.50	66.00	70.55	**79.50**
Transformer	-10	4.50	8.68	8.55	**12.64**
	0	27.50	50.27	51.50	**68.18**
	10	39.82	60.73	70.32	**80.64**
	18	38.41	61.50	71.36	**80.23**

A. Dataset, Scenarios, Target Models, and Attack Methods

1) Dataset: We evaluate signal augmentation methods with the public dataset RML2016.10a [4] and RML2018.01a [1].

The RML2016.10a dataset comprises 11 modulation styles, including 8PSK, BPSK, CPFSK, GFSK, PAM4, AM-DSB, AM-SSB, 16QAM, 64QAM, QPSK, and WBFM, with SNR ranging from -20 to 18 dB. The dataset comprises 220,000 signal samples, with 1,000 samples at each SNR for each modulation type. Each sample has IQ channels, consisting of 128 points.

The RML2018.01a comprises 24 distinct modulation types, including OOK, 4ASK, 8ASK, BPSK, QPSK, 8PSK, 16PSK, 32PSK, 16APSK, 32APSK, 64APSK, 128APSK, 16QAM, 32QAM, 64QAM, 128QAM, 256QAM, AM-SSB- WC, AM-SSB-SC, AM-DSB-WC, AM-DSB-SC, FM, GMASK, and OQPSK. The SNR ranges from $-20:2:30$ dB, with a total of 2,555,904 samples and 1024 sampling points each.

2) Scenarios: We set up three scenarios with different numbers of samples to verify our method.

- Scenario I: We divide the dataset into training set, validation set, and test set following a 6:2:2 ratio. The training set comprises 132000 samples, while the validation and test sets each contain 44000 samples.
- Scenario II: To simulate the scenario of limited training data, we extract 10% of samples from each class in the complete dataset. This yields 100 samples for each modulation type at every SNR level. The dataset is then divided into training, validation, and test sets following a ratio of 1:1:2. Specifically, there are 22000 training samples, 22000 validation samples, and 44000 test samples.
- Scenario III: We extract 100 samples from each class of modulation styles at each SNR in the RML2018.01a dataset as the training set, and the ratio of training, validation, and test sets is 1:1:1, with 62,400 samples respectively.

3) Target models and attack methods: We employ the AMR models based on CNN [2], LSTM [3], CLDNN [13], and transformer [15]. Adversarial attacks employ FGSM and PGD, where FGSM is a simple yet effective first-order attack method, and PGD has been proven to be the strongest first-order attack method. In Scenario I, the recognition accuracy of the target model under FGSM and PGD attacks is shown in Fig. 3. It can be seen that the different AMR models have a large loss of accuracy under adversarial attacks. So it is necessary to take certain defensive means.

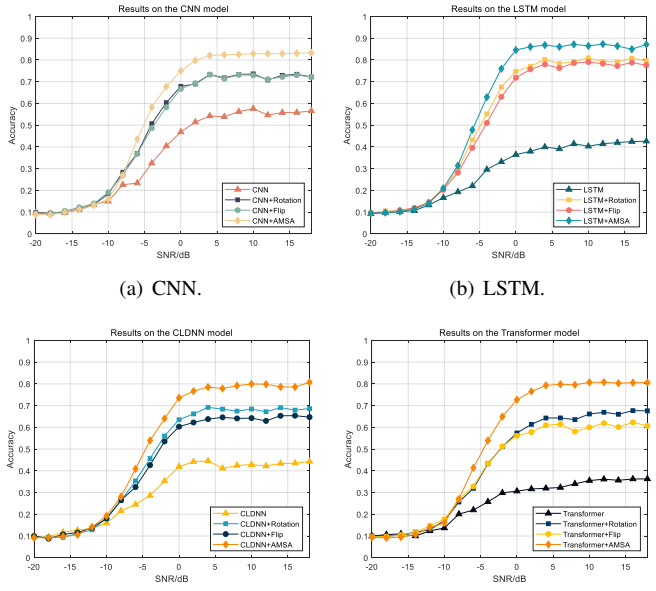
All experiments use the TensorFlow backend on a single NVIDIA Corporation GPU.

B. Comparison of Defense Ability of Data Augmentation Methods

To verify the effectiveness of the proposed method in improving the model's adversarial robustness, models are subjected to FGSM attack and PGD attack (white-box attack) with an attack coefficient $\alpha = 0.01$. We compare AMSA with traditional data augmentation methods, including rotation and flip, which are proven to be effective methods [17]. We test the recognition accuracy of the model after the attack, and the recognition results on the full RML2016.10a dataset are shown in Table I, with the best defense results bolded.

