Adversarial Robust Modulation Recognition Guided by Attention Mechanisms

Quanhai Zhan, Xiongwei Zhang, Meng Sun*, Lei Song, and Zhenji Zhou

Abstract Deep neural networks have demonstrated considerable effectiveness in recognizing complex communications signals through their applications in the tasks of automatic modulation recognition. However, the resilience of these networks is undermined by the introduction of carefully designed adversarial examples that compromise the reliability of the decision processes. In order to address this issue, an Attention-Guided Automatic Modulation Recognition (AG-AMR) method is proposed in this paper. The method introduces an optimized attention mechanism within the Transformer framework, where signal features are extracted and filtered based on the weights of the attention module during the training process, which makes the model to focus on key features for the task. Furthermore, by removing features of low importance where adversarial perturbations may appear, the proposed method mitigates the negative impacts of adversarial perturbations on modulation classification, thereby it improves both accuracy and robustness. Experimental results on benchmark datasets show that AG-AMR obtains a high level of accuracy on modulation recognition and exhibits significant robustness. Furthermore, when working together with adversarial training, it is shown that AG-AMR effectively resists several existing adversarial attacks, which thus further validates its effectiveness on defending against adversarial sample attacks.

Index Terms—Adversarial Attacks, Adversarial Training, Attention Mechanisms, Automatic Modulation Recognition

I. Introduction[1]

The rapid advancement of wireless communication technologies has led to the widespread application of automatic modulation recognition (AMR) in areas such as cognitive radio [1] and electronic countermeasures [2]. However, the emergence of increasingly diverse modulation types and signal variants in modern communication systems has posed challenges to the high performance of AMR systems. Traditional AMR methods are generally divided into two main categories: likelihood-based methods [3] and feature-based ones [4]. Likelihood-based methods perform recognition by computing the maximum likelihood over all modulation categories given the input signals, which typically requires ideal channel conditions and incurs high computational cost. Feature-based methods involve manual construction of feature representations, which can achieve suboptimal recognition results with lower complexity but often rely on extensive manual experience in feature extraction. In contrast, AMR methods based on machine learning automatically extract features from signals, reducing the dependence on manual expertise and achieving better recognition performance at relatively low costs. In recent years, AMR methods based on deep learning have been extensively explored [5-9]. Due to their superior performance and effectiveness, deep learning methods have increasingly become the dominant technology in wireless signal recognition tasks.

However, deep neural networks have been shown to be vulnerable to adversarial attacks, where deep learning models produce incorrect decisions when small perturbations are applied to the input data [10, 11]. A great deal of work has been dedicated to studying adversarial attacks on modulation recognition tasks in wireless signals. In these studies, fine-tuned perturbations are added to the original signals as additive noises, causing the model to misclassify the inputs. To address the challenges posed by adversarial attacks, various defense strategies based on detection [12, 13] and mitigation [14, 15] have been proposed. Detection-based methods primarily use statistical techniques to identify and prevent adversarial examples. Mitigation-based methods are essentially adversarial training, which retrain deep learning models on the data augmented with adversarial examples to improve the models' resilience to adversarial attacks. In general, these methods are important for defending against adversarial attacks, ensuring the stability and reliability of deep learning-based AMR models in practical applications.

Compared to the detection methods, adversarial training can correctly classify adversarial examples into their true categories, rather than detecting and discarding them. However, adversarial training methods rely on a large number of examples during the training phase, which inevitably increases computational complexity when

Quanhai Zhan, Xiongwei Zhang, Meng Sun, Lei Song, Zhenji Zhou are with the Lab of Intelligent Information Processing, Army Engineering University, Nanjing, 210007, China. (Corresponding: Meng Sun, sunmeng@aeu.edu.cn.)

This work is supported by the Natural Science Foundation of China (62071484,62371469) and the Natural Science Foundation of Jiangsu Province for Excellent Young Scholars (BK20180080).

generating these examples and retraining the model. Additionally, retraining with adversarial examples has a potential impact on the model's performance on clean data, potentially reducing accuracy. More importantly, in complex models like Transformers, the large number of parameters can lead to overfitting as the adversarial training progresses. Therefore, there is a need for new algorithms that effectively balance accuracy, robustness, and computational efficiency in the presence of adversarial attacks.

In order to address the challenges above, an Attention-Guided Automatic Modulation Recognition (AG-AMR) method is proposed in this paper. This method integrates an optimized attention mechanism within the Transformer model, which extracts and filters signal features based on attention weights during training. With the attention mechanisms, the model achieves high recognition accuracy, reduces complexity, and improves robustness compared to traditional approaches. Furthermore, by integrating the proposed method with adversarial training, the model's resistance to adversarial perturbations is further enhanced, which effectively mitigates their negative impacts on classification performance. Consequently, the AG-AMR method makes a balance in accuracy, robustness, and computational efficiency for automatic modulation recognition under adversarial conditions. The main contributions of this paper can be summarized as follows.

1) An AG-AMR method based on an optimized attention mechanism is proposed. By weighting features with the attention module, the method enables the network structure to focus on critical information, thereby improves the recognition accuracy, reduces the model complexity, and enhances the adversarial robustness.
2) Extensive validation and comparative experiments are conducted on the benchmark datasets RML2016.10a and RML2018.01a. The results demonstrate that the AG-AMR method outperforms other typical modulation recognition methods in terms of both robustness and stability when processing adversarial examples.
3) Robust recognition in adversarial environments is achieved. By integrating adversarial training techniques with the model during training, the model's ability to resist adversarial perturbations is further enhanced, which ensures high stability and accuracy in the presence of adversarial examples.

The remaining sections of this paper are organized as follows: Section II provides a comprehensive overview of automatic modulation recognition and its adversarial defenses, as well as efficient training within Transformers. Section III presents the prior knowledge relevant to this study, including the AMR model, adversarial perturbations, and an introduction to adversarial training. Section IV describes the AG-AMR method and its underlying principles in details. Section V describes the experimental setup and results to validate the effectiveness of the AG-AMR. Finally, Section VI concludes the paper with a summary and outlines future research directions.

II. Related Works

In this section, existing deep learning-based modulation recognition methods are introduced firstly, followed by a discussion of existing adversarial defense techniques for deep learning-based AMR models. Finally, related works on efficient training methods are explored within the Transformer framework as a potential idea to make a balance between the accuracy and the robustness of the models.

A. Automatic Modulation Recognition

With the rapid advancement of wireless communication technologies, modulation recognition is becoming an important task of signal sensing to handle complex and diverse inputs. Deep learning, with its powerful abilities on pattern recognition, has become an effective solution compared to traditional methods. O'Shea et al. apply CNN networks into modulation recognition tasks, whose results demonstrate that CNNs are able to effectively extract and classify features of modulation signals [16]. Subsequently, West [17] has further explored the applications of deep learning architectures to modulation recognition, and Karra et al. [18] propose hierarchical deep neural networks for radio modulation recognition. The temporal nature of radio modulation signals has also been extensively explored as a critical feature by many researchers. Rajendran [19] have used Long Short-Term Memory (LSTM) networks to improve accuracy, which demonstrate the advantages of RNN-based modulation recognition methods in capturing temporal information. In addition, Xu and Zhang [20, 21] combine the spatial feature extraction capabilities of CNNs with the temporal analysis strengths of RNNs to provide a superior hybrid model for modulation recognition. With the advents of attention mechanisms and Transformer models, Chen and Cai [8, 22] effectively address long-range dependencies in signals using attention mechanisms.

The works above suggest that deep learning-based models exhibit excellent performance in modulation recognition tasks, which relies on statistical features learned from large datasets for decision making. However, advanced deep learning method have been shown vulnerable to adversarial attacks.

