Received 9 February 2023, accepted 23 February 2023, date of publication 1 March 2023, date of current version 6 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3250644

RESEARCH ARTICLE

Deep Learning Based Cognitive Radio Modulation Parameter Estimation

WENXUAN MA AND ZHUORAN CAI
College of Physics and Electronic Information, Yantai University, Yantai 264005, China
Corresponding author: Zhuoran Cai (caizhuoran@ytu.edu.cn)

ABSTRACT Automatic Modulation Classification (AMC) is a critical issue in electromagnetic spatial perception. Currently traditional recognition techniques are difficult to adapt to complex signal situations. Most existing modulation classification algorithms ignore the complementarity between different features and the importance of feature fusion. Based on this, we proposed a method for image feature fusion for AMC that fully uses the complementarity between different image features. The original signal is converted into an image by the Gramian Angular Field (GAF) method, and the GAF image is used as the input to the network, meanwhile the received signal is converted from the Inphase-Quadrature (I-Q) domain to the r-θ domain using the Accumulated Polar Feature conversion technique, and the original signal is feature coded from the r-θ domain and then converted into an image. The fused features of the two images are used as input to the neural network for model training to achieve automatic modulation classification of multiple types of signals. In the evaluation phase, the differences in the recognition effectiveness of the proposed method by different neural networks are discussed. Experiments show that the best performance is achieved using the Swin-Transformer network model, with a more than 90% recognition rate for the modulation method at signal-to-noise ratios greater than 4dB.

INDEX TERMS Modulation classification, Gramian angular field, accumulated polar feature, feature fusion.

I. INTRODUCTION

With the rapid development of wireless communication technology, the limited spectrum resources can not meet the increasing demand for wireless communication devices. Cognitive Radio (CR) technology can improve spectrum utilization efficiency by adjusting transmission parameters in Real-time through intelligent learning and aware-ness of the spectrum environment. AMC is one of the basic schemes in CR, which can identify multiple modulations from unknown signals and is an essential component of non-cooperative communication systems. With the increase in the number of transmitter types and interference sources and the influence of complex wireless environments on the transmitted signals, the identification algorithm in a single scenario is no longer applicable. Therefore, it is necessary to develop modulation classification algorithms adapted to the complex environment of wireless communication, which can not only automatically extract deep features of the signal to improve accuracy but also have high recognition accuracy with a low signal-to-ratio [1], [2].

There are two main methods for the current modulation classification algorithm. The first is simple manual calculation classification algorithm. The second is deep learning based automatic modulation classification algorithm [3], [4]. Manually computed classification algorithms in modulation classification are divided into Likelihood ratio-Based (LB) and Feature extraction-Based (FB) methods. The LB [5] algorithm calculates and compares the likelihood function value between signals. Although apparent classification accuracy can be obtained, the likelihood function is very complex, resulting in a considerable amount of calculation, which is extremely difficult to deal with some complex signal types. The FB [6] algorithm makes up for the shortcomings of the LB algorithm, the calculation amount reduces a lot, and the extraction of signal features also achieves good results. However, the FB method extracts feature based on specific signal samples and then sets the decision-making

The associate editor coordinating the review of this manuscript and approving it for publication was Yun Lin.

method. Hence, it is easy to make mistakes in a poor channel environment.

In recent years, the rapid development of Deep Learning (DL) techniques has led to remarkable achievements in areas such as computer vision [7], natural language processing [8], and information resource allocation for mobile networks [9], [10]. The development of DL techniques in the field of communication is also attracting more and more attention. Compared with the two previously mentioned manual computational classification methods, DL automatically learns radio features in the I/Q data in the signal and classifies the signal modulation based on these features, so DL techniques have been widely used in the field of automatic modulation classification. The two most basic network structures of DL technology are Convolutional Neural Network (CNN) and Recur-rent Neural Network (RNN). In order to adapt to the complex communication environment, some enhanced network structures like Convolutional Long term Deep Neural Network (CLDNN) and Residual Network (ResNet) are also applied to AMC.

A. RELATED WORK

Use of deep learning techniques to achieve modulation classification has become the mainstream direction of the subject. Modulation classification of signals using deep learning requires a large amount of communication signal data to train the neural network architecture, and communication signal data is readily available on the receiver side. Thus deep learning techniques have a lot of exploration value in the wireless communication field.