Due to increased sample diversity, both flip and rotation augmentation methods exhibit certain defensive capabilities against FGSM attacks. Although the adversarial training component in the AMSA method employs the PGD approach, the adversarial samples generated by PGD possess transferability, thereby demonstrating good defensive performance against the FGSM attack. As shown in Table I, AMSA significantly enhances the model's robustness under low and high SNR conditions, with respective improvements of 24.55%, 45.00%, 29.32%, and 40.68% compared to the original model at 0 dB. Moreover, compared to the better one of flip and rotation, AMSA shows improvements of 6.91%, 9.19%, 10.78%, and 16.68%.

The comparison of model recognition accuracy under PGD attack is shown in Fig. 4. The performance of rotation and



(a) CNN. (b) LSTM.

(c) CLDNN. (d) Transformer.

Fig. 4. Recognition accuracy of different models varies with SNR under PGD attack.

flip is similar, while AMSA has obvious advantages when the SNR surpasses -10 dB, culminating in its highest accuracy at 4 dB.

C. Comparison of Generalization Ability of Data Augmentation Methods

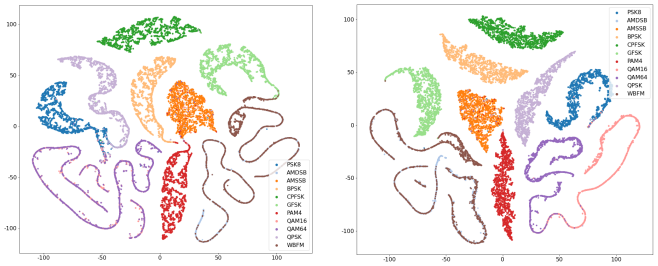
In Scenario I, to demonstrate the effectiveness of AMSA in mitigating the performance degradation caused by adversarial training, Table II presents the recognition results of clean samples using AMR models based on CNN, LSTM, CLDNN, and Transformer. (AT represents adversarial training.)

From the perspective of overall recognition accuracy, although the AMSA method applied to CNN and CLDNN models did not reach the baseline for clean samples, it showed significant improvement compared to adversarial training and was relatively close to the benchmark. For LSTM-based models, the AMSA method yielded remarkable results, surpassing recognition accuracy without augmentation. In Transformer models, AMSA's performance was slightly better than the baseline.Regarding specific SNR, AMSA demonstrated pronounced benefits when SNR was around 0 dB, outperforming other methods across various models. Under lower SNRs, the differences in recognition accuracy among methods were minimal, with non-augmented methods performing relatively better. Under high SNRs, AMSA enhanced the recognition accuracy over adversarial training for CNN and CLDNN models, while for LSTM and Transformer models, it surpassed the baseline.

To further investigate the enhancement brought by AMSA, we employed T-SNE to visualize the feature distributions of the model before and after the application of AMSA, as depicted in Fig. 5. It is evident that after the augmentation with AMSA, the data distribution within the same class became more clustered, and the classification boundaries between

TABLE II
RECOGNITION RATE OF DIFFERENT MODELS USING ADVERSARIAL TRAINING AND AMSA ON CLEAN SAMPLES (%).

SNR (dB)	-18	-10	0	10	18	Overall accuracy
CNN	**9.18**	**22.41**	82.82	**86.05**	**86.36**	**58.49**
+AT	9.09	20.14	82.50	84.86	84.95	56.86
+AMSA	9.14	18.45	**83.36**	85.55	85.55	57.95
LSTM	9.14	**24.14**	79.91	85.22	85.50	57.65
+AT	9.09	23.55	83.14	84.91	84.41	57.44
+AMSA	**9.36**	23.00	**90.45**	**92.45**	**92.64**	**62.72**
CLDNN	**9.41**	18.18	78.82	**85.27**	**84.55**	**56.94**
+AT	9.14	16.18	79.86	83.55	83.73	55.56
+AMSA	9.09	**22.55**	**79.91**	83.64	84.36	56.39
Transformer	**9.50**	**22.64**	78.91	83.45	83.95	56.52
+AT	9.09	20.50	80.77	83.32	82.95	55.92
+AMSA	9.14	21.59	**81.00**	**84.09**	**84.27**	**56.53**



(a) model without augmentation. (b) model with AMSA.