B. Adversarial Defense in Automatic Modulation Recognition

With the emerging of adversarial attacks, the secure design of deep learning methods for AMR must consider new issues, i.e., detecting and defending against adversarial attacks [23]. Modifications for the purpose of adversarial attack can significantly alter the data distribution of signals, which potentially affects the model's performance on modulation recognition. In order to improve the models' robustness, researchers have proposed various defense strategies. Detection-based methods have been explored by Zhang et al. [24], who implement the detection of adversarial

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2025.3526577

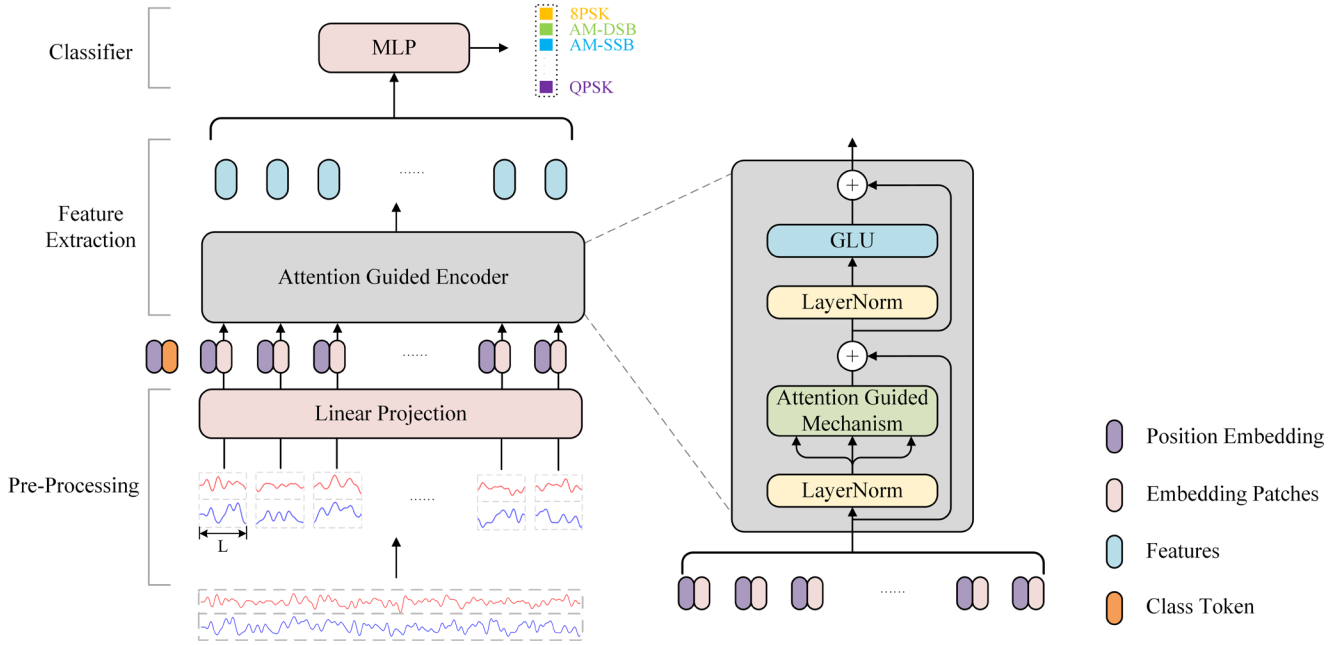

Fig. 1. The architecture of the AG-AMR

examples using neural repulsion techniques complemented by label smoothing and Gaussian noise injection. Similarly, Kokalj-Filipovic et al. [13] utilize the signal-to-noise ratio as a critical metric to detect adversarial examples through softmax outputs. Xu et al. [12] develop a multi-feature fusion technique by extracting Local Intrinsic Dimensionality (LID) and Constellation Diagram (CD) features for adversarial example detection. Mitigation-based methods aim to enhance the models' robustness against adversarial attacks. Chen and Tang et al. [25, 26] propose methods to mitigate adversarial attacks through adversarial distillation techniques. In addition, adversarial training, which introduces adversarial examples during the training process, is considered one of the most effective methods to improve robustness. Kim et al. [14] have used noise-enhanced training data based on random smoothing to improve the robustness of modulation classifiers against adversarial perturbations. Kokalj-Filipovic et al. [15] have implemented a defense method against adversarial attacks by pre-training the target classifier using autoencoders.

Despite these efforts, existing defense strategies in AMR either increase computational cost or have negative impacts on the performance of recognition. Therefore, there is a need for an ideal defense method that effectively makes a balance between adversarial robustness and efficiency without compromising accuracy.

C. *Efficient Training in Transformers*

Transformer has shown great success in pattern recognition, but there are few related works on modulation recognition, where computational complexity would further make its adversarial training more complex. To improve this efficiency of Transformer models, extensive research has discussed the potential of efficient training by appropriately discarding image patches during the training or testing process. Tang et al. [27] introduce the patch slimming method, which reduces computational complexity by reducing the number of input image patches, thereby increasing the processing speed and computational efficiency of the model without significantly compromising the performance. Rao et al. [28] use a dynamic token sparsification approach to adjust the number of tokens based on each token's contribution to the final prediction, thereby optimizing resource allocation. This allows the model to maintain high accuracy while significantly reducing the computational load. These methods demonstrate that strategically discarding non-essential information can effectively improve training efficiency and mitigate the negative impacts of adversarial perturbations, which is in line with the motivation of adversarial defense.

In this paper, the AG-AMR method is designed to effectively optimize the Transformer model's attention mechanism, to reduce the computational complexity and to enable the integration with adversarial training, which achieves enhanced resilience to adversarial attacks while maintaining high accuracy.

III. PRIOR KNOWLEDGE

In this section, the model of the automatic modulation recognition system is introduced firstly. Then, an overview of the concepts of adversarial perturbations and adversarial training are provided in the context of modulation recognition tasks.

A. *Model of Automatic Modulation Recognition*

In general, a typical communication signal model can be represented as:

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

where $s(t)$ is the baseband signal from the transmitter, $h(t)$ is the impulse response of the wireless channel, and $n(t)$ signifies the additive white Gaussian noise during transmission. The modulation process begins by selecting a modulation scheme m, then using a modulator F to map the baseband signal $s(t)$ into a modulated signal, which is then transformed into a form suitable for transmission through the channel. The goal of the AMR task is to extract features from the received signal $r(t)$ without prior information and to estimate the modulation scheme m.

B. Adversarial Perturbations

In modulation recognition tasks, adversarial perturbations are designed to affect the input signal to make the model output incorrect predictions, thereby reducing the accuracy of the recognition. Such perturbations should be imperceptible, i.e., indistinguishable from normal noises, and no interfering with the normal operation of the receiver. To this end, the maximum amplitude of the adversarial perturbations is constrained with the l_∞-norm to ensure that the generated adversarial examples remain the characteristics of the original signals. The definition is given as follows:

$$\rho^* = \arg\max_{\rho} L_{y,\theta}(x+\rho) \quad \text{s.t.} \quad \|\rho\|_\infty \le \varepsilon \tag{2}$$

In this definition, ρ^* is the adversarial perturbation. The term $L_{y,\theta}(x+\rho)$ denotes the loss value between the predicted label and the true label y when the input signal is the perturbed signal $x+\rho$ and the model parameters are θ. The notation $\|\cdot\|_\infty$ denotes the l_∞-norm.

C. Adversarial Training

The goal of adversarial training is to find an optimal set of parameters such that the model minimizes the impact of adversarial perturbations. Specifically, adversarial training involves adjusting the model parameters using N training samples to minimize the loss function. This process is illustrated as follows:

$$\min_{\theta} \frac{1}{N} \sum_i \max L_{y_i,\theta}(x_i+\rho_i) \quad \text{s.t.} \quad \|\rho_i\|_\infty \le \varepsilon \tag{3}$$

In the formula, $\max L_{y_i,\theta}(x_i+\rho_i)$ represents the maximum perturbation under a constraint, and the overall goal of the formula is to obtain model parameters that minimizes the impact of adversarial perturbations on the classification loss. This formulation can be seen as an extension of clean sample training, since it corresponds to the objective function of the modulation recognition model under clean-sample training when $\rho_i = 0$.

IV. Modulation Recognition Guided by Attention Mechanism

This section provides a comprehensive overview of the proposed AG-AMR method, which includes the Attention-Guided Encoder (AG-Encoder), and the network structure, as illustrated in Fig. 1 The following subsections present the components in details.

