Wireless Interference Recognition With Multimodal Learning

Pengyu Wang, Ke Ma, *Graduate Student Member, IEEE*, Yingshuang Bai, Chen Sun, *Senior Member, IEEE*, Zhaocheng Wang, *Fellow, IEEE*, and Sheng Chen, *Life Fellow, IEEE*

***Abstract*— In non-cooperative communications, malicious electromagnetic interference attacks communication systems and causes higher probability of communication disruption. In order to address the challenges posed by electromagnetic interference, the wireless interference recognition technique has emerged, which identifies the interference signals without priori information. In recent years, the success of deep learning (DL) has sparked interest in introducing DL in the field of wireless interference recognition. However, most DL-based interference identification methods improve accuracy by dramatically increasing network sizes while ignoring the important effect of network inputs. For this reason, we extensively investigate the impact of different signal transformation forms of interference (called signal modalities) on performance. The artificial features of the interference signal are also utilized as one of the refined modalities, which breaks the inherent concept that artificial features are only used in the methods of feature extraction. Convolution and transformer are combined in the extraction of different modal features. In order to reduce the complexity of transformer, a dual transformer module (DTM) is proposed. Furthermore, to overcome the imbalance of modal optimization during the training process, an adaptive gradient modulation (AGM) strategy is proposed, which leads to better convergence for the multimodal training. Finally, modal information selection mechanism (MISM) selects the most appropriate modalities for each input sample, which saves computational costs. Extensive experiments demonstrate that combining multiple interference modalities is more effective than trying different networks.**

***Index Terms*— Wireless interference recognition, multimodal learning, convolutional neural networks, anti-interference communication.**

Received 4 May 2024; revised 28 July 2024; accepted 24 September 2024. Date of publication 7 October 2024; date of current version 12 December 2024. This work was supported in part by the National Natural Science Foundation of China under Grant U22B2057, in part by the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (CPSF) under Grant GZB20240387, and in part by China Postdoctoral Science Foundation under Grant 2024T170492. The associate editor coordinating the review of this article and approving it for publication was F. Chiariotti. *(Corresponding author: Zhaocheng Wang.)*

Pengyu Wang, Ke Ma, and Zhaocheng Wang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: wangpengyu@mail.tsinghua.edu.cn; ma-k19@mails.tsinghua.edu.cn; zcwang@tsinghua.edu.cn).

Yingshuang Bai and Chen Sun are with the Wireless Network Research Department, Sony China Research Laboratory, Beijing 100022, China (e-mail: Yingshuang.Bai@sony.com; Chen.Sun@sony.com).

Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, U.K. (e-mail: sqc@ecs.soton.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2024.3470244.

Digital Object Identifier 10.1109/TWC.2024.3470244

I. Introduction

With the emergence of new services in wireless communications and the rapid development of satellite communications, the scarcity of wireless spectrum resources has become a pressing issue [1], [2]. The regulation of radio magnetic spectrum is of great practical importance as it can effectively and reasonably allocate spectrum resources to meet the needs of as many legitimate users as possible [3], [4]. At the same time, the regulation of spectrum enables the detection of illegal users and safeguarding the legitimate users. Therefore, it is imperative to strengthen monitoring and management of spectrum resources [5], [6].

In traditional collaborative communications, both the transmitter and receiver have a priori communication information, such as transmission bandwidth, frame format and modulation method. Unlike collaborative communications, non-collaborative wireless signal awareness has to make detection and analysis with almost no a priori knowledge at all, which is a key aspect of magnetic spectrum regulation [7]. The wireless interference signal recognition refers to detection and identification of electromagnetic signals released by non-cooperative users under the condition that the parameters of these signals are completely unknown, so as to obtain and understand the information of the electromagnetic environment and to ensure safe and reliable communication [8], [9].

Information confrontation is a form of modern warfare. How to effectively identify and detect the non-cooperative wireless signals becomes more and more critical [10]. In a complex and changing electromagnetic environment, wireless interference identification can achieve the discovery, monitoring and reconnaissance of non-cooperative signals [11], [12]. The anti-interference technology must be aware of interference released by non-partners in order to select an appropriate technical means to eliminate or mitigate the effects of the interference. Communication anti-interference technology usually includes interference recognition, interference suppression, anti-interference decision and interference avoidance [13], [14], [15]. Among these, wireless interference recognition is the foundation of anti-interference communication [16]. If the type of interference signal can be efficiently and accurately identified, the corresponding subsequent anti-interference measures can be developed to minimize the damage to communication quality. Therefore, wireless interference signal cognition is of great significance for anti-interference communication [17].

In civil communication, the main application scenarios of wireless interference cognition technology are in the fields of cognitive radio, spectrum monitoring and radio management [18], [19], [20]. Wireless interference cognition technology can regulate the use of spectrum and ensure the communication of all kinds of legal users. In addition, for public frequency bands, wireless interference recognition technology can effectively schedule the use of spectrum and avoid communication collisions. Therefore, in both civil and military communications, wireless interference identification technology plays an important role [21], [22].

Traditional identification methods can generally be divided into two categories: maximum likelihood-based and feature-based methods. The likelihood-based methods utilize a Bayesian minimum error probability criterion to construct the test statistic and develop an optimal judgement threshold to yield identification results, which theoretically achieves the best performance. The interference identification based on Naive Bayes classifier was investigated in [23]. The simulation results reported in [23] indicated that the proposed method achieved a better average accuracy compared with other traditional methods. The study [24] used generalized likelihood ratios to cope with deception interference signals, and simulation results showed that the generalized likelihood ratio approach had a high identification probability for detecting deception interference. The work [25] studied adaptive coherent estimator and the generalized likelihood ratio test for detecting and classifying jamming signals. The mean likelihood ratio-based identification method was adopted in [26] for classifying binary phase shift keying (BPSK) and quadrature phase shift keying signals. The work [27] studied joint detection and automatic classification of multiple interference signals using a generalized dynamic Bayesian network. However, the aforementioned maximum likelihood-based approaches require full knowledge of the wireless channel information, which is not available in most communication scenarios.

The feature-based approach relies heavily on feature extraction. Feature extraction is used to obtain features that can distinguish different interfering signals, and these features essentially reflect the inherent differences between the diverse signals. The study [28] investigated the bispectral and singular spectrum analysis of interfering signals and classifies these interferences by means of artificial neural networks. This method improved the classification recognition rate compared to conventional algorithms. The work [29] used fourth-order cumulants as features for signal extraction, which effectively distinguished between multiple electromagnetic signals and was robust to noise. The authors of [30] considered the recognition of wireless signals in communication scenarios with high speed movement and impulse noise. They used cumulants as extracted signal features and analyzed the recognition accuracy of higher order cumulants in this communication scenario. However, the aforementioned feature extraction methods require expert knowledge with hand-crafting features.

Deep learning (DL) is one of the most rapidly developing research areas in the recent years, and it shows an extraordinary potential in the areas such as image recognition, intelligent transportation, natural language processing and sentiment analysis [31], [32], [33]. Moreover, DL has been introduced into the field of wireless communication [34], [35], [36]. Thanks to its powerful feature extraction and data mining capabilities, DL offers a completely new solution in the field of interference signal recognition and has shown amazing recognition performance [37], [38]. The convolutional neural network (CNN) with an attention mechanism was proposed in [39] for deception interference recognition. The experimental results reported in [39] showed that the proposed method achieved a higher recognition accuracy and faster convergence speed than conventional methods. The work [40] introduced a novel distributed few-shot learning method for interference identification, with multiple sub-networks adopting federated learning to achieve global optimization. The simulation results of [40] showed that the proposed method can achieve good performance even with small training data set. The authors of [41] proposed a two-stage training strategy for CNN based signal recognition methods, and the experimental results showed that the proposed CNN network training method outperforms manual feature extraction methods based on higher-order cumulants. In [42], a complex CNN network was designed, which can better model the interrelationship between the homogeneous and orthogonal components compared with the traditional CNN network. The study [43] proposed a weighted ensemble CNN with transfer learning for classifying radar active deception interference signals. A denoising diffusion probability model was utilized in [44] to identify diverse types of interference in real-time communication scenarios. To utilizing the capability of the self-attention mechanism in transformer in capturing global features [45], [46], networks with the self-attention mechanism were designed to classify interference signals [47].

However, the current DL-based methods for interference recognition focus on changing model structures to obtain gains and ignore the effect of different transform domain forms of interference on the performance, resulting in low recognition performance. Drawing inspiration from multimodal learning, we consider the various transformed forms of the signal as distinct modalities in the context of multimodal learning, thereby enabling networks to enhance performance. Multimodal learning integrates information from different sources, such as images, text, and audio, to create a more comprehensive and holistic understanding of the data. In this paper, we refer to the transform-domain forms of signal (e.g., time-frequency domain information, frequency domain sequences, artificial features) as signal modalities. To this end, we propose adaptive multi-modal networks (AMN), which combine different modal information as the network input and intelligently select the appropriate modal information for processing, ensuring recognition accuracy while reducing the computational costs of multi-modal information. The contributions of this paper can be summarized as follows.

1) We break away from the traditional separate view of feature-based methods and DL-based methods by using artificial features as a special modal information as the input to deep neural networks. In fact, artificial features can be regarded as a refined modality that can

TABLE I
COMPARISON BETWEEN PROPOSED AMN AND CONVENTIONAL AS WELL AS EXISTING DL-BASED METHODS

Methods	Refs.	Artificial features	Automatic features selection by networks	Modal information selection	Key points
Feature-based	[28]– [30]	√	×	×	Require expert knowledge with hand-crafting features
Existing DL-based	[37]– [47]	×	√	×	Leverage the powerful feature extraction capabilities of DL and trade off prediction accuracy with complexity
The proposed	AMN	√	√	√	Integrate DL with feature-based method and balance prediction accuracy and complexity by MISM

provide complementary information to DL approaches for wireless interference identification. The performance of artificial features as well as different modal information, for deep learning networks is analyzed. To the best of our knowledge, this is the most extensive study on different modal information for DL-based interference recognition.

2) In order to extract features from different modal information, this paper adopts a network structure with joint convolution and transformer, which allows the network to capture both local and global features. To reduce the complexity of multi-headed self-attention in transformer, a low-complexity dual transformer module (DTM) is proposed. Furthermore, we propose an adaptive gradient modulation (AGM) strategy during multimodal training, resulting in better fusion performance.
3) To reduce the computational overhead of processing multiple modalities simultaneously, we propose modal information selection mechanism (MISM). This MISM applies reinforcement learning methods to construct novel reward functions that encourage the network to select the most appropriate modal information based on the current input samples rather than using all modal information, which can effectively reduce the computational overhead.
4) Extensive experiments verify that utilizing multiple interference modalities effectively improves recognition accuracy compared to the traditional uni-modal approaches. The proposed AMN has advantages in terms of both recognition accuracy and computational complexity, achieving a balance between accuracy and complexity.