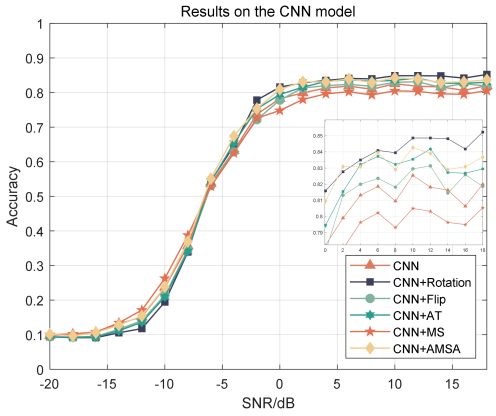
Fig. 5. T-SNE result of distribution calculated by AMR model based on LSTM. (a) Distribution calculated by the model without augmentation. (b) Distribution calculated by the model with AMSA.

different classes became clearer. Additionally, AMSA helped alleviate the confusion between 16QAM and 64QAM.

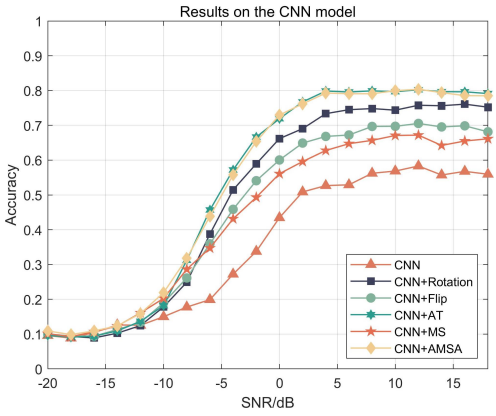
D. Performance Comparison on Small Datasets

To verify the enhancement effect and defense performance of the algorithm when the sample quantity is insufficient, 10% of samples from the RML2016.10a are randomly selected for data augmentation. The results of the AMR model based on CNN are shown in Fig. 6., with similar outcomes observed on AMR models based on LSTM, CLDNN, and Transformer. We also conducted experiments on the extracted RML2018.01a dataset, and the results are shown in Fig. 7. AT, MS represents adversarial training, mixing signals, respectively.

With fewer training samples, mixing signal improves recognition accuracy at low SNRs but yields adverse effects at high SNRs. Employing adversarial training has no advantage at low SNRs, but at high SNRs, instead of degrading the accuracy on clean samples, it improves adversarial robustness. AMSA, as a combination of the two, shows superior performance at all SNRs. It not only maintains high robustness from adversarial training but also achieves a recognition rate on clean samples that is superior to adversarial training and mirror augmentation, approaching that of rotation augmentation. It can be observed that AMSA demonstrates clear advantages in scenarios with limited sample sizes.

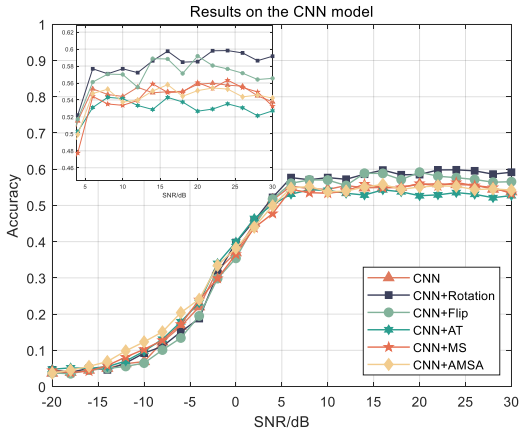


(a) Recognition rate on clean samples.

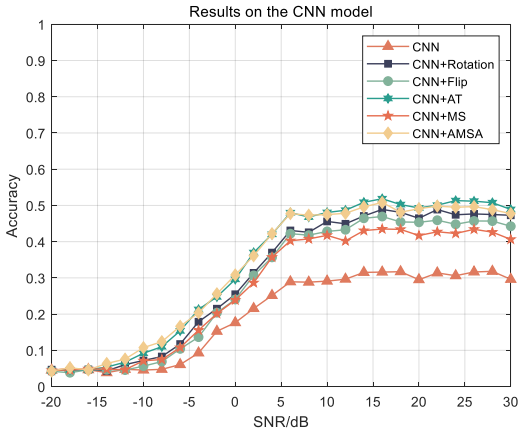


(b) Recognition rate on adversarial samples.

Fig. 6. Recognition accuracy of the AMR model based on CNN varies with SNR in Scenario II.



(a) Recognition rate on clean samples.



(b) Recognition rate on adversarial samples.