A. Data Preprocessing and Feature Embedding

The long-range dependency and parallel processing capabilities of the Transformer are essential for effective AMR tasks [22]. In the process of data preprocessing and feature embedding, this method implements data segmentation and incorporates positional embeddings along with class tokens to achieve a comprehensive aggregation of the signals. This approach does not take structures like CNN and RNN, which thus avoid their limitations in focusing local features and high computational costs. By focusing on global patterns, the method enhances the model's ability to capture essential information from the input signals while keeping low computational complexity.

The input signal is treated as a two-channel image, where each channel corresponds to the real and imaginary parts of the signal. Specifically, each sample can be represented as a $H \times W \times C$ matrix, where each element contains information from the I/Q channels. To preserve the integrity of the original modulated signal and to take advantage of the feature extraction capabilities of the Transformer network, the input signal is normalized and framed. The input signal examples $x \in \mathbb{R}^{H\times W\times C}$ are segmented into a series of signal sequence frames $x_p \in \mathbb{R}^{N\times(H\cdot L\cdot C)}$ with frame length L, where (H, W, C) is the number of rows and columns of the original input signal matrix, $(H \times L)$ is the resolution of each sequence frame, and $N = W / L$ is the number of sequence blocks obtained, which is also the effective sequence processing length of the multi-head attention processing module. Then, these sequential frames are flattened one by one into one-dimensional columnwise vectors $x_t \in \mathbb{R}^{2L\times 1}$ with length $2L$. The vectors are horizontally concatenated to form the feature matrix $X \in \mathbb{R}^{N\times 2L}$. This process can be represented as:

$$\begin{aligned} x_t = [&r_I(tL), r_I(tL+1), \ldots, r_I(tL+L-1), \\ &r_Q(tL), r_Q(tL+1), \ldots, r_Q(tL+L-1)]^T \end{aligned} \tag{4}$$

$$X = [x_1, x_2, \ldots, x_n]^T \tag{5}$$

To generate the initial input token X_0 for training within the Transformer architecture, a combination of linear projection and positional embedding are used. First, the input features X are transformed into an intermediate token sequence $X_{proj} \in \mathbb{R}^{N\times D}$ via a linear projection, where D is the embedding dimension. Then, a class token $x_{cls} \in \mathbb{R}^{1\times D}$ is introduced to capture the global information of the entire signal sequence. This token is prepended to the front of each sequence, resulting in an extended sequence $[x_{cls}; X_{proj}]$, with x_{cls} as the first element of the sequence. Finally, to enable the model to capture the temporal information within the sequence, each element, including the class token, is

augmented with learnable positional embeddings $X_{pos} \in \mathbb{R}^{(N+1)\times D}$. In summary, the generation of the input sequence X_0 can be represented as:

$$X_{proj} = XW_{proj} \tag{6}$$

$$x_{cls} = \text{LearnableParameter}(D) \tag{7}$$

$$X_{(pos,n)} = \begin{cases} \sin(pos/10000^{2i/D}), n = 2i \\ \cos(pos/10000^{2i/D}), n = 2i+1 \end{cases} \tag{8}$$

$$X_0 = [x_{cls}; X_{proj}] + X_{pos} \tag{9}$$

where W_{proj} denotes the linear projection weight matrix, pos indicates the token sequence positions, and n represents the value of the corresponding feature dimension.

B. Attention-Guided Encoders

The Attention-Guided Encoder (AG-Encoder) is a critical component of the proposed method, which is designed to enhance the processing of input data through a structured framework with multiple layers. The AG-Encoder consists of a Multi-Head Self-Attention (MSA) layer and a Gated Linear Unit (GLU) layer. The MSA layer enables the model to focus on key regions of the input by dynamically adjusting attention weights based on the significance of the input features. This attention mechanism prioritizes important information, while minimizing the influence of irrelevant one, thereby enhancing the model's robustness to adversarial perturbations. The GLU layer, which replaces the conventional Multi-Layer Perceptron (MLP), addresses the computational intensity associated with feature extraction. By substituting the MLP with GLU, the model reduces computational load while retaining essential sequential information, which makes it to be particularly suitable for temporal tasks. This combination of attention and gating mechanisms optimizes the feature embedding process, which results in improved robustness against adversarial attacks, as it ensures that the most relevant features are preserved, and redundant or irrelevant features are discarded. The AG-Encoder can be expressed mathematically as follows.

1) Guided Multi-Head Attention Mechanism: In the MSA, the input sequence X_0 first undergoes layer normalization to standardize its values. It is then transformed into query (Q), key (K), and value (V). For multi-head attention, the matrices K, Q, V with shape $(N+1)\times D$ are divided into h heads, where each head now has the shape $(N+1)\times D/h$. Taking the first head as an example, V^1 is then reweighted by the attention scores computed from A^1. This process can be described as follows:

$$[K_{(N+1)\times D},\ Q_{(N+1)\times D},\ V_{(N+1)\times D}] = \text{LayerNorm}(X_0)W^1_{D\times 3D} \tag{10}$$

$$\text{Attn}(K^1_{(N+1)\times \frac{D}{h}}, Q^1, V^1) = SoftMax(\frac{Q^1K^1}{\sqrt{D}} + B)V^1 = A^1_{(N+1)\times(N+1)}V^1 \tag{11}$$

In this process, B represents a learnable bias term. Notably, all the column vectors in the matrix A^1 are normalized by the softmax function, ensuring that the sum of each column vector is equal to 1. Finally, the outputs $X^{'}$ of the MSA module are produced by concatenating and transforming the outputs A and V of each head.

$$X^{'}_{(N+1)\times D} = \text{Concat}(A^1V^1, A^2V^2, \ldots, A^hV^h)_{(N+1)\times D}W^2_{D\times D} \tag{12}$$

The magnitude of attention can reveal the importance of embedded features. Therefore, in our approach, this characteristic is used to filter out typical features while removing noises and interferences from the signals. Specifically, $\sum_{i=1}^{h} A^i / h$ is obtained by summing the attention matrix A and generating an index a, where the sum is always equal to 1, and its component represents the degree of receptivity to the embedded information. Conversely, the row vector of this result reveals the extent to which each input embedding influences the output embedding. Thus, the multi-head guided attention mechanism selects the top k embedded features based on the magnitude of the values in index a. Consequently, the output of the MSA, previously defined by Equation 12, is transformed as follows:

$$X^{'}_{k\times D} = \text{MaskBy}[\text{Concat}(A^1V^1, \ldots, A^hV^h), \text{Topk}(a)]_{k\times D}W^2_{D\times D} \tag{13}$$

2) Gated Linear Unit (GLU): In transformer networks, the feed-forward neural network (FFN) typically processes each token sequence independently and identically. In the AG-AMR, the Gated Linear Unit (GLU) is used as basic units, which is better suited for temporal tasks. Specifically, the GLU first uses its linear component to transform the input token sequences from the Multi-Head Self-Attention (MSA) mechanism, generating an intermediate output. Next, the gating component performs another linear transformation on the input and calculates a gating value ranging between 0 and 1 using an activation function such as a sigmoid. These gating values are used to modulate each element of the output from the linear transformation, thus dynamically controlling the amount of information flow. Finally, at the output stage of the GLU, the results of the linear transformation are multiplied elementwise by the gating values to produce the final output, as follows:

$$GLU(X^{'}) = (W_L X^{'} + b_L) \otimes (W_G X^{'} + b_G) \tag{14}$$

In the formula, W_L and W_G are the weight matrices for the linear and gating components, respectively, while b_L and b_G are the corresponding bias terms for these components.

C. Principle and Advantages

This section analyzes the principles and advantages of the proposed AG-AMR method. As shown in Fig. 2, which further emphasizes the attention-guided feature selection strategy based on the structure presented in Fig. 1, the model dynamically evaluates and filters data segments using the Guided Multi-Head Self-Attention mechanism.