Table I offers a brief comparison of the proposed AMN with conventional feature-based methods and existing DL-based methods. A feature-based approach typically consists of two components, namely, feature extraction and classification. The common manually made features includes high order statistics, cyclic feature, and so on. Support vector machine or decision tree are generally adopted as the classifier. Existing DL-based approaches by contrast leverage deep neural networks to automatically extract features from the input data, eliminating the need for manual feature engineering. The proposed AMN combines the advantages of both feature-based and DL-based approaches, and an MISM is seamlessly incorporated in the AMN for choosing the most appropriate modal information conditioned on the inputs.

The rest of the paper is organized as follows. Section II briefly defines the wireless interference recognition and commonly encountered interfering signals. Section III details the proposed AMN. Section IV performs simulations and analyzes the advantages of the proposed algorithm, in terms of accuracy and complexity. Section V summarizes this paper.

II. PROBLEM DEFINITION AND INTERFERENCE SIGNAL MODEL

The interference signal emitted by non-cooperative party, denoted as $s(t)$, reaches the receiver after passing through the wireless channel. The received signal $r(t)$ can be mathematically expressed as

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

where $h(t)$ denotes the unit impulse response of the wireless channel, the notation '$*$' denotes convolution operation, and $n(t)$ refers to the additive white Gaussian noise which is independent of the interference signal $s(t)$.

A. Problem Definition

The objective of wireless interference recognition is to blindly recognize the category of interference. Assume that there are M categories of $s(t)$, namely, $s(t) \in Y = \{y_i\}_{i=1}^{M}$, where Y is the candidate pool of interfering signals, and y_i corresponds to the i-th interference category. The final determination of the category index $i^\star$ can be obtained by the maximum-a-posterior (MAP) criterion, which can be formulated as

$$i^\star = \arg \max_{i \in \{1,2,\cdots,M\}} G\big(y_i | r(t)\big), \tag{2}$$

where $G\big(y_i|r(t)\big)$ is the conditional probability distribution of y_i given the observed $r(t)$.

B. Interference Signal Model

Non-cooperative parties frequently employ various patterns of interference signals to target wireless communication systems. These interference patterns can be categorized into different types, such as aimed interference, partial band noise (PBN) interference, comb interference and sweeping interference.

1) Aimed interference encompasses signals like modulated signals and continuous wave (CW), among others. Binary frequency shift keying (BFSK) is a common type of modulation used in interference signals. It employs different frequencies

to represent distinct information. Specifically, the transmission of binary bits 0 and 1 in BFSK is achieved by utilizing two frequencies, namely, f_p and f_q, respectively. The mathematical model for BFSK can be formulated as

$$s(t) = \sqrt{P_J}\, m_J(t)\, e^{\mathrm{j}(2\pi f_c t + \theta_J)}, \tag{3}$$

where P_J is the transmit power, $\mathrm{j} = \sqrt{-1}$, and $m_J(t)$ is the baseband modulated signal, while f_c and θ_J are the carrier frequency and initial phase of BFSK interference, respectively. Here θ_J is uniformly distributed within $[0,\, 2\pi]$, and $m_J(t)$ can be expressed as

$$m_J(t) = \cos\left(2\pi f_p t\, m(t) + 2\pi f_q t\, \overline{m}(t)\right), \tag{4}$$

where $m(t)$ represents the unipolar digital baseband signal, and the opposite of $m(t)$ is denoted as $\overline{m}(t)$.

BPSK is another widely used modulation interference, which represents binary information by employing two different phases, and the mathematical model of BPSK interference can be formulated as

$$s(t) = \sqrt{P_J}\, e^{\mathrm{j}(2\pi f_c t + \theta_P + \theta_J)}, \tag{5}$$

where θ_P denotes the modulation phase, which typically takes the values of 0 or π.

CW interference can be modeled as a signal with a fixed frequency, which can be expressed as

$$s(t) = \sqrt{P_J}\, e^{\mathrm{j}(2\pi f_c t + \theta_J)}. \tag{6}$$

2) PBN interference is a form of interference that operates in the frequency domain. It utilizes bandlimited noise to suppress the target frequency band, and it exhibits noise characteristics in the time domain. The mathematical model of PBN interference can be represented as

$$s(t) = U_n(t)\, e^{\mathrm{j}(2\pi f_c t + \theta_J)}, \tag{7}$$

where $U_n(t)$ is the baseband bandlimited noise, and the interference bandwidth is defined to be equal to the bandwidth of $U_n(t)$.

3) Multi-tone (MT) interference involves combining multiple independent CW signals with different frequencies and phases. Mathematically it can be expressed as

$$s(t) = \sum_{l=1}^{N_T} \sqrt{P_{J_l}}\, e^{\mathrm{j}\left(2\pi f_{c_l} t + \theta_{J_l}\right)}, \tag{8}$$

where N_T is the number of tones, P_{J_l} is the power of the l-th tone, f_{c_l} represents the center frequency of the l-th tone, and θ_{J_l} denotes the initial phase of the l-th tone.

4) Unlike CW interference, which consists of a single tone at a fixed frequency, frequency modulation (FM) interference spans a range of frequencies during the modulation process. In FM interference, the carrier frequency is controlled by a specific modulating signal. At a specific moment, FM interference resemble CW interference, while FM interference is a band-limited form of interference that covers a certain frequency range over a period of time.

In particular, the carrier frequency of linear frequency modulation (LFM) interference linearly changes with time, and the LFM interference can be mathematically formulated as

$$s(t) = \sqrt{P_J}\, e^{\mathrm{j}\left(2\pi\left(f_L + \frac{f_H - f_L}{2T_{\mathrm{sw}}} t\right) t + \theta_J\right)}, \tag{9}$$

where T_{sw} is the sweep period, and f_L and f_H are the start and cut-off frequencies of the interference band, respectively. The sweep bandwidth W_{sw} can be calculated as $W_{\mathrm{sw}} = f_H - f_L$, and the center frequency can be obtained as $f_c = \frac{f_H - f_L}{2}$.

Similarly, the carrier frequency of sinusoid frequency modulation (SFM) interference changes with time in a cosine manner. Mathematically, SFM can be represented as

$$s(t) = \sqrt{P_J}\, e^{\mathrm{j}2\pi\left(f_c t + K_{\mathrm{FM}} \int_0^t \cos(2\pi f_m \tau) d\tau\right)}, \tag{10}$$

where f_m is the modulation frequency of the baseband modulated signal, and K_{FM} is the frequency modulation proportional constant. In (10), θ_J is omitted for simplification.

III. The Proposed Method

Our AMN exploits modal information from various transformation domains of the interference signals. These diverse modalities contain artificial feature (AF), time-frequency image (TFI), frequency sequence (FS) and differential sequence (DS). For each modal information, the AMN separately extracts the information and then interacts with the modal information through the information fusion module. This process allows the network to effectively combine and integrate the different modalities. Furthermore, the MISM plays a crucial role in selecting and prioritizing the most relevant modal information, allowing the network to allocate its computational resources more effectively and make better decisions based on the prioritized information. Consequently, the network's performance and processing efficiency are enhanced. We now provide a detailed explanation of this proposed AMN.

A. Overall Structure of AMN

The overall framework of the proposed AMN is illustrated in Fig. 1. As can be seen from Fig. 1, the proposed network comprises four main parts. (i) Multi-modal information preprocessing: In this stage, the interference signals undergo signal pre-processing to obtain a multi-modal representation. The modal information in this part includes AF, FS, DS and TFI of the interference signals. (ii) Modal information extraction module: This module consists of four separate branches, each dedicated to extracting features from a specific modality. Each branch incorporates convolutional layers and transformer layers. The convolutional operations enhance the local feature extraction capability, while the multi-head self-attention (MSA) mechanism provides the global feature extraction capability. (iii) Information fusion module: This module facilitates the interaction of information among different modal features. By learning the complementarity between modal features, it helps to improve the accuracy of classification. (iv) MISM: This module selects and activates the most appropriate modal processing branch for the current input signal, assisting the

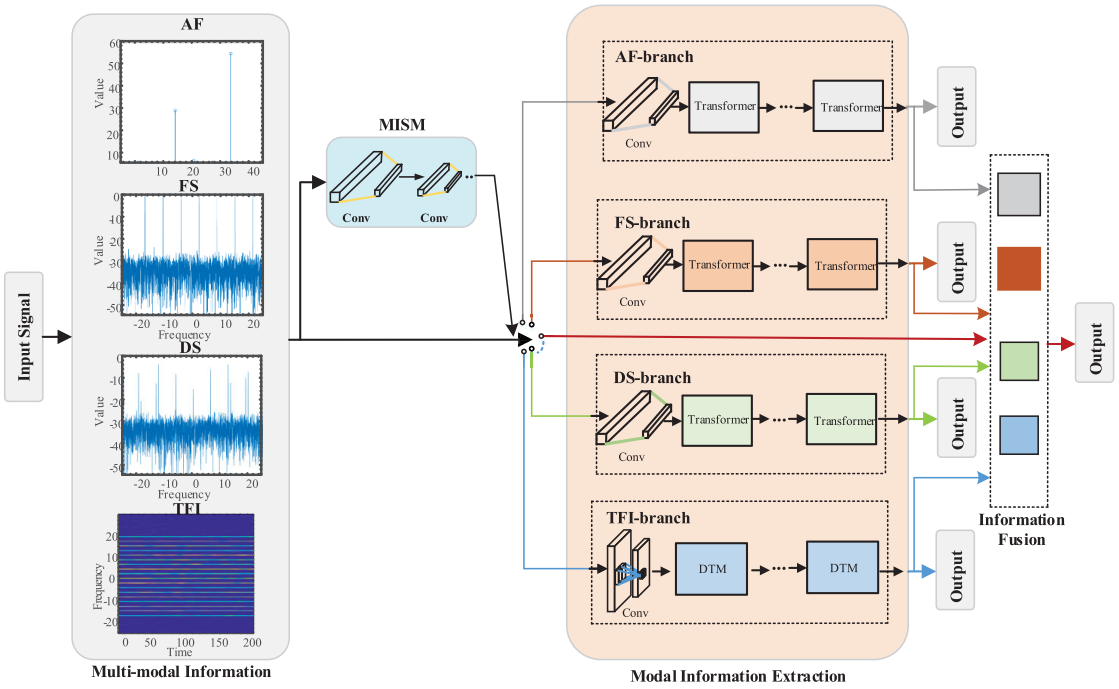


Fig. 1. Pipeline of the AMN. The input to the AMN contains four modal information, namely, AF, FS, DS and TFI, of interference signals. Our model consists of four modules: multi-modal information pre-processing, MISM, modal information extraction, and information fusion. Multi-modal information pre-processing extracts four types of information. MISM, taking all or partial modal information as input, selects and activates the most appropriate modal processing branch for the current input signal. Modal information extraction, taking all the modal information as input, is responsible for conducting feature extraction of the corresponding modal information branch. Information fusion module is adopted to thoroughly explore and fuse information of different modal information.

network in determining the optimal modal information input for different samples of interference signals. This effectively reduces the computational complexity of the network, while maintaining accuracy.