For example, the literature [5] proposes converting the signal domain into a graph domain, identifying the modulation type using the geometric relationship of the constellation graph, constructing constellation graphs of different signals, and directly extracting image features using a simple neural network. The classification of signals using simple convolutional neural networks was proposed in [11] and [12], and this method directly performs feature extraction on the original data signal with fast classification and small network models, which can be directly applied to hardware devices. However, the disadvantage of low recognition rate and significant error brought by the simple network model cannot be ignored. The literature [13] proposed a CLDNN structure for signal modulation for classification and achieved excellent results.

In order to train deeper networks, ResNet is widely used in the field of modulation classification. In addition to the recognition of raw I/Q data, considering that CNN is proposed in the field of image recognition, [14] proposed to transform the signal into a two-dimensional image by Spectral Correlation Function (SCF) and extract complex features from the SCF image of the received signal by CNN network. This method takes full advantage of convolutional neural networks, but the recognition accuracy of a single feature is always weaker than that of multiple features.

When using neural networks to classify signal modulation, many scholars have considered preprocessing and manual feature descriptions of the original signal in an appropriate form to improve the correctness of the features extracted by the neural network [15]. The uniform drawback of the work mentioned above is that it only considers the extraction of a single feature, ignoring the complementarity between different features. Multimodal approaches have been proposed in computer vision to recognize images, and a multimodal fusion model is proposed in the field of automatic modulation classification [16], [17] to further improve the classification performance in order to solve the problem of degraded classification performance in a low signal-to-noise environment caused by a single feature.

Automatic modulation classification combined with multiple features has an excellent performance, for example, the combination of two images: a circular spectrum image and a constellation diagram [18], a time frequency diagram and a transient autocorrelation image [19], and two time frequency diagrams of the signal combining Smooth Pseudo Wigner-Ville Distribution (SPWVD) and Born-Jordan Distribution (BJD) images [20]. The methods for combining signal features and sequences are higher-order accumulation and IQ sequences [21]. Two methods of combining sequences: Discrete Orthogonal S-Transform (DOST) sequences and IQ sequences [22].

Although the above methods fuse multiple features, they have a common limitation: they do not consider the complementarity between different image features and do not integrate them with an appropriate fusion mechanism. Most of the existing DL frame-work-based AMC methods mentioned above try to characterize the original modulated signal, but rarely consider the relationship between the signal features and the network structure, and the recognition effect of different deep learning methods on the signal varies greatly. Therefore, this paper uses a variety of networks to recognize the transformed images in order to investigate which network is the most superior for signal classification.

B. CONTRIBUTION

The innovation of this study can be summarized as follows:

Fully considers the complementary nature of the two signals image features, encodes the signal as a two dimensional image by the Gramian Angular Field method, and simultaneously maps the complex symbols onto the polar coordinate image using the cumulative polar coordinate feature transform. Features are extracted from both images simultaneously using a neural network structure. The extracted features are fed into a neural network classifier using the Multi-task Multi-sensor Fusion (MMF) fusion method.

The stage of experimental simulation firstly used two convolutional neural networks and two Transformer vision networks to extract features from two images and found that the classification accuracy of the Transformer vision network was better than that of the convolutional neural network.

Compared the classification accuracy among seven Transformer vision networks. The Swin-Transformer network was found to achieve the best results.

After introducing the related work and the paper's contribution, Section II presents the modulation method of the signal, the principle of GAF, and the cumulative polar coordinate constellation image. Section III presents the selected neural network characteristics and the image feature fusion method after signal conversion. Section IV presents the data set used in this paper and the results of simulation experiments. A summary and discussion of the proposed method are carried out in Section V.

II. METHOD

A. SIGNAL MODEL

Automatic modulation classification is the link between signal demodulation. The module that handles modulation classification is usually deployed on the receiving side of the overall communication system, transmitting the received modulation information to the demodulator, which then demodulates it.