Specifically, in the process of data preprocessing and feature embedding, the modulation signal is treated as a

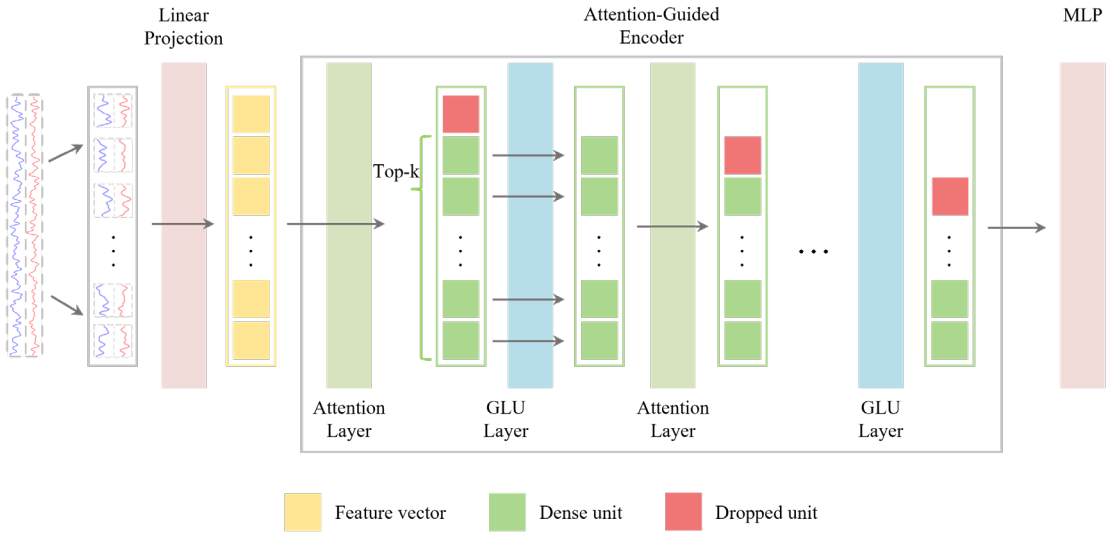

Fig. 2. The illustration of the Attention Guided Mechanism

two-channel image with real and imaginary parts. The signal undergoes standardization and framing to maintain its integrity. It is then segmented into multiple frames, flattened, and transformed into token sequences through linear projection, which enables effective aggregation of the information. The introduction of class tokens and positional embeddings further enhances the model's ability to capture global information and interpret temporal data. These steps provide the Transformer network with structured and comprehensive inputs, which promotes high accuracy on the tasks of modulation recognition.

The AG-AMR method enhances the model's performance by integrating the dynamic filtering capabilities of the attention mechanism with the gating control functions of the GLU. The MSA mechanism further refines the model's feature extraction by progressively reducing the number of embeddings from $N-1$ to k, where N represents the initial number of embeddings and k corresponds to the retained essential embeddings. This reduction minimizes redundancy and significantly reduces the computational burden by discarding non-essential information. Moreover, with the increasing of the number of layers of the network, many more embeddings are progressively discarded, which ensures computational efficiency while maintaining high-quality features for the final decision-making process. For instance, retaining 90% of the data in a typical 12-layer Transformer network can increase computational efficiency by approximately 40%.

Moreover, the GLU plays a crucial role in controlling information flow during the transformation process. It achieves data control through two pathways: the linear transformation path provides a new representation of the data, while the gating signal path generates weights ranging from 0 to 1 through a sigmoid function, which depicts the transmission strength of each element in the transformed data. By performing element-wise multiplication of these two outputs, the GLU finely controls the flow of the information, thereby it enhances the model's ability to capture and process complex data relationships.

The overall process of the proposed method is comprehensively described in Algorithm 1, which outlines the steps from signal preprocessing to embedding, feature selection, and training. The advantages of proposed method are presented as follows: Firstly, the method enhances the model's ability to process signals in parallel, which enables efficient handling of large volumes of data during the feature embedding phase. This parallel processing improves computational efficiency, which enables the model to scale well with increasing data sizes. Secondly, the attention-guided approach enables the model to focus on the most relevant features while dynamically filtering out redundant or irrelevant data, which enhances robustness against adversarial perturbations. Finally, the use of the GLU instead of traditional MLP layers reduces computational overhead without compromising model performance.

Algorithm 1 The AG-AMR Model

Input: Modulation signal with I/Q components, x
Output: Modulation type, M

$x' \leftarrow \text{Normalize}(x)$
$X \leftarrow \text{Frame}(x')$
$X_{proj} \leftarrow \text{LinearProjection}(X)$
$x_{cls} \leftarrow \text{GenerateClassToken}(D)$
$X_{pos} \leftarrow \text{PositionEmbedding}()$
$X_o \leftarrow [x_{cls}; X_{proj}] + X_{pos}$
for each layer in AG-Encoder **do**
    $K, Q, V \leftarrow \text{MSA}(X_0)$
    $X' \leftarrow \text{Select}(K, Q, V)$
    $G \leftarrow \text{GLU}(X')$
end for
$M \leftarrow \text{Classify}(G)$

V. EXPERIMENTS

In this section, extensive experiments are conducted to validate the effectiveness of the AG-AMR method. The proposed method is trained and tested as well as other typical models using publicly available datasets under adversarial attacks. All experiments are performed on the NVIDIA GeForce RTX 3090 GPU.

A. Experimental Setup

1) Datasets: In this study, automatic modulation recognition is conducted on the datasets of RML2016.10a and RML2018.01a. The RML2016.10a dataset has 220,000 modulation signals of 11 different modulation types. This dataset considers various factors of channel environments, including additive white Gaussian noise (AWGN), selective fading (Rician and Rayleigh), central frequency offset, and sampling rate offset. The signal-to-noise ratio (SNR) of samples is uniformly distributed between -20 dB and 18 dB at intervals of 2 dB. Each sample consists of 128 sampling points from in-phase (I) and quadrature (Q) components. The RML2018.01a dataset extends the range of modulation types and SNR values, as listed in Table I. Each modulation-SNR combination contains 4,096 frames, with each frame comprising 1,024 complex time series samples. Samples are represented as floating-point numbers for the in-phase and quadrature (I/Q) components, where each frame has a shape of (1024, 2). In total, the dataset includes 2,555,904 frames. During the experiments, stratified sampling is employed to partition the dataset into training, validation, and test sets in a 3:1:1 ratio for different modulation types.

Table I PARAMETERS OF THE BENCHMARK DATASETS

Datasets Name	RML 2016.10a	RML 2018.01a
Signal format	In-phase and Quadrature	
Number of samples	220000	2555904
Data Dimension	(2, 128)	(2, 1024)
Range of SNR	-20dB: 2dB: 18dB	-20dB: 2dB: 30dB
Modulations	8 Digital Modulation 8PSK, BPSK, CPFSK, GFSK, PAM4, 16QAM, 64QAM, QPSK 3 Analog Modulation AM-DSB, AM-SSB, WBFM	19 Digital Modulation OOK, 4ASK, 8ASK, BPSK, QPSK, 8PSK, 16PSK, 32PSK, 16APSK, 32APSK, 64APSK, 128APSK, 16QAM, 32QAM, 64QAM, 128QAM, 256QAM, OQPSK, GMSK 5 Analog Modulation AM-SSB-WC, FM AM-SSB-SC, AM-DSB-WC, AM-DSB-SC
Channel Conditions	Noise (AWGN) Selective fading (Ricean&Rayleigh) Carrier frequency offset (CFO) Symbol rate offset (SFO)	

2) Models: To assess the performance of the AG-AMR method on the aforementioned datasets, it is compared against several competitive models proposed in recent years, including MCLDNN[21], PET-CGDNN[29], LSTM[19], and GRU[30] models. MCLDNN employs a complementary processing approach by decomposing IQ data into three input streams. It utilizes convolutional modules for feature extraction and fusion, followed by two LSTM layers and a fully connected network for further feature extraction. The convolutional kernel sizes are set to (8, 50), (2, 8, 50), (1, 8, 50), and (2, 5, 100), while the number of units in the LSTM and fully connected layers are set to 250 and 128, respectively. The LSTM model leverages amplitude/phase features, which are processed through two LSTM layers, and a single fully connected layer and output the classification results. The PET-CGDNN and GRU models combine convolutional layers with gated recurrent units to recognize the IQ data. The convolutional kernels are set to (2, 8, 75) and (1, 5, 25), with the number of gated recurrent units set to 128. The Transformer model employs the original multi-head attention mechanism with 12 layers. All models use the same training principles and configurations. The batch size is set to 128, and the initial learning rate is 0.01. If the validation loss does not decrease for 10 consecutive epochs, the learning rate is halved. If the validation loss does not decrease for 50 consecutive epochs, training is terminated early.