B. Multi-Modal Information Pre-Processing

The multi-modal information pre-processing module is responsible for extracting the four modal features, namely, AF, TFI, FS and DS, from the interference signals. By obtaining these modal representations, the network can leverage the complementarity between different modalities and enhance its overall recognition performance.

1) AF Modal Information: Nine AFs are elaborately designed to enhance the representation of interference signals and provide valuable insights into key properties of these signals. By treating AF as the distinct signal modal information, their characteristics can be leveraged to improve the analysis and understanding of the signals by the AMN.

Time-domain kurtosis (TDK) can be formulated as

$$\mathrm{TDK} = \frac{\mathbb{E}\left((r_{\mathrm{re}} - \mu_{\mathrm{re}})^4\right) + \mathbb{E}\left((r_{\mathrm{im}} - \mu_{\mathrm{im}})^4\right)}{(\sigma_{\mathrm{re}}^4 + \sigma_{\mathrm{im}}^4)}, \tag{11}$$

where $\mathbb{E}(\cdot)$ denotes the expectation operator, r_{re} and r_{im} represent the real and imaginary parts of the received signal, respectively. The mean values of r_{re} and r_{im} are denoted as μ_{re} and μ_{im}, and the corresponding variances are denoted as σ_{re}^2 and σ_{im}^2, respectively.

The 3 dB bandwidth factor $\mathrm{BF}_{\mathrm{3dB}}$ can be written as

$$\mathrm{BF}_{\mathrm{3dB}} = \frac{W_{\mathrm{3dB}}}{W_{\mathrm{cogn}}}, \tag{12}$$

where W_{3dB} and W_{cogn} represent the 3 dB bandwidth of the signal and the cognitive bandwidth, respectively.

Jamming detection bandwidth factor (JDBF) can be formulated as

$$\mathrm{JDBF} = \frac{W_j}{W_{\mathrm{cogn}}}, \tag{13}$$

where W_j indicates the estimated interference bandwidth.

Frequency-domain kurtosis (FDK) can be formulated as

$$\mathrm{FDK} = \frac{\mathbb{E}\left((P(k) - \mu_P)^4\right)}{\sigma_P^4}, \tag{14}$$

where $P(k)$ is the power spectral density, also known as FS, of the received signal, while μ_P and σ_P are the mean and standard deviation of $P(k)$, respectively.

Average spectral flatness coefficient (ASFC) can be written as

$$\mathrm{ASFC} = \sqrt{\frac{1}{L_a}\sum_{k=0}^{L_a-1}\left(P_p(k) - \overline{P_p}\right)^2}, \tag{15}$$

$$P_p(k) = P(k) - \frac{1}{L_s}\sum_{i=0}^{L_s-1} P^{\mathrm{Cir}}(k+i), \tag{16}$$

where L_a and L_s are the predefined lengths of ASFC and sliding window, respectively, while $P^{\mathrm{Cir}}(k+i)$ is the circular shift sequence of $P(k)$, and $\overline{P_p}$ denotes the mean of $P_p(k)$.

The ASFC serves as an indicator for determining whether a significant impulsive component exists in the spectrum.

Square spectrum bandwidth factor (SSBF) can be expressed as

$$\mathrm{SSBF} = \frac{W^s_{3\mathrm{dB}}}{W_{\mathrm{cogn}}}, \tag{17}$$

where $W^s_{3\mathrm{dB}}$ is the 3 dB bandwidth of the squared received signal.

Second spectrum kurtosis (SSK) can be written as

$$\mathrm{SSK} = \frac{\mathbb{E}\left((P_s - \mu_s)^4\right)}{\sigma_s^4}, \tag{18}$$

where P_s is the power spectrum obtained from the fourth power of the received signal, while μ_s and σ_s are the mean and standard deviation of P_s, respectively.

Quartic spectrum bandwidth factor (QSBF) can be written as

$$\mathrm{QSBF} = \frac{W^q_{3\mathrm{dB}}}{W_{\mathrm{cogn}}}, \tag{19}$$

where $W^q_{3\mathrm{dB}}$ refers to the 3 dB bandwidth obtained from the fourth power of the received signal.

Differential signal spectrum kurtosis (DSSK) can be expressed as

$$\mathrm{DSSK} = \frac{\mathbb{E}\left((P_q - \mu_q)^4\right)}{\sigma_q^4}, \tag{20}$$

where P_q is the power spectrum of the differential signal, while μ_q and σ_q are the mean and standard deviation of P_q, respectively.

The temporal domain amplitude distribution of the input signal can be represented by the definition of the TDK. The value of the TDK reflects the sharpness of the distribution or the level of concentration of the data around its center. A higher TDK value suggests a sharper distribution, while a lower TDK value indicates a flatter distribution.

The concept behind defining frequency domain parameters is to capture the key characteristics of the interference signal in a concise manner, aiming to reduce computational complexity. This can be achieved by considering commonly used frequency domain parameters such as signal bandwidth and ripple. Additionally, the distribution of frequency domain components can be reflected through parameters like FDK.

Considering that the bandwidth is a crucial and easily understandable characteristic of a signal in frequency domain analysis, it is reasonable to prioritize the bandwidth factor when selecting frequency domain features for an interference signal. To distinguish between broadband and narrowband jamming, it is possible to define and extract a parameter known as the JDBF and $\mathrm{BF}_{3\mathrm{dB}}$ from the jamming signal.

To assess the degree of energy concentration of an interfering signal in the frequency domain and determine whether it contains an impulse, we can use the ASFC. This measure helps to reflect this characteristic by quantifying the variability or spread of the impulse component within the frequency spectrum of the signal.

These nine features are concatenated together to form a one-dimensional AF modal information as $X_A \in \mathbb{R}^{L_{a0}}$, where $L_{a0} = 9$ is the length of AF. These AF modal information can provide valuable insights and contribute to a more comprehensive analysis of the signals, leading to enhanced understanding and potentially better decision-making.

2) TFI Modal Information: The TFI reflects the macroscopic characteristics of the signal's frequency changes over time. To capture the time-frequency characteristics of the interference signals, we employ the short-time Fourier transform (STFT), formulated as

$$X_T = \left| \sum_{n=0}^{N_{\mathrm{STFT}}-1} r(n) w^{\mathrm{H}}(n-m)\, e^{-\mathrm{j}\frac{2\pi k n}{N_{\mathrm{STFT}}}} \right|^2, \tag{21}$$

where m and k denote the discrete indices of time and frequency, respectively, N_{STFT} is the number of points in STFT, and the window function $w(n)$ used is the Hamming window, while $w^{\mathrm{H}}(n)$ denotes conjugation of $w(n)$.

The modal TFI, denoted as $X_T \in \mathbb{R}^{H_{t0} \times W_{t0} \times C_{t0}}$, can be converted into RGB images for input, where $H_{t0} \times W_{t0}$ represents the resolution of the image, and C_{t0} is the dimension of feature channels.

3) FS Modal Information: FS, denoted as $X_F \in \mathbb{R}^{L_{f0}}$, describes the microscopic characteristics of the signal spectrum. Specifically, the sequence X_F is the squared moduli of the N_{FFT}-points fast Fourier transform (FFT) of $r(n)$:

$$X_F = \frac{1}{N_{\mathrm{FFT}}} \left| \sum_{n=0}^{N_{\mathrm{FFT}}-1} r(n)\, e^{-\mathrm{j}\frac{2\pi k n}{N_{\mathrm{FFT}}}} \right|^2, \tag{22}$$

for $0 \le k \le N_{\mathrm{FFT}} - 1$. In this case, the length $L_{f0} = N_{\mathrm{FFT}}$.

4) DS Modal Information: DS exhibits frequency invariance of the signal over time, which can be obtained as

$$X_D = \frac{1}{N_{\mathrm{FFT}}} \left| \sum_{n=0}^{N_{\mathrm{FFT}}-1} d(n)\, e^{-\mathrm{j}\frac{2\pi k n}{N_{\mathrm{FFT}}}} \right|^2, \tag{23}$$

for $0 \le k \le N_{\mathrm{FFT}} - 1$, where the differential signal $d(n)$ is calculated from the received signal $r(n)$ according to

$$d(n) = r(n + \Delta)\, r^{\mathrm{H}}(n). \tag{24}$$

The offset Δ is set to 256 in this paper. DS can be expressed as a one-dimensional (1D) sequence $X_D \in \mathbb{R}^{L_{d0}}$, with length $L_{d0} = N_{\mathrm{FFT}}$.

C. Modal Information Extraction Module

This module comprises four branches: the AF branch, TFI branch, FS branch, and DS branch. CNNs are effective in capturing local features by sliding convolutional kernels over the input but they have limitation in capturing global features. Transformer networks by contrast enjoy significant superiority in capturing global features. Therefore, we leverage both the global feature capturing capabilities of transformer networks and the local extraction properties of CNNs.

1) Feature Extraction for AF: The feature extraction process for AF, as depicted in the upper part of Fig. 1, involves stacking 1D convolutional layers and transformer layers. In particular, the processing of the convolutional layers can be expressed as

$$X_{A'} = f_{\text{1d-conv}}(X_A; \varphi), \tag{25}$$

where $f_{\text{1d-conv}}(\cdot;\varphi)$ represents a series of 1D convolution operations, and $X_{A'} \in \mathbb{R}^{L_a \times C_a}$ is the output with length L_a and C_a feature channels, while φ stands for trainable parameters of 1D convolutional layers. After the convolutional operations, $X_{A'}$ is further processed through stacked transformer layers.

Transformer is composed of an MSA module and a feed-forward network (FFN). The MSA module is responsible for capturing global feature information, while the FFN facilitates information interaction between feature channels. Specifically, the MSA consists of h heads, and each head calculates the information in the same manner, expressed as

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_t}}\right)V. \tag{26}$$

In (26), $(\cdot)^{\mathrm{T}}$ denotes the transpose operator, $d_t = \frac{C_a}{h}$, the query matrix $Q \in \mathbb{R}^{L_a \times d_t}$, the key matrix $K \in \mathbb{R}^{L_a \times d_t}$ and the value matrix $V \in \mathbb{R}^{L_a \times d_t}$ are given by $Q = X_{A'}W_q$, $K = X_{A'}W_k$ and $V = X_{A'}W_v$, respectively, where $W_q, W_k, W_v \in \mathbb{R}^{C_a \times d_t}$ are learnable matrices. To ensure the diversity of extracted features, the MSA module computes information using a total of h heads and then combines them together. This process, which helps to capture different aspects of the input and enhances the overall representation, can be expressed as

$$A_i = \text{Attention}\big(Q_i, K_i, V_i\big),\ 1 \le i \le h, \tag{27}$$

$$X_{A''} = \text{concat}\big(A_1, A_2, \cdots, A_h\big)W, \tag{28}$$

where $A_i \in \mathbb{R}^{L_a \times d_t}$ is the i-th head, $X_{A''} \in \mathbb{R}^{L_a \times C_a}$ is the output of the MSA, and $W \in \mathbb{R}^{C_a \times C_a}$ is a learnable matrix

To facilitate inter-channel information interaction, the MSA is followed by the FFN. The FFN, which consists of two layers of multi-layer perceptron (MLP), can be expressed as

$$\text{FFN}\big(X_{A''}\big) = \omega\big(X_{A''}W'\big)W'', \tag{29}$$

where the activation function ω is chosen to be the Gaussian error linear unit (GELU), $W' \in \mathbb{R}^{C_a \times (r \cdot C_a)}$ and $W'' \in \mathbb{R}^{(r \cdot C_a) \times C_a}$ are trainable matrices, in which r is referred to as the expanding ratio.