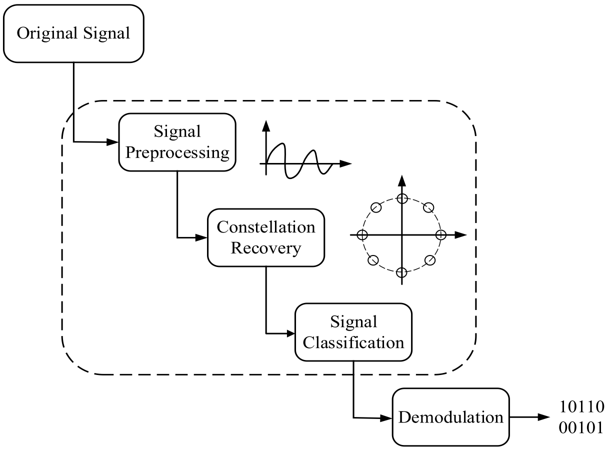


FIGURE 1. Cognitive radio receiver operation.

In a wireless communication system, the received signal is expressed as:

$$q(t) = a(t) * z(t) + w(t) \tag{1}$$

The $a(t)$ represents the signal that is modulated for transmission, and $z(t)$ the represents the impulse response to external variations in the wireless channel. $w(t)$ is the most basic noise and interference model: additive Gaussian white noise. This model increases broadband noise with constant spectral density and Gaussian distributed amplitude. $q(t)$ is the signal received by the receiving end, which generally represents information in I/Q format and is sampled n times by the analog-to-digital converter at a specific rate, where the real and imaginary parts denote the I, Q vectors, respectively. By specific expansion, $q(t)$ is also expressed as:

$$q(t) = s(t)\, e^{j(2\pi f_0 + \theta_0(t))} a(t) + w(t) \tag{2}$$

The $s(t)$ represents the Rayleigh fading channel in the wireless communication channel environment. f_0 represents the different Local Oscillators (LO) changing the signal frequency. $\theta_0(t)$ represents the phase shift of the signal due to the Doppler effect. Digital modulation technology has three primary forms: Amplitude Shift Keying (ASK), the amplitude of the carrier is modulated by the message signal to take different values, susceptible to gain changes. Frequency-Shift Keying (FSK), the frequency of the carrier, takes different values according to the signal data, well anti-interference but occupies a large bandwidth. Phase Shift Keying (PSK), the phase of the carrier is changed, and it can also be used as timing information to synchronize the clock. Various other modulation methods are improvements or combinations of the above methods, and different modulation methods have different mathematical expressions. $a(t)$ can be expressed as:

$$\begin{aligned} a(t) = &\left[E_m \sum\nolimits_n r_n d(t - nT_S)\right] \\ &\times \cos(2\pi(f_c + f_m)t + \varphi_0 + \varphi_m) \end{aligned} \tag{3}$$

The E_m represents the modulated amplitude, r_n represents the symbol sequence, and $d(t - nT_S)$ represents the signal pulse. f_m represents the modulation frequency, and f_c represents the carrier frequency. φ_0 represents the initial phase, and φ_m represents the modulation. Quadrature Amplitude Modulation (QAM) is a combination of amplitude modulation and phase modulation. It has two orthogonal carriers modulated r_n and v_n, so that $a(t)$ can be expressed as:

$$\begin{aligned} s(t) = &\left[A_m \sum\nolimits_n r_n g(t - nT_S)\right] \cos(2\pi f_c t + f_0) \\ &+ \left[A_m \sum\nolimits_n v_n g(t - nT_S)\right] \sin(2\pi f_c t + f_0) \end{aligned} \tag{4}$$

B. SIGNAL CONVERSION

1) GAF BASED TWO DIMENSIONALIZATION OF TIME SERIES SIGNALS

Using the existing knowledge of communication experts, it is possible to transform the I/Q domain into the r/θ domain before deep learning to classify the signals. We can deal with channel fading directly in the r/θ domain. The amplitude can be mapped to the r-axis and the phase can be represented by the theta-axis. It is easier to learn the modified parameters (Δr, $\Delta\theta$) to eliminate the channel fading in the r/θ domain.