3) Adversarial Attacks: To evaluate the robustness of the models against adversarial examples, extensive adversarial attack tests are conducted. These tests involve various gradient-based and optimization-based adversarial attack methods. The Fast Gradient Sign Method (FGSM) generates adversarial examples by adding small perturbations to the original data in the direction of the loss function gradient. Projected Gradient Descent (PGD), an iterative version of FGSM, creates stronger adversarial examples through updates with multiple small steps. For optimization-based attack methods, the Carlini & Wagner (C&W) attack iteratively minimizes the distance to the decision boundary to mislead the classifier into making incorrect decisions. Additionally, AutoAttack (AA) is an automated attack framework that combines multiple attack strategies to generate robust and reliable adversarial examples. These attack methods pose widely recognized threats to deep learning models and are commonly used to test a model's resilience against potential attacks. All the attack methods are implemented using the *Adversarial Robustness Toolbox*. By employing these validation techniques, a comprehensive evaluation and analysis of the models' robustness is ensured when being confronted with adversarial sample attacks.

B. The Analysis of Optimal Structural Parameters

To reveal the impact of frame length for feature embedding and the network layers of the AG-Encoder on model performance, experiments are conducted with varying frame lengths and network depths. The results on the two datasets are presented in Table II. It is evident that, with the same frame length, deeper network layers provide the model with stronger feature extraction capabilities, resulting in better recognition performance. Notably, the recognition performance of AG-AMR demonstrates sensitivity to the

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2025.3526577

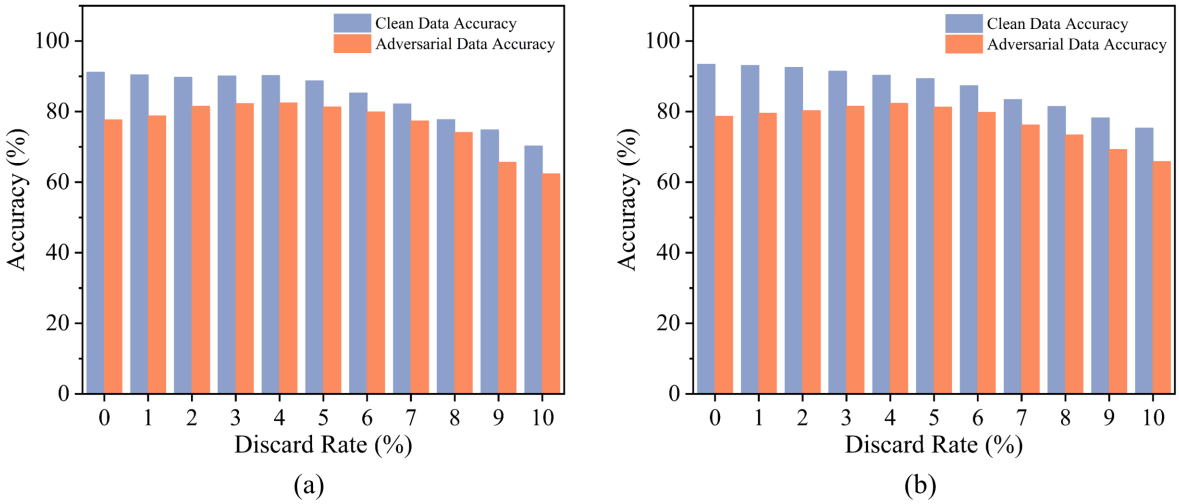

Fig. 3. Accuracy comparison of different discard rate on (a) RML2016.10a, (b) RML2018.01a

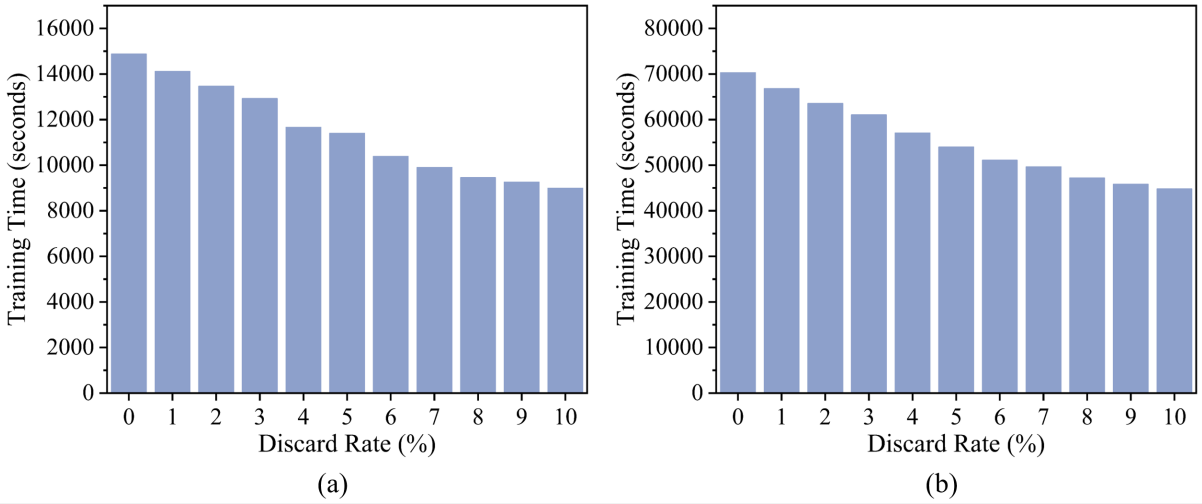

Fig. 4. Training time comparison of different discard rate on (a) RML2016.10a and (b) RML2018.01a

Table II MODEL PERFORMANCE ACROSS DIFFERENT FRAME LENGTHS AND NUMBER OF BLOCKS

Frame Length (L)	Number of blocks (H)	RML2016.10a			RML2018.01a		
		Parameters	Training time (sec/epoch)	Accuracy (%)	Parameters	Training time (sec/epoch)	Accuracy (%)
2	4	10.9M	33	59.60	8.9M	762	32.09
2	8	11.1M	70	60.70	9.1M	1628	33.54
2	12	11.3M	112	**61.32**	9.3M	2681	33.93
8	4	4.7M	16	58.29	3.4M	347	38.74
8	8	4.9M	31	59.84	3.6M	525	40.26
8	12	5.1M	45	60.25	3.8M	1058	41.67
16	4	3.8M	13	57.20	**3.4M**	304	49.03
16	8	3.9M	25	57.25	3.6M	573	50.48
16	12	4.1M	36	57.66	3.8M	837	50.82
32	4	**3.5M**	12	55.00	5.0M	187	59.41
32	8	3.7M	21	57.08	5.2M	354	61.74
32	12	3.9M	31	57.35	5.3M	517	**64.03**
64	4	3.8M	**11**	50.48	8.9M	**146**	49.51
64	8	3.9M	20	51.62	9.1M	270	50.04
64	12	4.1M	28	52.66	9.3M	398	50.22

Table III COMPARISON OF MODEL SIZE AND COMPLEXITY ON TWO DATASETS: RML2016.10a (denoted by A) and RML2018.01a (denoted by B)

Model	Learning parameters		Training time (sec/epoch)		Training epochs		Minimum validation loss	
Dataset	A	B	A	B	A	B	A	B
AG-AMR	11261259	5349976	112	517	121	134	1.0878	1.1191
GRU	151179	152856	9	313	103	162	1.1124	**1.1124**
LSTM	201099	202776	11	497	**89**	106	1.1004	1.1192
MCLDNN	406199	407876	17	662	103	**92**	**1.0612**	1.1313
PET-CGDNN	**71871**	**75340**	**6**	**208**	97	311	1.0945	1.1185
Transformer	15765744	9547657	161	2796	156	169	1.0852	1.1182

input frame length, achieving the highest accuracy with specific frame lengths on both datasets. Due to the differences in input sample dimensions across different datasets, the model training process should aim to extract feature information adequately through appropriate frame lengths. Specifically, short frames carrying less information are prone to overfitting due to over-interpretation, while long frames carrying too much information might not be fully utilized for feature extraction, leading to suboptimal recognition performance. Moreover, the training time of the model decreases with an increase in frame length and increases with an increase in network depth. Shorter frames generate more training tokens, necessitating longer training times. To validate the effectiveness of AG-AMR, we adopted the optimal structural parameters, which yielded the best accuracy, for further experiments on the RML2016.10a and RML2018.01a.