To recap, the processing of AF is summarized as follows

$$\begin{cases} X_{A'}^{(0)} = f_{\text{1d-conv}}\big(X_A;\varphi\big), \\ X_{A''}^{(l)} = \text{MSA}\left(\text{NL}\big(X_{A'}^{(l-1)}\big)\right) + X_{A'}^{(l-1)}, \\ X_{A'}^{(l)} = \text{FFN}\left(\text{NL}\big(X_{A''}^{(l)}\big)\right) + X_{A''}^{(l)}, \end{cases} \tag{30}$$

where $X_{A''}^{(l)} \in \mathbb{R}^{L_a \times C_a}$ and $X_{A'}^{(l)} \in \mathbb{R}^{L_a \times C_a}$ are the outputs of the l-th layer MSA and FFN, respectively, and $\text{NL}(\cdot)$ represents the layer normalization. Let the number of transformer layers for the AF branch be N_a. Then the final output of the AF branch is $X_{A'}^{(N_a)}$.

2) Feature Extractions for FS and DS: As can be seen from the middle part of Fig. 1, the FS and DS branches adopt the same structure as the AF branch. Therefore, they have similar feature extraction processes as that of AF, and we denote their output features by $X_{F'}^{(N_f)} \in \mathbb{R}^{L_f \times C_f}$ and $X_{D'}^{(N_d)} \in \mathbb{R}^{L_d \times C_d}$, respectively, where N_f and N_D are the numbers of transformer layers in the FS and DS branches, and L_f and L_d denote the output lengths of FS and DS, respectively, while C_f and C_d represent the corresponding feature dimensions.

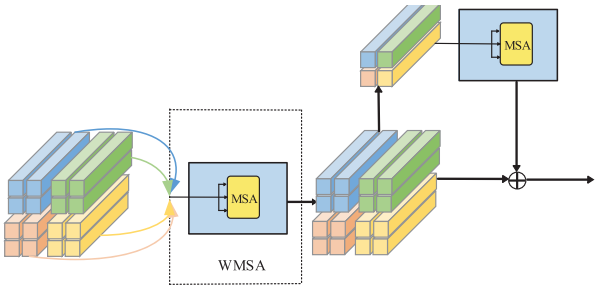


Fig. 2. Structure of DTM. Different colors signify various windows, and WMSA conducts MSA operations within each of these designated segments.

3) Feature Extraction for TFI: As shown in the bottom part of Fig. 1, several 2D convolutional layers are employed to extract the details and local features of X_T, which can be expressed as

$$X_{T'} = f_{\text{2d-conv}}\big(X_T;\eta\big), \tag{31}$$

where $f_{\text{2d-conv}}(\cdot;\eta)$ represents this feature extraction operation with learnable parameters η, and $X_{T'} \in \mathbb{R}^{H \times W \times C_t}$ denotes the output features.

To adapt the input form for the transformer, $X_{T'}$ can be divided into multiple feature blocks, which are sequentially concatenated to form a sequence. Specifically, the size of each feature block is defined as $p \times p$, and $X_{T'} \in \mathbb{R}^{H \times W \times C_t}$ can be evenly partitioned into $E = W \times H/p^2$ feature blocks. The i-th feature block is denoted as $x_i \in \mathbb{R}^{C_{t'}}$, where $C_{t'}$ is given by $C_{t'} = p^2 \times C_t$. These feature blocks are concatenated to form a sequence $X_p = \big[x_1, x_2, \cdots, x_E\big] \in \mathbb{R}^{E \times C_t'}$. The computational complexity $\Psi(\cdot)$ of MSA can be written as

$$\Psi(\text{MSA}) = 4EC_{t'}^2 + 2E^2C_{t'}. \tag{32}$$

It can be seen that the computational complexity of MSA and FFN grows quadratically with the number of feature blocks E. When the size of the TFI input is large, the number of feature blocks E increases dramatically, resulting in a substantial computational cost for the MSA operation.

To address the aforementioned challenge, a low-complexity dual transformer module (DTM) is introduced. Drawing inspiration from [46], the DTM incorporates a window-based multi-head self-attention (WMSA) operation, which partitions the input into multiple windows to effectively reduce computational complexity.

As illustrated in Fig. 2, the DTM consists of both WMSA and MSA operations. The WMSA focuses on capturing characteristics within each window, while the MSA aims to extract global characteristics between windows. The input tensor, denoted as X_p, is divided into E/r^2 windows, where each window has a size of r^2. Thus, X_p can be written as $X_p \in \mathbb{R}^{E/r^2 \cdot r^2 \times C_t'}$. MSA operations are performed in each window. Since r^2 is considerably smaller than E, the WMSA operation significantly reduces the computational complexity, compared to the standard MSA. This approach enables the DTM to achieve a low-complexity solution.

However, the WMSA operations performed within each window pose a challenge in terms of effectively interacting with information between windows, which inevitably impacts the prediction performance. In other words, WMSA operations

deviate from the original intention of designing global feature extraction. To address this issue, the MSA is employed to enhance the flow of information between windows. A representative of each window can be selected, resulting in a total of E/r^2 representatives. For ease of implementation, a representative is randomly selected from each window in this paper. These representatives can interact with each other through the MSA operation. Finally, each representative is fused back into its original window by element-wise summation, facilitating the exchange of information between different windows. During the summation, each representative is broadcasted (copied) to the widow size. The computational complexity of this low-complexity DTM can be written as

$$\Psi(\mathrm{DTM}) = 4\left(1+\frac{1}{r^2}\right)EC_{t'}^2 + 2\,r^2\left(1+\frac{E}{r^6}\right)EC_{t'}. \tag{33}$$

Let the number of DTMs be N_t. We obtain the final output of the TFI branch as $X_{T'}^{(N_t)} \in \mathbb{R}^{L_t \times C_t}$, where L_t and C_t are the length and feature dimension of the output, respectively. When E is set to 600, r to 5, and $C_{t'}$ to 64, the ratio $\Psi(\mathrm{DTM})/\Psi(\mathrm{MSA})$ is found to be 8%, highlighting the computational complexity advantage that DTM exhibits.

D. Information Fusion Module

The output features of the AF, FS, DS and TFI branches are denoted as $X_{A'}^{(N_a)}$, $X_{F'}^{(N_f)}$, $X_{D'}^{(N_d)}$ and $X_{T'}^{(N_t)}$, respectively. To obtain the recognition probability for each branch, the obtained features are fed into the globe average pooling (GAP) operation along spatial dimension and the fully connected layers, followed by a softmax function. For example, the output probability of the TFI-branch can be expressed as

$$G_t = \mathrm{FC}\left(X_{T'}^{(N_t)}\right), \tag{34}$$

where G_t is the recognition probability obtained from the TFI-branch, and we have omitted the GAP function. Similarly, the probabilities corresponding to the AF, DS, and FS can be obtained as G_a, G_d and G_f, respectively.

We employ two straightforward approaches, namely, probabilistic fusion (PF) and feature fusion (FF), to fuse the information from the four modalities. Specifically, the PF approach with equal weighting can be expressed as

$$G_{\mathrm{PF}} = \frac{1}{4}\left(G_a + G_t + G_d + G_f\right), \tag{35}$$

where G_{PF} denotes the probability after fusion.

The FF method first combines the features from different branches by concatenating them. The concatenated features are then passed through a fully connected layer with a softmax function, to calculate the final probability according to

$$G_{\mathrm{FF}} = \mathrm{FC}\left(X_{A'}^{(N_a)}, X_{F'}^{(N_f)}, X_{D'}^{(N_d)}, X_{T'}^{(N_t)}\right). \tag{36}$$

In (34) and (36), we have omitted the softmax function.

1) Gradient Imbalance Phenomenon: Gradient imbalance in modal optimization is an inevitable phenomenon [48]. To illustrate this, consider the PF method as an example. For the FF method, the derivation and conclusion are similar.

Denote the four modal extraction functions as $X_{A'}(\cdot|\theta_A)$, $X_{F'}(\cdot|\theta_F)$, $X_{D'}(\cdot|\theta_D)$ and $X_{T'}(\cdot|\theta_T)$, where θ_A, θ_F, θ_D and θ_T are learnable parameters for the AF, FS, D and TFI, respectively. Let the weights of the corresponding fully connected classifiers be W_A, W_F, W_D and W_T. The PF probability output for the k-th input x_k can be expressed as

$$\begin{aligned} G(x_k) = {} & W_A X_{A'}(x_k|\theta_A) + W_F X_{F'}(x_k|\theta_F) \\ & + W_F X_{D'}(x_k|\theta_D) + W_T X_{T'}(x_k|\theta_T). \end{aligned} \tag{37}$$

For simplicity, the bias terms are omitted. The true label of the input x_k is denoted as y_k, and the cross-entropy loss function $\mathcal{L}$ is expressed as

$$\mathcal{L} = -\frac{1}{N}\sum_{k=1}^{N}\log\left(\frac{e^{G(x_k)_{y_k}}}{\sum_{m=1}^{M} e^{G(x_k)_m}}\right) \tag{38}$$

where N and M are the numbers of samples and categories, respectively, while $G(x_k)_m$ denotes the output for class m.

The parameters associated with each modality are updated using gradient descent. We take the TFI update as an example (other modalities are similar), which can be expressed as

$$\begin{aligned} W_T^{t+1} &= W_T^t - \mu\frac{\partial\mathcal{L}\left(W_T^t\right)}{\partial W_T} \\ &= W_T^t - \mu\frac{1}{N}\sum_{k=1}^{N}\frac{\partial\mathcal{L}}{\partial G\left(x_k\right)}X_{T'}\left(x_k|\theta_T\right), \end{aligned} \tag{39}$$

$$\begin{aligned} \theta_T^{t+1} &= \theta_T^t - \mu\frac{\partial\mathcal{L}\left(\theta_T^t\right)}{\partial\theta_T} \\ &= \theta_T^t - \mu\frac{1}{N}\sum_{k=1}^{N}\frac{\partial\mathcal{L}}{\partial G\left(x_k\right)}\frac{\partial\left(W_T^t X_{T'}\left(x_k|\theta_T\right)\right)}{\partial\theta_T^t}, \end{aligned} \tag{40}$$

where the superscript ^{t} denotes the iteration index, W_T^t is the weight at the t-th iteration, and μ is the learning rate. It can be observed that the update of each modal parameters is largely independent of the other modalities, except for the term $\frac{\partial\mathcal{L}}{\partial G(x_k)}$. This implies that when a particular modality exhibits high confidence, it will play a dominant role in fusion process. As a result, the network will prioritize this modality, paying more attention to this modality, and potentially ignore the other modalities, making the other modalities insufficiently updated.