Fig. 7. Recognition accuracy of the AMR model based on CNN varies with SNR in Scenario III.

E. Influence of Sample Interpolation Mixing and Mixed Labels on Adversarial Training Performance

The AMSA method has introduced certain improvements to adversarial training by using mixed labels and interpolation of adversarial samples.

To observe the impact on the performance of adversarial training in Scenario II, we employed adversarial training, adversarial training with one-time interpolation mixing of adversarial samples with original signals, and adversarial training with two-time interpolation mixing (including the use of genuine labels and mixed labels). We conducted recognition experiments on clean samples and adversarial samples, and the overall accuracy of the CNN-based AMR model is presented in Table III.

Based on the table, it can be observed that mixing adversarial samples with original samples can simultaneously enhance the robustness and generalization of adversarial training. In addition, the effectiveness of interpolating twice surpasses that of interpolating once, as multiple interpolations can cover a larger space and learn more distribution characteristics. However, excessive interpolation can also lead to increased computational costs, hence the approach adopted in this paper is a two-time interpolation. The use of mixed labels tends to elevate the recognition accuracy of the model on clean samples, albeit sometimes at the expense of adversarial robustness, although such losses are typically minimal.

F. Determination of the Proportion of Data from Adversarial Augmentation

In the aforementioned experiments, all samples augmented by mixing samples of the same types are added to the dataset, with the proportion of virtual samples generated by adversarial augmentation primarily determined by the adversarial samples. In the adversarial training phase, experimental results indicate that an excessive or insufficient proportion of adversarial samples has detrimental effects. When adversarial samples are scarce, it becomes challenging for the model to learn sufficient features, leading to inadequate adversarial robustness. However, an excessive number of adversarial samples not only fails to enhance robustness but also induces overfitting, resulting in a simultaneous decline in robustness and generalization.

Taking the CNN-based AMR model as an example, Table IV illustrates the overall recognition accuracy obtained with different proportions of adversarial augmentation. In Scenario I, the number of adversarial samples added to the dataset per round is randomly determined to be between 50% and 75% of the original data quantity, while in Scenario II, this proportion ranges from 75% to 100%, which is determined based on a better balance of robustness and generalization.

TABLE III
THE OVERALL RECOGNITION ACCURACY OF THE MODEL UNDER DIFFERENT ADVERSARIAL TRAINING METHODS (%).

	AT	One-time interpolation mixing		Two-time interpolation mixing	
		True label	Mixed label	True label	Mixed label
Ori.	56.49	56.78	57.01	56.91	57.45
Adv.	52.90	52.92	53.00	53.16	53.10

TABLE IV
OVERALL RECOGNITION ACCURACY OF MODELS AFTER ADVERSARIAL TRAINING WITH DIFFERENT ADVERSARIAL AUGMENTATION PROPORTION (%).

	Scenarios I		Scenarios II	
	Ori.	Adv.	Ori.	Adv.
0%	58.49	36.55	55.98	35.38
0%-25%	58.00	51.29	56.38	48.00
25%-50%	57.60	53.30	56.39	51.46
50%-75%	**57.22**	**53.92**	56.33	52.25
75%-100%	56.84	54.01	**56.49**	**52.90**
100%	56.86	53.83	56.08	52.52

V. CONCLUSION

Aiming at the possible adversarial attacks faced by deep learning-based AMR models, this paper introduces a radio signal augmentation method AMSA for improving the adversarial robustness and generalization of the models. The method combines adversarial training with signal mixing. It makes the model adversarial robust by adding adversarial samples and uses data mixing to adjust the data distribution and generate diverse data. The results demonstrate that AMSA can address the lack of improvement of models' adversarial robustness through traditional signal augmentation methods and alleviate the degradation of the model's performance on clean data due to adversarial training. We believe it can make neural networks used in real-world applications more reliable and secure. However, this paper only considers scenarios involving gradient-based white-box attacks. Exploring how to maintain stable defense performance against various attacks is an important research direction for defending against adversarial attacks.

REFERENCES

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

[2] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Proc. EANN*, 2016.

[3] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deeplearning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 3, pp. 433–445, 2018.

[4] T. J. O'Shea and N. West, "Radio machine learning dataset generation with gnu radio," in *Proc. GNU Radio Conf.*, 2016.

[5] K. Tekbıyık, A. R. Ekti, A. Görçin, G. K. Kurt, and C. Keçeci, "Robust and fast automatic modulation classification with CNN under multipath fading channels," in *Proc. IEEE VTC*, 2020.