Additionally, the impact of the attention mechanism on accuracy is explored by conducting experiments with different inter-layer discard rates in the guided multi-head attention mechanism. The results are shown in Fig. 3. The model's accuracy is evaluated using clean examples with an SNR above 0 dB and adversarial examples generated by FGSM white-box attacks with a perturbation intensity of 0.001. As observed, increasing the dropout rate leads to a decrease in accuracy for clean examples, while the accuracy for adversarial examples gradually improves. Analysis indicates that when the guided inter-layer dropout rate is set to 4%, the model achieves the most balanced performance on both datasets, maintaining 90.2% and 91.4% accuracy for clean sample recognition and reaching 82.45% and 82.35% robustness accuracy against adversarial attacks, respectively.

The comparison of training times on the two datasets is illustrated in Fig. 4. Compared to the original model that retains all information, the guided attention model with a 4% dropout rate reduced computational load by 19.31% and shortened the training time by an average of 22.60%. This demonstrates that the AG-AMR method not only excels in balancing accuracy and adversarial robustness but also reduces the total amount of data involved in the forward pass, allowing the model to focus on key information and achieve faster convergence.

C. Recognition Performance on Clean Examples

In order to comprehensively evaluate the performance of AG-AMR, two AG-AMR curves are presented in Fig. 5. The first one, labeled by AG-AMR (with best performance), represents the performance achieved by the model with the best configuration. The second one, labeled by AG-AMR (with most efficiency), demonstrates the performance of the most efficient configuration in terms of computational complexity. These curves are compared with the original Transformer model and other typical models as described in Section 5.A.

In the experiments on RML 2016.10a, the overall accuracy of GRU is 58.42%, with an accuracy above 0 dB of 85.52%, which is significantly lower than the results of the baseline models. The overall accuracy of the PET-CGDNN and LSTM across all SNRs is 60.53% and 60.45%, respectively. However, the LSTM model performs well above 0 dB, with both models achieving 89.91% and 89.75%, respectively. The MCLDNN and Transformer have average accuracies of 61.48% and 61.39%, respectively, and recognition rates of 91.06% and 91.39% above 0 dB SNR. The AG-AMR (with best performance) method achieves an average accuracy of 61.49%, which surpasses all models. This model significantly improves accuracy around 0 dB SNR and approximates the optimal models at high SNRs. The experiments on the RML 2018.01a dataset show similar results. The AG-AMR (with best performance) method achieves an overall accuracy of 64.03%, which outperforms the MCLDNN and PET-CGDNN models, with a notable advantage around 0 dB SNR and competitive performance at high SNRs.

By comparing the experimental results of AG-AMR (with best performance) and AG-AMR (with most efficiency), it is obvious to see that, in terms of performance, particularly the overall recognition accuracy, the results of the most efficient one are significantly lower than those of the best performance model. Specifically, in the RML 2016.10a dataset, the most efficient model achieves an overall accuracy of 55.00% and a recognition rate above 0 dB of 80.88%, which is lower than both the best performance

models of 61.49% and 90.65%. Across the entire SNR range, the most efficient performs worse than the optimal model, especially at higher SNRs, where the accuracy gap is particularly pronounced. Similarly, in the RML 2018.01a dataset, the most efficient one also shows lower accuracy, with an overall accuracy of 49.03%, far below the best one 64.03%. Additionally, it is further observed that, although the most efficient model exhibits lower accuracy than the best performance, it demonstrates significant advantages in efficiency. Specifically, in the experiments on both datasets, the most efficient model reduces the number of parameters by 7.8M and the training time by 2.6 hours on the RML 2016.10a dataset, and on the RML 2018.01a dataset, the number of parameters is reduced by 0.9M and training time by 6.1 hours.

The analysis of the comparison of model size and complexity across two datasets, as shown in Table III, reveals that AG-AMR has a larger number of parameters compared to other advanced models. This is primarily due to the use of the Transformer's Encoder structure, which inherently comes with a substantial number of parameters. However, despite this, AG-AMR achieves a reduction in both model volume and training time when compared to the original Transformer model. Specifically, AG-AMR has 11.3M parameters, which is lower than 15.8M of the original Transformer model, although it is still higher than other lightweight models such as GRU (0.15M) and LSTM (0.2M). Regarding the training time, AG-AMR takes 3.76 hours, which is shorter than the original Transformer's by 3.21 hours in RML2016.10a. By incorporating the improved attention-guided mechanism and optimized structural design, AG-AMR makes a good balance between training efficiency and high performance, which demonstrates that structural optimization can significantly improve computational efficiency without sacrificing model accuracy.

These experiments indicate that the AG-AMR with an attention-guided mechanism can effectively discard some detailed information in signals. This discarding allows the model to remove noise components from the input signal, thereby increases the model's focusing on critical information. Additionally, by fine-tuning the model structure and the attention-guided mechanism, the AG-AMR method achieves a modulation recognition model that weakens the impact of noises while maintaining high accuracies.

D. Robust Performance Against Adversarial Examples

The robustness of the AG-AMR method under different adversarial attacks is evaluated using several common adversarial attack methods, including FGSM, PGD, C&W, and AutoAttack. In order to ensure the validity of the attack results, the RML2018.01a dataset is pre-processed by screening and selecting examples that can be correctly classified by the model. Since each model exhibits different accuracy on the test set, the correctly classified examples also vary across models. By following the method in [26], a total of 255,590 correctly classified examples are selected to form the attack test set.

For each model, various levels of adversarial attacks are performed on the selected examples, and the models' robustness is compared based on their accuracy under different attack intensities. The experimental results are depicted in Fig. 6. It is obvious to see that, all models exhibit a decreasing trend in accuracy as the attack intensity increases. Specifically, under common attacks of FGSM and PGD, when the perturbation intensity reaches 0.005, the accuracy of baseline models such as GRU, LSTM, and MCLDNN drops rapidly to a value below 20%, which indicates that these models are highly vulnerable to adversarial perturbations. In contrast, the AG-AMR method consistently maintains an accuracy of around 30% under the same conditions, which demonstrates its robustness.

Under the AA attack, although all models exhibit a significant decrease in accuracy, the AG-AMR shows a clear advantage of performance over other models under the same attack intensity. For instance, with an attack intensity of 0.005, AG-AMR maintains an accuracy of about 30%, while other models show accuracy below 20%. This phenomenon demonstrates the effectiveness of the attention-guided feature selection strategy in improving the model's robustness. Moreover, AG-AMR's robustness is particularly evident in the face of high-intensity C&W attack, where adversarial perturbations are strategically crafted to bypass model defenses. Under this attack, the accuracy of AG-AMR remains at around 23%, whereas other models' accuracy values drop to below 17%. The trends shown in Fig. 6 highlight that, while all models experience a certain level of performance degradation under strong adversarial attacks, AG-AMR's accuracy drops insignificantly, which demonstrates its ability to withstand stronger adversarial perturbations. This result further validates the effectiveness of its attention-guided mechanism, which allows the model to dynamically focus on key features of the signal while disregarding irrelevant noise components, thus reducing the impact of adversarial perturbations.

E. Defense Performance Comparison with Adversarial Training

Adversarial training is an effective means of enhancing model robustness. In this section, the models are fine-tuned using adversarial training on the RML2018.01a dataset, evaluating the advancement of the proposed method through the implementation of various adversarial attack methods.

Due to the difficulty in achieving high accuracy with low SNR examples (modulated signals below -2 dB), it is unnecessary to perform adversarial attacks on these examples. This experiment primarily selects signal with SNRs greater than -2 dB as test subjects. At each SNR level, 20% of the training set are sampled to create the adversarial training dataset. For generating adversarial training examples, the most common PAT method [31] for gradient-based attack methods is used. This method employs adversarial examples generated by PGD attacks for training, setting the maximum perturbation coefficient epsilon to 0.01, with a perturbation step size of 0.002 and 5 steps. For

optimization-based attack methods, examples generated by the C&W attack method with 5 query rounds are used as the training set.