2) Adaptive Gradient Modulation: Our propose AGM strategy specifically addresses the aforementioned gradient imbalance problem and enables the network to updating different modalities simultaneously. We take the single modality $v \in \{A, F, D, T\}$ update as an example. The stochastic gradient descent (SGD) can be expressed as

$$\theta_v^{t+1} = \theta_v^t - \mu\frac{\partial\mathcal{L}\left(\theta_v^t\right)}{\partial\theta_v} = \theta_v^t - \mu g\left(\theta_v^t\right), \tag{41}$$

where $g\left(\theta_v^t\right)$ denotes the full gradient. An unbiased estimation of $g\left(\theta_v^t\right)$ can be formulated as

$$g\left(\theta_v^t\right) = \frac{1}{B_t}\sum_{x\in B_t}\frac{\partial l_{B_t}\left(\theta_v^t\right)}{\partial\theta_v}, \tag{42}$$

where B_t denotes mini-batch, and l_{B_t} is the cross-entropy loss function over B_t.

We define a performance factor χ_v^t of modality v, which reflects the relative convergence speed of the individual modality v during joint multimodal training, and it is expressed as

$$\chi_v^t = \frac{\sum_{k\in B_t}\gamma_{v,k}^t}{\chi_{mean}^t}, \tag{43}$$

$$\gamma_{v,k}^t = \sum_{i=1}^{M}\mathbb{I}_{i=y_k}\,\mathrm{softmax}\left(W_v^t X_{v'}\left(x_k|\theta_v^t\right)\right)_i, \tag{44}$$

where $\mathbb{I}$ denotes indicator function, $\gamma_{v,k}^t$ reflects the recognition confidence of modality v, and $\mathrm{softmax}(\cdot)_i$ denotes the i-th term of softmax output, while χ_{mean}^t represents the average convergence speed of the different modalities, formulated as

$$\chi_{mean}^t = \frac{1}{4}\sum_{k\in B_t}\left(\gamma_{A,k}^t+\gamma_{F,k}^t+\gamma_{D,k}^t+\gamma_{T,k}^t\right). \tag{45}$$

Therefore, a gradient adjustment factor at the t-th iteration corresponding to the mode v can be introduced as

$$k_v^t = \begin{cases} 1-\kappa\left(\mathrm{sigmoid}\left(\chi_v^t\right)-0.5\right), & \chi_v^t > 1,\\ 1, & \text{otherwise}, \end{cases} \tag{46}$$

where κ is a hyper-parameter, which is set to 1 in this paper. With k_v^t, the SGD update of the modal v (41) is modified as

$$\theta_v^{t+1} = \theta_v^t - \mu k_v^t g\left(\theta_v^t\right). \tag{47}$$

E. Modal Information Selection Mechanism

The fusion of the multimodal information enjoys significant superiority in terms of prediction accuracy. However, processing multiple modal information imposes higher computational cost. Our MISM addresses this issue by selecting the appropriate modal branch based on the current input sample, allowing the AMN to reduce computational complexity while maintaining high prediction accuracy. In other words, the MISM enables the network to handle multimodal data with improved efficiency and effectiveness by utilizing computational resources more wisely. To achieve this, the MISM incorporates a policy network as a selection switch to choose the suitable modal branches for each sample. This policy network is trained using reinforcement learning technique so that it learns to make optimal decisions on which modal branches to activate, based on the characteristics of the input sample.

We choose a low-overhead policy network with partial modal information as its input so that the cost of running it is negligible compared to modal extraction branches. This policy network takes partial modal information x as input and produces an output of $(K+1)$-dimensional vector as:

$$\mathbf{o} = \left[o_1,\cdots,o_k,\cdots,o_K,o_{K+1}\right] = f_{\text{policy}}(x;\phi), \tag{48}$$

where f_{policy} is the policy network with trainable parameters ϕ, o_k is the probability of executing the k-th modal branch, $1\le k\le K$, and o_{K+1} is the probability of extracting all the input modal information and performing modal fusion. An action vector is defined as $\mathbf{u} = \left[u_1,\cdots,u_k,\cdots,u_{K+1}\right]$. If $u_k = 1$, the k-th modal branch is executed, while if $u_k = 0$, the k-th branch is not executed. The action u_k is selected based on o_k, and the output distribution of the policy network is given by

$$\pi_\phi(\mathbf{u}|x) = \prod_{k=1}^{K+1}\left(1-o_k\right)^{1-u_k}o_k^{u_k}. \tag{49}$$

The selection of the modal processing branch for the current sample is determined by the action vector $\mathbf{u}$ according to

$$k^\star = \arg\max_{k\in\{1,2,\cdots,K+1\}} u_k, \tag{50}$$

where $k^\star$ denotes the selected branch. In this paper, $K=4$. The final predicted category is determined according to

$$\widehat{y} = \arg\max_{y\in\{1,2,\cdots,M\}} G\big(X_A,X_T,X_F,X_D|k^\star\big), \tag{51}$$

where $G\big(X_A,X_T,X_F,X_D|k^\star\big)\in\mathbb{R}^M$ is the prediction probability by the $k^\star$-th modal branch determined by the policy network. For simplicity, the branch index k^* is omitted below.

Depending on the chosen branch determined by the policy network, the reward signal $R(\mathbf{u})$ can be obtained, which reflects the network accuracy and computational efficiency associated with the decision. $R(\mathbf{u})$ can be expressed as

$$R(\mathbf{u}) = \begin{cases} 1-\widehat{C}(\mathbf{u}), & \widehat{y}=y,\\ \rho, & \text{otherwise}, \end{cases} \tag{52}$$

where $\widehat{C}(\mathbf{u})$ denotes the normalized computational cost, which is the ratio of the utilized computing resources, associated with action $\mathbf{u}$, to the total available computing resources. The reward function encourages the network to minimize computational costs when the predicted category is correct under the decision $\mathbf{u}$. Conversely, the reward function applies a penalty ρ when the predicted category is incorrect. The policy network can be optimized by maximizing the following expected reward:

$$\mathcal{J} = \mathbb{E}_{\mathbf{u}\sim\pi_\phi}(R(\mathbf{u})), \tag{53}$$

Since the reward function $R(\mathbf{u})$ is non-differentiable, a policy gradient method [49] is employed to perform the optimization.

F. Training and Testing Procedure

The training samples are represented as $\{X_i,y_i\}_{i=1}^N$, where X_i is the i-th training sample and y_i is the corresponding label, while N is the number of training samples. $\{X_i\}_{i=1}^N$ are processed by multimodal pre-processing to yield $\{X_{A,i},X_{T,i},X_{F,i},X_{D,i}\}_{i=1}^N$, where $X_{A,i}$, $X_{T,i}$, $X_{F,i}$ and $X_{D,i}$ are the i-th AF, TFI, FS and DS training samples, respectively. During the training process, the parameters of the AMN are updated using the cross-entropy loss function:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i\log\left(G(X_{A,i},X_{T,i},X_{F,i},X_{D,i})\right), \tag{54}$$

where $G(X_{A,i},X_{T,i},X_{F,i},X_{D,i})$ denotes the prediction probability of the AMN, which can be the function of single modality or multiple modalities, determined by the MISM module. Algorithm 1 summarizes the training and testing procedure for the proposed AMN.

Algorithm 1 Training and Testing Procedure of AMN

Input: Training data $\{X_i, y_i\}_{i=1}^{N}$, total training epochs E_p, B mini-batches in each epoch, testing data $\{X_{t_i}\}_{i=1}^{N'}$

Output: Predicted interference types $\{\widehat{y}_{t_i}\}_{i=1}^{N'}$ for test samples $\{X_{t_i}\}_{i=1}^{N'}$.

Process multimodal information $\{X_{A,i}, X_{T,i}, X_{F,i}, X_{D,i}\}_{i=1}^{N}$ from $\{X_i\}_{i=1}^{N}$ by (11)–(24);
Construct modal extraction module $X_{A'}(\cdot|\theta_A)$, $X_{F'}(\cdot|\theta_F)$, $X_{D'}(\cdot|\theta_D)$ and $X_{T'}(\cdot|\theta_T)$;
Construct modal fusion PF by (35) or FF by (36);
if *end-to-end training* **then**
  for $i = 1, 2, \cdots, E_p$ **do**
    for $j = 1, 2, \cdots, B$ **do**
      Obtain modal features $X_{A'}^{(N_a)}$, $X_{F'}^{(N_f)}$, $X_{D'}^{(N_d)}$ and $X_{T'}^{(N_t)}$, and G_{PF} or G_{FF} during forward propagation;
      Calculate gradient adjustment factor k_v^t for different modalities by (46);
      Update parameters θ_v by (47);
    end
  end
  Fix parameters of θ_A, θ_F, θ_D and θ_T;
  Construct policy network $f_{\text{policy}}(\cdot;\phi)$;
  for $i = 1, 2, \cdots, E_p$ **do**
    for $j = 1, 2, \cdots, B$ **do**
      Calculate $R(\mathbf{u})$ by (52);
      Update parameters ϕ by maximizing (53);
    end
  end
  Save all the learnable parmeters of AMN;
else
  Load the AMN model;
  for $i = 1, 2, \cdots, N'$ **do**
    Process multimodal information $\{X_{A,t_i}, X_{T,t_i}, X_{F,t_i}, X_{D,t_i}\}_{t=1}^{N'}$ from $\{X_t\}_{t=1}^{N'}$;
    Obtain predicted $\widehat{y}_{t_i}$ by $\widehat{y}_{t_i} = \arg\max_{y\in\{1,2,\cdots,M\}} G\big(X_{A,t_i}, X_{T,t_i}, X_{F,t_i}, X_{D,t_i}\big)$;
  end
end

IV. Simulation Study

Extensive experiments are conducted to validate the effectiveness of the proposed AMN. We also perform a series of simulations to assess the impact of different modalities on accuracy, evaluate the superiority of the DTM, and demonstrate the effectiveness of the MISM.

A. Simulation Setup

The simulation dataset is generated by Matlab2018. It includes seven types of interference patterns: BFSK, MT, BPSK, CW, LFM, PBN and SFM. The parameters corresponding to the center frequency and bandwidth of each interference sample are randomly varied. The interference signal is often emitted by a highly mobile unmanned aerial vehicle (UAV) or jamming vehicle, resulting in a direct path and high jamming power. This type of signal significantly impacts the cooperative communication link, leading to a Rice channel model for the interference signal. For the channel model, a single-path Rice channel with a Rice factor of 10 dB is considered.

In the experiments, the batch size of 64 and the SGD optimizer are adopted. The maximum number of epochs is set to 50. The initial learning rate is 0.001, and learning rate decays by a factor of 0.1 for every 30 epochs. The training and test samples are 1000 and 100, respectively, under each interference type and each interference-to-noise ratio (INR). The INR value varies from -20 dB to 10 dB at 2 dB intervals.