The coordinate system generally used to represent a time series is the Cartesian coordinate system, which intersects two number axes at the origin, constituting a planar affine coordinate system, and the common two-dimensional coordinate system position is determined by a pair of numbers. A polar coordinate system is one in which a certain point in the plane, called the pole, is taken, a ray, called the polar axis, is drawn, and a unit of length and the positive direction of the angle are selected. The Gramian Angular Field (GAF) uses a polar coordinate system to represent the time series, and in the GAF matrix, each element is the cosine of the sum of the angles [24], [25].

Suppose a time series $T = \{t_1, t_2, t_3, \ldots.., t_n\}$, each time has a definite value. The time series T is transformed by the

Piecewise Aggregation Approximation (PAA) method, which converts long series data into short series data by averaging the time series into multiple segments and each segment is represented by its corresponding mean value. Reduce all the values in the transformed time series T to between [0,1] by the following equation.

$$\overline{t_x} = \frac{t_x - \min(T)}{\max(T) - \min(T)} \tag{5}$$

Reduce all the values in the transformed time series T to between [−1,1] by the following equation.

$$\overline{t_x} = \frac{(t_x - \max(T) + (t_x - \min(T)))}{\max(T) - \min(T)} \tag{6}$$

The time series $\hat{T}$ is obtained after the reduction. The values are then encoded as angular cosines and the timestamps as radii, so that $\hat{T}$ can be expressed in polar coordinates with the following equation.

$$\begin{aligned} \Phi &= \arccos\left(\overline{t_x}\right), 0 \leq \overline{t_x} \leq 1, \overline{t_x} \in \hat{T} \\ r &= \frac{g_i}{N} \end{aligned} \tag{7}$$

where g_i is the time stamp, r is the radius, and the interval is divided into N equal parts to regulate the span in the polar coordinate system.

The method described above is a new way of representing time series whose corresponding values change between different angles within the circle of polar coordinates as the time series increases. The conversion of the normal time series values in equation (7) has two important properties. The first property is that it is bijective because when $\boldsymbol{\Phi} \in [0, \pi]$, $\cos(\boldsymbol{\Phi})$ are monotonic, a time series transformed produces one and only one result in polar coordinates with a unique inverse mapping; the second property is that unlike the Cartesian coordinate system, the polar coordinate system preserves the absolute time relationship [26].

In the cosine function $y = \cos x$. The value field is [0,1], which corresponds to the definition field $[0, \pi/2]$, and when the value field is [−1,1], which corresponds to the definition field $[0, \pi]$. This indicates that rescaled data in different intervals have different angular constraints, which provides different Gramian Angular Field information granularity for the classification task. Gramian Angular Difference (GADF) has a precise inverse mapping to convert the information in the polar coordinate system into images.

There are two categories of Gramian Angular Field: The Gramian Summation Angular Field (GASF) and The Gramian Difference Angular Field (GADF). Both methods are used to determine the temporal correlation of different time intervals by the triangular sums or differences between each point after converting the time series to a polar coordinate system. The equations for both categories are as follows.

$$\begin{aligned} GASF &= [\cos(\boldsymbol{\Phi}_a + \boldsymbol{\Phi}_b)] \\ &= \cos\boldsymbol{\Phi}_a \cdot \cos\boldsymbol{\Phi}_b - \sin\boldsymbol{\Phi}_a \cdot \sin\boldsymbol{\Phi}_b \\ &= \cos\arccos\left(\overline{t_a}\right) \cdot \cos\arccos\left(\overline{t_b}\right) \\ &\quad - \sin\arccos\left(\overline{t_a}\right) \cdot \sin\arccos\left(\overline{t_b}\right) \\ &= \overline{t_a} \cdot \overline{t_b} - \sqrt{1 - \overline{t_a}^2} \cdot \sqrt{1 - \overline{t_b}^2} \end{aligned} \tag{8}$$

$$\begin{aligned} GASF &= [\sin(\boldsymbol{\Phi}_a - \boldsymbol{\Phi}_b)] \\ &= \sin\boldsymbol{\Phi}_a \cdot \cos\boldsymbol{\Phi}_b - \cos\boldsymbol{\Phi}_a \cdot \sin\boldsymbol{\Phi}_b \\ &= \sin\arccos\left(\overline{t_a}\right) \cdot \cos\arccos\left(\overline{t_b}\right) \\ &\quad - \cos\arccos\left(\overline{t_a}\right) \cdot \sin\arccos\left(\overline{t_b}\right) \\ &= \sqrt{1 - \overline{t_a}^2} \cdot \overline{t_b} - \overline{t_a} \cdot \sqrt{1 - \overline{t_b}^2} \end{aligned} \tag{9}$$