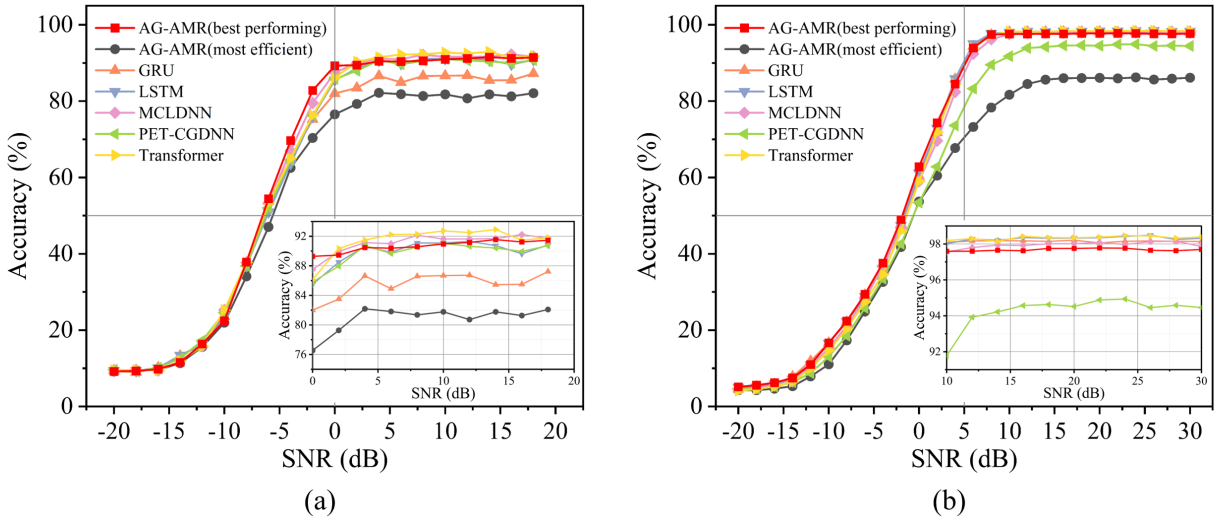

Fig. 5. Comparison on accuracy of models on (a) RML2016.10a, (b) RML2018.01a

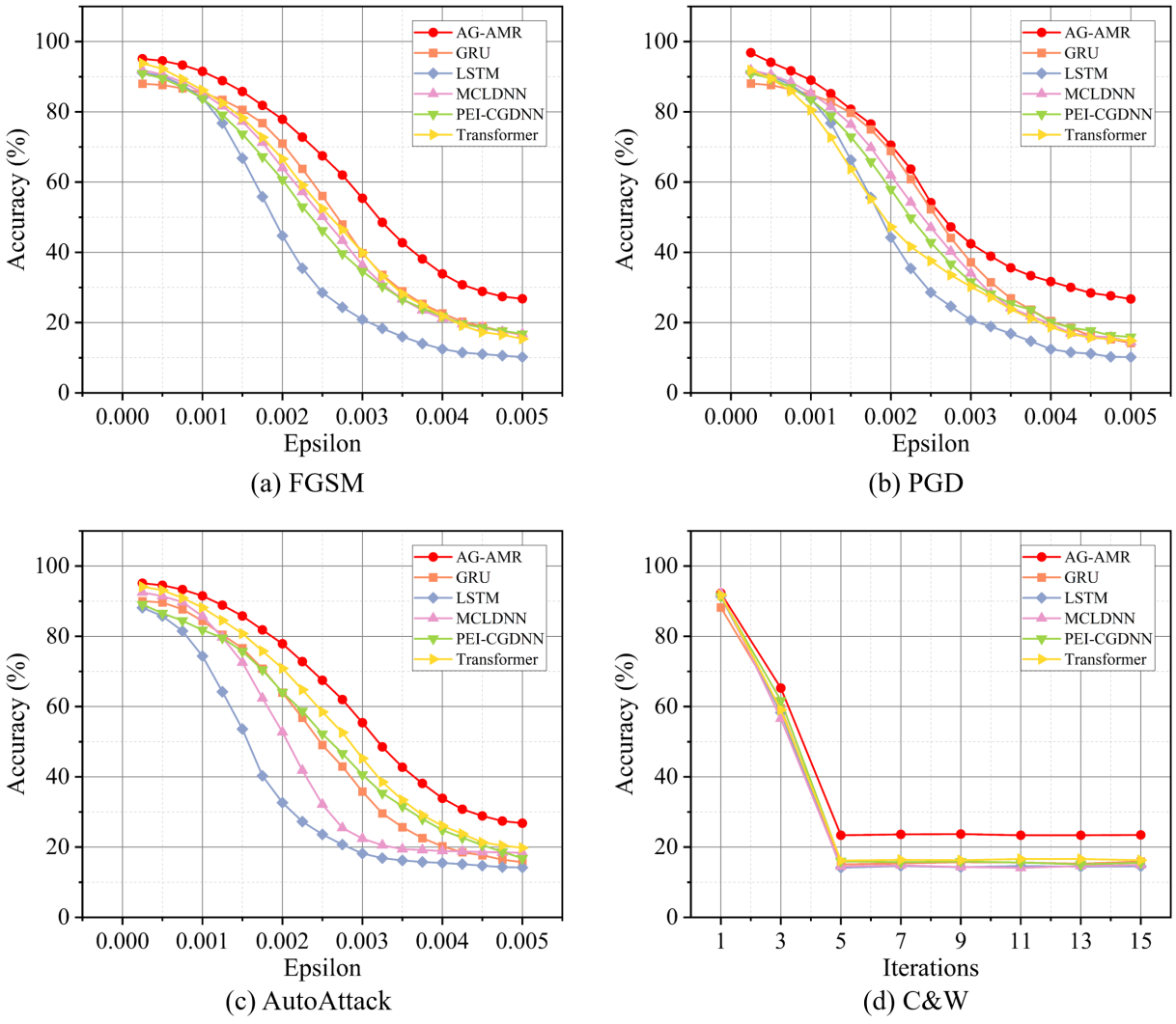

Fig. 6. Different degrees of different attack on AMR models on RML2018.01a

Table IV ACCURACY OF DIFFERENT MODELS UNDER DIFFERENT ATTACKS

Models	Average Accuracy from -2dB to 30dB (%)				
	Clean	PGD	FGSM	C&W	AA
Transformer (baseline)	**90.53**	8.95	11.49	15.11	8.69
Adversarial Training	80.45	47.30	55.88	26.75	12.74
Random Smoothing	77.37	34.62	47.46	27.83	11.97
Gradient Masking	79.62	32.38	41.82	24.65	10.29
AG-AMR	85.19	**72.69**	**74.58**	**64.90**	**21.02**

After fine-tuning with adversarial training, the adversarial trained models are tested using clean examples and adversarial examples generated by PGD, FGSM, C&W, and AutoAttack methods. To make the attack effects more evident, the attack intensity is increased, setting epsilon to 0.03 for PGD, FGSM, and AutoAttack, and setting iterations to 30 for C&W attacks. Since the proposed method is also based on the Transformer architecture, the baseline model for comparison is the original Transformer model. This ensures that the improvements achieved by the proposed method are due to the optimizations introduced within the Transformer structure itself. Various adversarial defense methods are selected for comparison, including the baseline Transformer with its adversarial training approach, as well as other defense techniques such as random smoothing and gradient masking. The experimental test results are shown in Table IV. The results indicate that, although the baseline model exhibits excellent recognition performance on clean examples, its accuracy drastically drops to 8.69% when facing adversarial attacks, especially during AutoAttack. Compared to the adversarial training method using the original Transformer, AG-AMR demonstrates better overall performance, exhibiting not only more prominent defense capabilities but also maintaining an 85.19% accuracy on clean examples. Furthermore, compared to typical random smoothing and gradient masking methods, AG-AMR combined with adversarial training shows superior accuracy, particularly in robustness against PGD attacks, achieving 2 to 3 times the robustness of other methods.