TABLE II
The Structures of Four Modalities in AMN

AF-modality	
Input	Input(9,1)
Conv1d	Conv1d (16,1,1)
Transformer	3×Transformer(MSA+FFN)
output	GAP+ Dense(7)+softmax
FS-modality	
input	Input(4096,1)
Conv1d	Conv1d (16,4,4)+BN+Relu+Maxpool(2)
Conv1d	Conv1d (16,3,2)+BN+Relu+Maxpool(2)
Transformer	Transformer(MSA+FFN)
output	GAP+ Dense(7)+softmax
DS-modality	
input	Input(3840,1)
Conv1d	Conv1d (16,4,4)+BN+Relu+Maxpool(2)
Conv1d	Conv1d (16,3,2)+BN+Relu+Maxpool(2)
Transformer	Transformer(MSA+FFN)
output	GAP+ Dense(7)+softmax
TFI-modality	
input	(3,40,40)
Conv2d	Conv2d(16,3×3,2) +BN+Relu+Maxpool(2)
Conv2d	Conv2d(16,3×3,1) +BN+Relu
DTM	2× DTM
output	GAP+ Dense(7)+softmax

B. Recognition Performance of Different Modalities

Table II summarizes the network structures of the four modalities in the AMN. The input to the AMN contains 4 multimodal information (AF, FS, DS and TFI). Conv1d$(16, 1, 1)$ represents a 1D convolutional layer with 16 feature channels, a convolutional kernel size of 1 and a stride of 1. Conv2d$(16, 3\times 3, 2)$ represents a 2D convolutional layer with 16 channels, a 3×3 convolutional kernel and a stride of 2. Maxpool(2) indicates maximum pooling with a size reduction of 2 times. Dense(7) denotes the fully connected layer with seven neurons. BN and rectified linear unit (Relu) denote the batch normalization and activation function, respectively. To extract global features, the automatic feature extraction of each type of modality contains the transformer structure. The transformer of the TFI uses the DTM to reduce the complexity.

The numbers of learnable parameters and floating-point operations (FLOPs) for different modal extraction branches are reported in Table III. It can be seen that the AF modality has the lowest number of FLOPs, compared to the other modalities, since the AF input contains only 9 features. As for the TFI modality, which involves a 2D image input, the DTM is employed to reduce the complexity.

TABLE III
COMPUTATIONAL COMPLEXITY OF EACH MODALITY

Modality	AF	FS	TFI	DS
FLOPs ($\times 10^3$)	49.5	983.15	777.33	890.99
Parameters ($\times 10^3$)	4.73	2.48	8.99	2.48

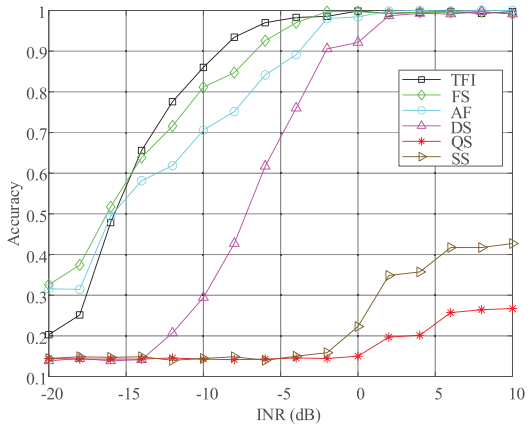


Fig. 3. Recognition accuracy of different modal information.

TABLE IV
AVERAGE RECOGNITION PERFORMANCE OF DIFFERENT MODAL INFORMATION

Modality	AF	FS	TFI	DS	QS	SS
Accuracy	77.97%	81.83%	81.70%	60.36%	17.33%	22.88%

Fig 3 depicts the recognition performance of the different modal information, where the square spectrum (SS) and quadratic spectrum (QS) of the interference signals are compared with AF, FS, DS and TFI. SS is the frequency sequence of the squared received signal, while QS is the frequency sequence of the quadratic term of the received signal. Both SS and QS can also serve as modal information for the interference signal. In general, the recognition accuracy increases as the INR increases. This is because as the INR increases, the interference signals are less affected by noise, making them easier to recognize. Clearly, the recognition performance varies across different modal information. Certain modalities are more susceptible to noise and fading, while others demonstrate better robustness. Specifically, SS and QS exhibit much lower identification capability compared to AF, FS, DS and TFI. In our experiments, we find that SS performs well in recognizing BPSK but it does not distinguish other interference signals effectively, resulting in poor overall recognition performance. Therefore, we do not use SS and QS as modal information in the subsequent experiments. For our four modalities, AF and FS outperform AS and DS, particularly at low INRs, indicating their advantageous performance under low INRs. DS achieves the worst recognition performance compared with the other three modalities. But at high INRs (greater than 0 dB), all the four modalities attain near 100% accuracy.

Table IV lists the average recognition accuracies for the different modal information. It can be seen that the recognition accuracies of FS and TFI are similar, both reaching approximately 81%. The recognition accuracy of AF reaches 77.97%, while imposing the lowest computational complexity. The average recognition accuracies of QS and SS are only 17.33% and 22.88%, respectively.

TABLE V
FLOPS OF DIFFERENT MODELS

Mothod	CNN	Resnet	ResNext	VIT	SWIN	Proposed
FLOPs ($\times 10^3$)	800.144	822.726	768.412	1040.048	777.71	777.33

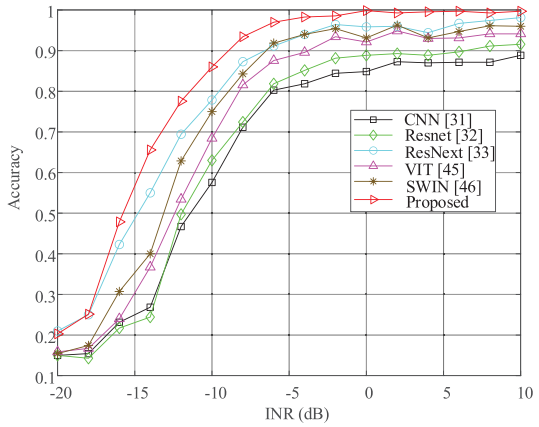


Fig. 4. Recognition accuracy comparison of different networks with TFI modal information.

TABLE VI
AVERAGE RECOGNITION PERFORMANCE OF DIFFERENT NETWORKS WITH TFI MODAL INFORMATION

Model	CNN	Resnet	ResNext	VIT	SWIN	Proposed
Accuracy	64.04%	65.95	77.37%	70.54%	73.53%	81.70%

C. Performance Comparison With Existing Methods

We now demonstrate the superiority of the proposed AMN over existing networks by evaluating the recognition performance of our AMN as well as different existing networks with the TFI modal information. The networks compared include CNN [31], ResNet [32], ResNext [33], VIT [45] and SWIN [46]. First Table V compares the computational complexity of these evaluated networks. It can be seen that the computational complexity of these models are approximately the same.

Fig. 4 depicts the recognition accuracies of different models with TFI. Observe that the proposed AMN demonstrates superior recognition performance compared to the existing models. At INR $= -10$ dB, the recognition accuracy of the proposed method surpasses those of ResNext, SWIN, VIT, Resnet and CNN by 7%, 12%, 18%, 23% and 30%, respectively. Furthermore, at high INRs, our method can reach approximately 100% accuracy, while the other methods cannot. This notable improvement can be attributed to the proposed DTM, which enables the extraction of both local and global features while effectively reducing computational complexity.

The average recognition performance of the six models with TFI are given in Table VI, where it can be seen that our AMN exhibits a higher average accuracy for wireless interference recognition compared to the other models. Furthermore, we eliminate the MSA for window information fusion from the

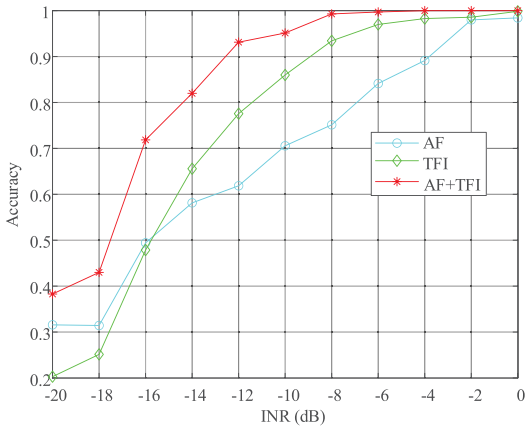


Fig. 5. Fusion performance of FF(AF+TFI) in comparison with the performance of using AF and TFI alone.

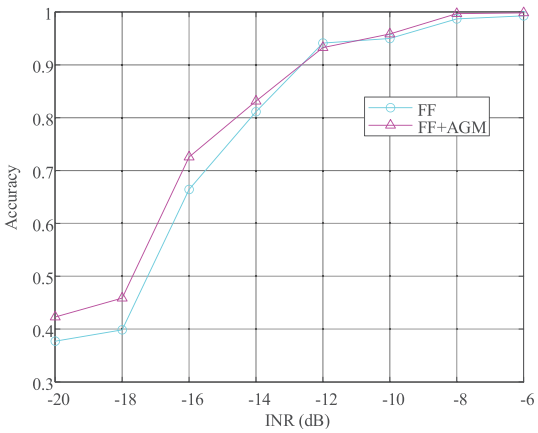


Fig. 7. Recognition accuracy of FF for all modalities (AF, FS, TFI, DS) without and with AGM.

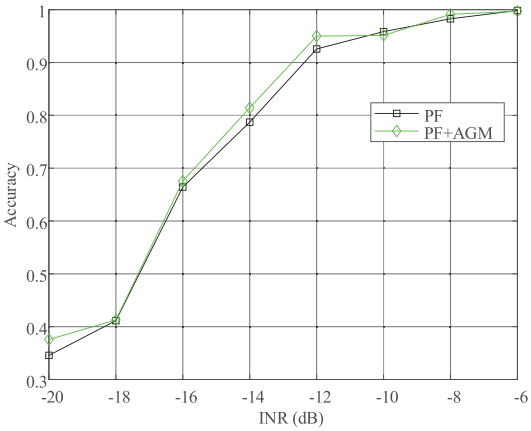


Fig. 6. Recognition accuracy of PF for all modalities (AF, FS, TFI, DS) without and with AGM.

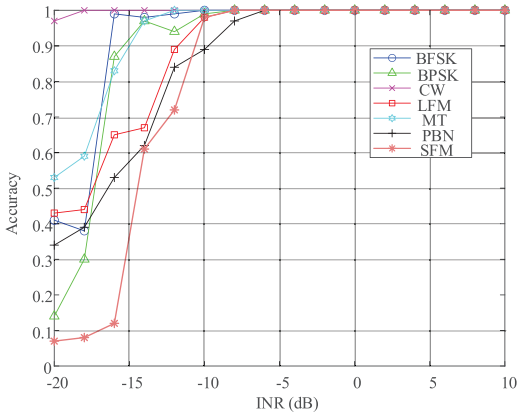


Fig. 8. Recognition accuracy of different interference patterns using FF+AGM.