[6] F. Zhang, C. Luo, J. Xu, Y. Luo, and F.-C. Zheng, "Deep learning based automatic modulation recognition: Models, datasets, and challenges," *Digit. Signal Process.*, vol. 129, p. 103650, 2022.

[7] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. ICLR*, 2014.

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

[9] Y. Shi, T. Erpek, Y. E. Sagduyu, and J. Li, "Spectrum data poisoning with adversarial deep learning," in *Proc. IEEE MILCOM*, 2018.

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2018.

[11] H. Zhang *et al.*, "Theoretically principled trade-off between robustness and accuracy," in *Proc. ICML*, 2019.

[12] A. Lamb *et al.*, "Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy," *Neural Netw.*, vol. 154, pp. 218–233, 2022.

[13] N. E. West and T. O'Shea, "Deep architectures for modulation recognition," in *Proc. IEEE DySPAN*, 2017.

[14] Y. Chen, B. Dong, C. Liu, W. Xiong, and S. Li, "Abandon locality: Frame-wise embedding aided transformer for automatic modulation recognition," *IEEE Commun. Lett.*, vol. 27, no. 1, pp. 327–331, 2023.

[15] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[16] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[17] L. Huang *et al.*, "Dataaugmentation for deep learning-based radio modulation classification," *IEEE Access*, vol. 8, pp. 1498–1506, 2020.

[18] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.

[19] S. Yun *et al.*, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF ICCV*, 2019.

[20] E. Harris *et al.*, "FMix: Enhancing mixed sample data augmentation," 2020, *arXiv:2002.12047*.

[21] X. Xu *et al.*, "Mixing signals: Data augmentation approach for deep learning based modulation recognition," 2022, *arXiv:2204.03737*.

[22] B. Tang, Y. Tu, Z. Zhang, and Y. Lin, "Digital signal modulation classification with data augmentation using generative adversarial nets in cognitive radio networks," *IEEE Access*, vol. 6, pp. 15713–15722, 2018.

[23] X. Yao, H. Yang, and Y. Li, "Modulation identification of underwater acoustic communications signals based on generative adversarial networks," in *Proc. OCEANS*, 2019.

[24] Z. Tang *et al.*, "Data augmentation for signal modulation classification using generative adverse network," in *Proc. IEEE ICEICT*, 2021.

[25] H. Zhou *et al.*, "Few-shot electromagnetic signal classification: A data union augmentation method," *Chin. J. Aeronaut.*, vol. 35, no. 9, pp. 49–57, 2022.

[26] I. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. NIPS*, 2014.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, 2015.

[28] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. ICLR*, 2016.

[29] A. Laugros, A. Caplier, and M. Ospici, "Addressing neural network robustness with mixup and targeted labeling adversarial training," in *Proc. ECCV Workshops*, 2020.

[30] X. Liu, F. Shen, J. Zhao, and C. Nie, "AugRmixAT: A data processing and training method for improving multiple robustness and generalization performance," in *Proc. IEEE ICME*, 2022.

[31] D. Hendrycks *et al.*, "AugMix: A simple data processing method to improve robustness and uncertainty," in *Proc. ICLR*, 2020.

[32] C. Si *et al.*, "Better robustness by more coverage: Adversarial training with mixup augmentation for robust fine-tuning," in *Proc. ACL-IJCNLP*, 2021.

Li Zhang is currently a Ph.D. student at the Information Engineering University. She received her B.E. degree from Information Engineering University, Zhengzhou, China in 2022. Scientific interests: intelligent signal processing, communication security on adversarial attack and defense, automatic modulation recognition, and artificial intelligence.

Gang Zhou Dr. at the Information Engineering University. He received his B.E. and M.A. Eng degrees from Information Engineering University in 1996 and 1999, Ph.D. degree from Beihang University in 2007. He was with the State key of the Laboratory of Mathematical Engineering and Advanced Computing, Information Engineering University, as a research fellow. His research interests are big data, network security, and artificial intelligence.

Gangyin Sun received the B.E. degree from Information Engineering University, Zhengzhou, China in 2022. He is currently pursuing the M.A. degree at the Information Engineering University. His research interests include radar signal processing and machine learning for radios.

Chaopeng Wu received the B.E. degree in Communication Engineering from Henan Polytechnic University in 2022 and is currently pursuing the M.A. degree in the Department of Data and Target Engineering at the Information Engi-neering University. His research interests include radar signal processing and machine learning for radios.