VI. CONCLUSION

In this paper, a robust modulation recognition method based on attention guidance, AG-AMR, is proposed. AG-AMR fully leverages the advantages of attention mechanisms in extracting key features, guiding and optimizing attention distribution during the training process, which enables the model to effectively filter and enhance critical signal features. Through extensive exploration of the model structure and numerous experiments on public datasets, this method is validated to outperform existing typical modulation recognition models and adversarial defense methods in terms of accuracy and robustness. Additionally, the performance comparison after adversarial training further demonstrates the advantage of the AG-AMR method in balancing precision and robustness. The AG-AMR method provides performance enhancement and security optimization solutions for deep learning-based modulation recognition models, enhancing the stability and reliability of AI-driven spectrum monitoring systems and non-cooperative communication systems in practical applications.

REFERENCES

[1] F. A. Bhatti, M. J. Khan, A. Selim, and F. Paisana, "Shared spectrum monitoring using deep learning," IEEE Transactions on Cognitive Communications and Networking, vol. 7, no. 4, pp. 1171-1185, 2021.

[2] W. Zhang, M. Feng, M. Krunz, and A. H. Y. Abyaneh, "Signal detection and classification in shared spectrum: A deep learning approach," in IEEE INFOCOM 2021-IEEE Conference on Computer Communications, 2021: IEEE, pp. 1-10.

[3] F. Hameed, O. A. Dobre, and D. C. Popescu, "On the likelihood-based approach to modulation classification," IEEE transactions on wireless communications, vol. 8, no. 12, pp. 5884-5892, 2009.

[4] S. Huang, Y. Yao, Z. Wei, Z. Feng, and P. Zhang, "Automatic modulation classification of overlapped sources using multiple cumulants," IEEE Transactions on Vehicular Technology, vol. 66, no. 7, pp. 6089-6101, 2016.

[5] Y. Chen, B. Dong, C. Liu, W. Xiong, and S. Li, "Abandon Locality: Frame-Wise Embedding Aided Transformer for Automatic Modulation Recognition," IEEE Communications Letters, vol. 27, no. 1, pp. 327-331, 2022.

[6] J. Ma, M. Hu, T. Wang, Z. Yang, L. Wan, and T. Qiu, "Automatic modulation classification in impulsive noise: Hyperbolic-tangent cyclic spectrum and multibranch attention shuffle network," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-13, 2023.

[7] S. Lin, Y. Zeng, and Y. Gong, "Learning of time-frequency attention mechanism for automatic modulation recognition," IEEE Wireless Communications Letters, vol. 11, no. 4, pp. 707-711, 2022.

[8] Z. Ke and H. Vikalo, "Real-time radio technology and modulation classification via an LSTM auto-encoder," IEEE Transactions on Wireless Communications, vol. 21, no. 1, pp. 370-382, 2021.

[9] T.-T. Dao, D.-I. Noh, M. Hasegawa, H. Sekiya, Q.-V. Pham, and W.-J. Hwang, "VT-MCNet: High-Accuracy Automatic Modulation Classification Model based on Vision Transformer," IEEE Communications Letters, 2023.

[10] C. Szegedy et al., "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[11] Y. Shi, Y. E. Sagduyu, K. Davaslioglu, and R. Levy, "Vulnerability detection and analysis in adversarial deep learning," Guide to

Vulnerability Analysis for Computer Networks and Systems: An Artificial Intelligence Approach, pp. 211-234, 2018.
[12] D. Xu, H. Yang, C. Gu, Z. Chen, Q. Xuan, and X. Yang, "Adversarial examples detection of radio signals based on multifeature fusion," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 68, no. 12, pp. 3607-3611, 2021.
[13] S. Kokalj-Filipovic, R. Miller, and G. Vanhoy, "Adversarial examples in RF deep learning: Detection and physical robustness," in 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2019: IEEE, pp. 1-5.
[14] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," IEEE Transactions on Wireless Communications, vol. 21, no. 6, pp. 3868-3880, 2021.
[15] S. Kokalj-Filipovic, R. Miller, N. Chang, and C. L. Lau, "Mitigation of adversarial examples in rf deep classifiers utilizing autoencoder pre-training," in 2019 International Conference on Military Communications and Information Systems (ICMCIS), 2019: IEEE, pp. 1-6.
[16] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in Engineering Applications of Neural Networks: 17th International Conference, EANN 2016, Aberdeen, UK, September 2-5, 2016, Proceedings 17, 2016: Springer, pp. 213-226.
[17] N. E. West and T. O'shea, "Deep architectures for modulation recognition," in 2017 IEEE international symposium on dynamic spectrum access networks (DySPAN), 2017: IEEE, pp. 1-6.
[18] K. Karra, S. Kuzdeba, and J. Petersen, "Modulation recognition using hierarchical deep neural networks," in 2017 IEEE international symposium on dynamic spectrum access networks (DySPAN), 2017: IEEE, pp. 1-3.
[19] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," IEEE Transactions on Cognitive Communications and Networking, vol. 4, no. 3, pp. 433-445, 2018.
[20] Z. Zhang, H. Luo, C. Wang, C. Gan, and Y. Xiang, "Automatic modulation classification using CNN-LSTM based dual-stream structure," IEEE Transactions on Vehicular Technology, vol. 69, no. 11, pp. 13521-13531, 2020.
[21] J. Xu, C. Luo, G. Parr, and Y. Luo, "A spatiotemporal multi-channel learning framework for automatic modulation recognition," IEEE Wireless Communications Letters, vol. 9, no. 10, pp. 1629-1632, 2020.
[22] J. Cai, F. Gan, X. Cao, and W. Liu, "Signal modulation classification based on the transformer network," IEEE Transactions on Cognitive Communications and Networking, vol. 8, no. 3, pp. 1348-1357, 2022.
[23] M. Z. Hameed, A. György, and D. Gündüz, "The best defense is a good offense: Adversarial attacks to avoid modulation detection," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1074-1087, 2020.
[24] L. Zhang, S. Lambotharan, G. Zheng, B. AsSadhan, and F. Roli, "Countermeasures against adversarial examples in radio signal classification," IEEE Wireless Communications Letters, vol. 10, no. 8, pp. 1830-1834, 2021.
[25] H. Tang, F. O. Catak, M. Kuzlu, E. Catak, and Y. Zhao, "Defending AI-Based Automatic Modulation Recognition Models Against Adversarial Attacks," IEEE Access, 2023.
[26] Z. Chen et al., "Learn to Defend: Adversarial Multi-Distillation for Automatic Modulation Recognition Models," IEEE Transactions on Information Forensics and Security, 2024.
[27] Y. Tang et al., "Patch slimming for efficient vision transformers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12165-12174.
[28] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," Advances in neural information processing systems, vol. 34, pp. 13937-13949, 2021.
[29] F. Zhang, C. Luo, J. Xu, and Y. Luo, "An efficient deep learning model for automatic modulation recognition based on parameter estimation and transformation," IEEE Communications Letters, vol. 25, no. 10, pp. 3287-3290, 2021.
[30] D. Hong, Z. Zhang, and X. Xu, "Automatic modulation classification using recurrent neural networks," in 2017 3rd IEEE International Conference on Computer and Communications (ICCC), 2017: IEEE, pp. 695-700.
[31] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, 2017.

Quanhai Zhan is currently pursuing his master's degree in the Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing China. His interests are machine learning and wireless security.

Xiongwei Zhang received the Ph.D. degree in signal and information processing from the Nanjing Institute of Communications Engineering, Nanjing, China, in 1992. He is currently a professor with the Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing. His research interests include speech signal processing, machine learning, and pattern recognition.

Meng Sun received his Ph.D. degree from the Department of Electrical Engineering in 2012, Katholieke University Leuven, Belgium. He is currently a professor with Army Engineering University, Nanjing, China. His research interests include signal processing, machine learning, and pattern recognition.

Lei Song received his Ph.D. degree from the National University of Defense Technology, Changsha, China, in 2001. He is currently an associate professor with Army Engineering University, Nanjing, China. His research interests include signal processing, wireless security, and sequential pattern recognition.

Zhenji Zhou received his Ph.D. degree in computer science from the PLA University of Science and Technology, Nanjing, China, in 2013. He is currently an associate professor with the Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing. His research interests include vulnerability mining, machine learning, and pattern recognition.