DTM and observe a performance decrease of approximately 2%. This finding underscores the critical role of the MSA in facilitating effective interaction among windowed information, thereby proving its significance within the proposed DTM framework.

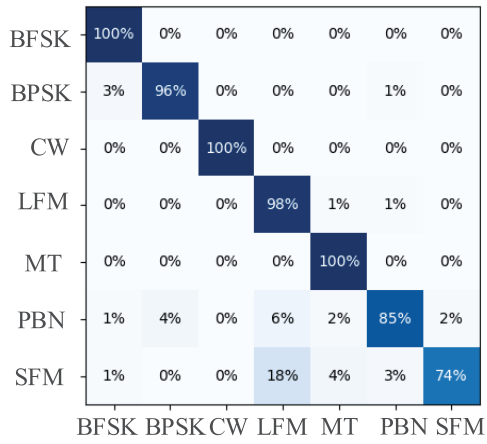
D. Fusion Performance of Modal Information

In the traditional algorithms, AF is commonly used as input for feature-based methods. We utilize AF as a modality and fuse it with other modalities to enhance the recognition performance. Fig. 5 illustrates the performance improvement achieved by fusing the AF and TFI modalities, where FF(AF+TFI) denotes the feature fusion of AF and TFI. The results show that AF can serve as valuable modal information for the interference signal and effectively enhance the recognition accuracy. Also AF modality has low complexity and does not introduce high computational cost to the fusion process. In contrast, existing methods treat AF as conventional features without considering its potential as refined information to boost the performance of DL-based methods.

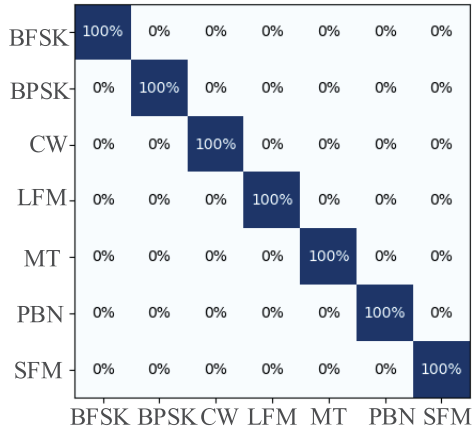
Next, we validate the effectiveness of the proposed AGM. The performance of PF and FF for all the modalities without and with the AGM are given in Figs. 6 and 7, respectively. It can be seen that the AGM strategy does improve the recognition accuracy. This performance enhancement can be attributed to the dynamic adjustment of gradients during the training process. Specifically, it increases the gradient of the under-optimized modality while suppressing the gradient of the better-optimized modality.

Fig. 8 presents the recognition accuracy of FF+AGM for each type of interference. It can be seen that the accuracy can reach 95% at INR $= -20$ dB for CW interference. However, the recognition accuracy of SFM is much poorer at low INRs, and the accuracy only reaches 95% at at INR $= -10$ dB. This is primarily due to SFM's vulnerability to be confused with LFM interference. The inherent similarity between these two FM signals leads to such confusion. Fig. 9 shows the confusion matrices for INRs of -12 dB and 0 dB. At INR $= -12$ dB, there appears to be confusion between LFM and SFM signals, while at INR $= 0$ dB, all the signals are correctly classified.

We present the performance outcomes following a two-by-two fusion process for various modalities, as depicted in Fig. 10. Observing the figure, it is evident that the combined performance matches or surpasses the best performance of the individual modality. Notably, in the scenario involving the fusion of TFI with either AF or FS, the fused performance exhibits a clear superiority. We have elucidated a fusion principle that governs the amalgamation of two operational



(a) INR $= -12$ dB



(b) INR $= 0$ dB

Fig. 9. Confusion matrix.

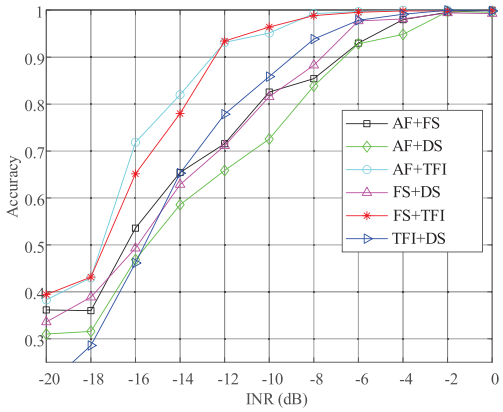


Fig. 10. Fusion performance for various modalities.

modalities. When one modality exhibits a pronounced performance superiority, the resultant combined performance closely aligns with that of the dominant modality. Conversely, in situations where each modality thrives in separate contexts (for example, AF and FS excel at low INR while TFI does so at high INR), the collaborative performance of these modalities significantly boosts overall effectiveness. TFI has the macroscopic characteristics of the interfering signal, and AF portrays

TABLE VII
MISM WITH DIFFERENT PENALTY VALUES ρ

ρ	AF	FS	DS	TFI	FF+AGM	Accuracy	FLOPs ($\times 10^6$)
-0.1	100%	0%	0%	0%	0%	77.97%	0.049
-1	85.1%	0%	0%	14.9%	0%	81.84%	0.158
-1.2	74.0%	0%	0%	17.4%	8.6%	85.51%	0.404
-1.5	72.3%	0%	0%	0%	27.7%	87.23%	0.783
-10	0%	0%	0%	0%	100%	89.5%	2.699

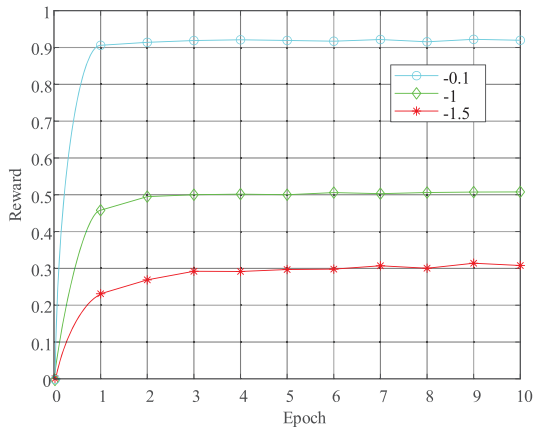


Fig. 11. Convergence speed of the reward function for MISM under various penalty factors.

the microscopic statistical characteristics, and the fusion of the two exhibits the best performance.

E. Modal Information Selection Mechanism

To achieve a fast computing speed and lightweight strategy, we choose to use a subset of modalities $x = [\text{TFI}, \text{AF}]$ as the input to the MISM. TFI goes through a 3×3 separable convolution with 16 channels and GAP, and AF is processed through a fully connected layer with 16 neurons. The outputs of both pathways are concatenated and passed through a fully connected layer with 5 neurons, representing the five strategies: AF, FS, DS, TFI, and FF+AGM. The MISM comprises only 587 learnable parameters, a number significantly smaller than the computational load of processing different modalities. Consequently, computational complexity of the policy network is almost negligible compared to the computational complexity of processing multimodal information. The proportion of network selection for each branch, recognition performance and computational complexity as the functions of penalty factor ρ are listed in Table VII. It can be seen that the penalty factor trades of performance with complexity. A lower ρ value enhances accuracy while imposing higher computational complexity. We simulate the convergence speed of the reward function for MISM under various penalty factors ($\rho = -0.1$, $\rho = -1$ and $\rho = -1.5$) as depicted in Fig. 11. The results reveal that MISM reward function converges within a few epochs, suggesting swift training speed for MISM.

Multimodal learning significantly enhances the interpretability of neural networks by leveraging the processing and integration of diverse information from multiple sources and in various forms. By merging features from different

modalities, multimodal learning enables the network to grasp the comprehensive context of a complex scenario. For instance, combining AF and TFI allows for a more precise understanding of entities and their meanings in both domains. This integration not only boosts the model's performance but also enhances interpretability by clearly indicating which AF information correlates with specific TFI features, offering a tangible framework for understanding decision-making. The neural network demonstrates varying behaviors upon integrating different modalities. Through the analysis of the differences between the pre-fusion and post-fusion conditions, we gain a deeper understanding of the neural network's operational mechanisms. Particularly, when each modality has its own strengths, the network shows a pronounced preference for harnessing the synergistic effects of the diverse modalities, thereby highlighting its tendency to learn from the complementary insights they offer.

V. Conclusion

We have proposed an AMN for wireless interference recognition, which combines different modal information, including notably AF, as the network input and intelligently selects the appropriate modal information for processing to achieve high recognition accuracy, while reducing the computational costs of multimodal information. Our network structure has adopted joint convolution and transformer to effectively extract both local and global features from different modal information. To reduce the complexity of MSA in transformer, DTM has been proposed. We have also introduced an AGM strategy during multimodal training to attain better network fusion performance. Additionally, MISM has been proposed, which strikes a balance between computational complexity and accuracy. The effectiveness of the proposed algorithms has been verified through extensive simulations.

References

[1] C. Qiu, Z. Wei, Z. Feng, and P. Zhang, "Joint resource allocation, placement and user association of multiple UAV-mounted base stations with in-band wireless backhaul," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1575–1578, Dec. 2019.

[2] T. Mao, J. Chen, Q. Wang, C. Han, Z. Wang, and G. K. Karagiannidis, "Waveform design for joint sensing and communications in millimeter-wave and low terahertz bands," *IEEE Trans. Commun.*, vol. 70, no. 10, pp. 7023–7039, Oct. 2022.

[3] F. Li, K.-Y. Lam, J. Hua, K. Zhao, N. Zhao, and L. Wang, "Improving spectrum management for satellite communication systems with hunger marketing," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 797–800, Jun. 2019.

[4] H. Bao, Y. Huo, X. Dong, and C. Huang, "Joint time and power allocation for 5G NR unlicensed systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 6195–6209, Sep. 2021.

[5] H. Zhou et al., "A cooperative matching approach for resource management in dynamic spectrum access networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 1047–1057, Feb. 2014.

[6] M. Karimi, S. M. S. Sadough, and M. Torabi, "Optimal cognitive radio spectrum access with joint spectrum sensing and power allocation," *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 8–11, Jan. 2020.

[7] H. Pirayesh and H. Zeng, "Jamming attacks and anti-jamming strategies in wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 767–809, 2nd Quart., 2022.

[8] C. Zhang, T. Mao, Z. Xiao, R. Liu, and X.-G. Xia, "Deceiving reactive jamming in dynamic wireless sensor networks: A deep reinforcement learning based approach," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 4455–4460.

[9] Y. Han, L. Duan, and R. Zhang, "Jamming-assisted eavesdropping over parallel fading channels," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 9, pp. 2486–2499, Sep. 2019.

[10] F. Gao, L. Guo, H. Li, J. Liu, and J. Fang, "Quantizer design for distributed GLRT detection of weak signal in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2032–2042, Apr. 2015.

[11] Y. Chen, J. Yang, W. Trappe, and R. P. Martin, "Detecting and localizing identity-based attacks in wireless and sensor networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 5, pp. 2418–2434, Jun. 2010.

[12] J. M. Moualeu, P. C. Sofotasios, D. B. da Costa, S. Muhaidat, W. Hamouda, and U. S. Dias, "Physical-layer security of SIMO communication systems over multipath fading conditions," *IEEE Trans. Sustain. Comput.*, vol. 6, no. 1, pp. 105–118, Jan. 2021.

[13] N. Hu, Y.-D. Yao, and J. Mitola, "Most active band (MAB) attack and countermeasures in a cognitive radio network," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 898–902, Mar. 2012.

[14] T. Mao and Z. Wang, "Physical-layer security enhancement for SIMO-MBM systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.

[15] I. Elleuch, A. Pourranjbar, and G. Kaddoum, "A novel distributed multi-agent reinforcement learning algorithm against jamming attacks," *IEEE Commun. Lett.*, vol. 25, no. 10, pp. 3204–3208, Oct. 2021.

[16] Q. Yan, H. Zeng, T. Jiang, M. Li, W. Lou, and Y. T. Hou, "Jamming resilient communication using MIMO interference cancellation," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1486–1499, Jul. 2016.

[17] Z. Ma, Y. Wu, M. Xiao, G. Liu, and Z. Zhang, "Interference suppression for railway wireless communication systems: A reconfigurable intelligent surface approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 11593–11603, Nov. 2021.

[18] Z. Shang, K. Huo, W. Liu, Y. Wang, and X. Li, "Interference environment model recognition for robust adaptive detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 4, pp. 2850–2861, Aug. 2020.

[19] C. Li, P. Qi, D. Wang, and Z. Li, "On the anti-interference tolerance of cognitive frequency hopping communication systems," *IEEE Trans. Rel.*, vol. 69, no. 4, pp. 1453–1464, Dec. 2020.

[20] B. Sun, Y. Zhou, J. Yuan, and J. Shi, "Interference cancellation based channel estimation for massive MIMO systems with time shifted pilots," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6826–6843, Oct. 2020.

[21] V. Kristem, A. F. Molisch, and L. Christen, "Jammer sensing and performance analysis of MC-CDMA ultrawideband systems in the presence of a wideband jammer," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3807–3821, Jun. 2018.

[22] W. Liu, X. Zhou, S. Durrani, and P. Popovski, "Secure communication with a wireless-powered friendly jammer," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 401–415, Jan. 2016.

[23] Y. Shi, X. Lu, Y. Niu, and Y. Li, "Efficient jamming identification in wireless communication: Using small sample data driven naive Bayes classifier," *IEEE Wireless Commun. Lett.*, vol. 10, no. 7, pp. 1375–1379, Jul. 2021.

[24] S. Zhao, Y. Zhou, L. Zhang, Y. Guo, and S. Tang, "Discrimination between radar targets and deception jamming in distributed multiple-radar architectures," *IET Radar, Sonar Navigat.*, vol. 11, no. 7, pp. 1124–1131, Jul. 2017.

[25] M. Greco, F. Gini, and A. Farina, "Radar detection and classification of jamming signals belonging to a cone class," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1984–1993, May 2008.

[26] A. Polydoros and K. Kim, "On the detection and classification of quadrature digital modulations in broad-band noise," *IEEE Trans. Commun.*, vol. 38, no. 8, pp. 1199–1211, Aug. 1990.

[27] A. Krayani, A. S. Alam, L. Marcenaro, A. Nallanathan, and C. Regazzoni, "Automatic jamming signal classification in cognitive UAV radios," *IEEE Trans. Veh. Technol.*, vol. 71, no. 12, pp. 12972–12988, Dec. 2022.

[28] D. Wei, S. Zhang, S. Chen, H. Zhao, and L. Zhu, "Research on anti-jamming technology of chaotic composite short range detection system based on underdetermined signal separation and spectral analysis," *IEEE Access*, vol. 7, pp. 42298–42308, 2019.

[29] A. Swami and B. M. Sadler, "Hierarchical digital modulation classification using cumulants," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 416–429, Mar. 2000.

[30] S. Kharbech, I. Dayoub, M. Zwingelstein-Colin, and E. P. Simon, "Blind digital modulation identification for MIMO systems in railway environments with high-speed channels and impulsive noise," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7370–7379, Aug. 2018.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Intl. Conf. Learning Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–14.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[33] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5987–5995.
[34] G. Gui, M. Liu, F. Tang, N. Kato, and F. Adachi, "6G: Opening new horizons for integration of comfort, security, and intelligence," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 126–132, Oct. 2020.
[35] H. Ye, G. Y. Li, and B.-H. Juang, "Deep learning based end-to-end wireless communication systems without pilots," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 3, pp. 702–714, Sep. 2021.
[36] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.
[37] Y. Wang, G. Gui, H. Gacanin, T. Ohtsuki, O. A. Dobre, and H. V. Poor, "An efficient specific emitter identification method based on complex-valued neural networks and network compression," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2305–2317, Aug. 2021.
[38] P. Wang, Y. Cheng, Q. Peng, B. Dong, and S. Li, "Low-bitwidth convolutional neural networks for wireless interference identification," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 557–569, Jun. 2022.
[39] Y. Kong, X. Wang, C. Wu, X. Yu, and G. Cui, "Active deception jamming recognition in the presence of extended target," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[40] M. Liu, Z. Liu, W. Lu, Y. Chen, X. Gao, and N. Zhao, "Distributed few-shot learning for intelligent recognition of communication jamming," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 3, pp. 395–405, Apr. 2022.
[41] F. Meng, P. Chen, L. Wu, and X. Wang, "Automatic modulation classification: A deep learning enabled approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10760–10772, Nov. 2018.
[42] Y. Tu, Y. Lin, C. Hou, and S. Mao, "Complex-valued networks for automatic modulation classification," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10085–10089, Sep. 2020.
[43] Q. Lv, Y. Quan, W. Feng, M. Sha, S. Dong, and M. Xing, "Radar deception jamming recognition based on weighted ensemble CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5107511.
[44] Y. Zhang, X. Ding, G. Li, Z. Zhang, and K. Yang, "Offline real-world wireless interference signal classification algorithm utilizing denoising diffusion probability model," *IEEE Signal Process. Lett.*, vol. 30, pp. 1132–1136, 2023.
[45] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–21.
[46] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[47] Z. Luo, Y. Cao, T.-S. Yeo, Y. Wang, and F. Wang, "Few-shot radar jamming recognition network via time-frequency self-attention and global knowledge distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5105612.
[48] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12692–12702.
[49] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. NIPS*, Denver, CO, USA, 1999, pp. 1057–1063.

Pengyu Wang received the Ph.D. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2023. He is a Post-Doctoral Researcher with the Department of Electronic Engineering, Tsinghua University, Beijing, China. His current research interests include artificial intelligence, cognitive radio, and machine learning on communication.

Ke Ma (Graduate Student Member, IEEE) received the B.S. degree (Hons.) from Tsinghua University, Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His current research interests include mmWave communication, beam management, and intelligent communication.

Yingshuang Bai received the M.S. degree in information and communication engineering from Harbin Engineering University, Harbin, China, in 2020. She joined Xiaomi, in April 2020, as a Standard Researcher, where her contributions primarily revolved around the advancement of new radio (NR) link simulation techniques. In August 2022, she joined Sony China Research Laboratory, Beijing, China. Her current research focuses on the domain of AI for wireless, with a specific emphasis on innovations within the physical layer.

Chen Sun (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2005. From August 2004 to May 2008, he was a Researcher with the ATR Wave Engineering Laboratories, Japan, working on adaptive beam-forming and direction-finding algorithms of parasitic array antennas and theoretical analysis of cooperative wireless networks. In June 2008, he joined the National Institute of Information and Communications Technology, Japan, as an Expert Researcher, working on distributed sensing and dynamic spectrum access in TV white space. Since then, he has been contributing to IEEE 1900.6 Standard, IEEE 802.11af Standard, and Wi-Fi Alliance Specifications for Wi-Fi Networks in TV White Space. In 2012, he joined Sony China, as a Research Manager, working on IEEE 802.19, ETSI, and 3GPP standards development. He is currently the Head of Beijing Laboratory, Sony Research and Development Center. He is the author of *Handbook on Advancements in Smart Antenna Technologies for Wireless Networks*, 60 international journals, and more than 120 conference papers. He is also the first inventor of 80 granted patents in U.S., EU, Japan, and China, including nine patents that are worth more than four million dollars. His research interests include AI for beamforming, wireless federated learning, blockchain-based spectrum sharing, wireless sensing, and V2X. He served as the 802.19 standards and rapporteur of ETSI standards.

Zhaocheng Wang (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees from Tsinghua University, in 1991, 1993, and 1996, respectively.

From 1996 to 1997, he was a Post-Doctoral Fellow with Nanyang Technological University, Singapore. From 1997 to 1999, he was a Research Engineer/a Senior Engineer with OKI Techno Centre (Singapore) Pte. Ltd., Singapore. From 1999 to 2009, he was a Senior Engineer/a Principal Engineer with Sony Deutschland GmbH, Germany. Since 2009, he has been a Professor with the Department of Electronic Engineering, Tsinghua University, where he is currently the Director of Broadband Communication Key Laboratory, Beijing National Research Center for Information Science and Technology (BNRist). He has authored or co-authored two books, which have been selected by IEEE Series on Digital and Mobile Communication and published by Wiley-IEEE Press. He has authored/co-authored more than 200 peer-reviewed journal articles. He holds 60 U.S./European granted patents (23 of them as the first inventor). His research interests include wireless communications, millimeter wave communications, optical wireless communications, and AI empowered wireless communications. He is a fellow of the Institution of Engineering and Technology. He was a recipient of the ICC2013 Best Paper Award, the OECC2015 Best Student Paper Award, the 2016 IEEE Scott Helt Memorial Award, the 2016 IET Premium Award, the 2016 National Award for Science and Technology Progress (First Prize), the ICC2017 Best Paper Award, the 2018 IEEE ComSoc Asia–Pacific Outstanding Paper Award, and the 2020 IEEE ComSoc Leonard G. Abraham Prize.

Sheng Chen (Life Fellow, IEEE) received the B.Eng. degree from the East China Petroleum Institute, Dongying, China, in January 1982, the Ph.D. degree from City University, London, in September 1986, both in control engineering, and the D.Sc. degree from the University of Southampton, Southampton, U.K., in August 2005. From 1986 to 1999, he held research and academic appointments with the University of Sheffield, University of Edinburgh, and University of Portsmouth, all in U.K. Since 1999, he has been with the School of Electronics and Computer Science, the University of Southampton, U.K., where he holds the post of Professor of intelligent systems and signal processing. He has published over 700 research papers. He has more than 20,000 Web of Science citations with an H-index of 63 and more than 39,000 Google Scholar citations with an H-index of 84. His research interests include adaptive signal processing, wireless communications, modeling and identification of nonlinear systems, neural network and machine learning, and evolutionary computation methods and optimization. He is a fellow of the United Kingdom Royal Academy of Engineering, Asia–Pacific Artificial Intelligence Association, and IET. He is one of the original ISI highly cited researchers in engineering in March 2004