After converting the normal time series into a polar coordinate system, the time series of each time step is used as a one-dimensional metric space. The metric space is a set with a distance function that defines the distance between all elements in the set. This distance function is called the metric on the set. What is described in the article is the representation of the distance in the polar coordinate system after the conversion of the I/Q domain to r/θ.

Using the GASF method as an example, $\hat{T}$ can then be converted to matrix $\mathbf{G}$:

$$[G(T\{t_1, t_2, t_3, \ldots, t_n\})] = \begin{pmatrix} \cos(\Phi_1 + \Phi_1) & \cos(\Phi_1 + \Phi_2) & \ldots & \cos(\Phi_1 + \Phi_n) \\ & \vdots & \ddots & \vdots \\ \cos(\Phi_n + \Phi_1) & \cos(\Phi_n + \Phi_2) & \cdots & \cos(\Phi_n + \Phi_n) \end{pmatrix}$$

The Gramian Angular Field provides a way to preserve temporal dependence, with time increasing with position from the top left to the bottom right. From the main diagonal, we can reconstruct the time series from the high level features learned by the deep neural network.

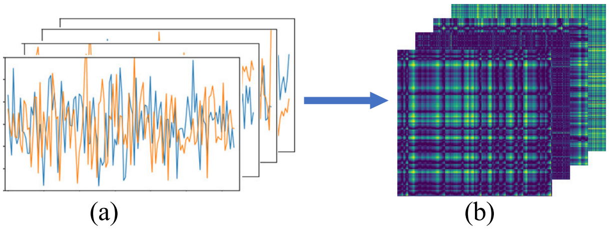


FIGURE 2. Overall Gramian Angular Field conversion method.(a)original signal. (b) The images obtained after GASF.

2) ACCUMULATED POLAR FEATURE CONVERSION

In digital communication systems, viewing the signal waveform directly in a right-angle coordinate system needs to be a clearer understanding of its modulation. Displaying the signal information on the complex plane provides an intuitive representation of the modulation type of the original signal, and this method is the principle of a constellation diagram.

Some literature [23] used CNN to identify constellation diagrams and thus achieve the purpose of identifying modulated signals. The identification process is simple. However, much temporal information is lost in the planar images, in addition to Quadrature Phase Shift Keying (QPSK) images are also easily misidentified as 8PSK and 16QAM images are also easily misidentified as 64QAM when the noise is too large.

The signals sent to the neural network are converted into more straightforward classifications before entering them for recognition. The proposed method of accumulated polar features includes three steps, the first step converts the signal from the I/Q domain to the r/θ domain, the second step converts the information on the coordinate axes into a gray image, and the third step converts the gray image into a color image to better utilize the neural network to extract features.

Step1: Similar to the approach proposed by GAF above, we similarly convert the signal from a cartesian coordinate system to a polar coordinate system. The specific knowledge in the communication system is learned in the polar coordinate system. Signal conversion by polar coordinate system features can improve the feature extraction part of the neural network with better performance [27].

The I/Q domain of the signal is converted to the r/θ domain before the neural network learning, which corresponds to the conversion of the I/Q axis to the r/θ axis obtained on the coordinate axes, where I, Q represent the real and imaginary parts of the complex symbols. r, θ represents the radius and polar position after conversion to the polar coordinate system, and L is the symbol length.

$$r = \sqrt{x^2 + y^2} \quad \theta = \arctan(\frac{y}{x}) \tag{10}$$

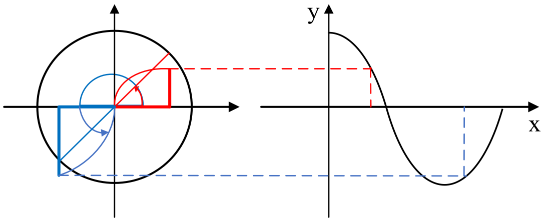


FIGURE 3. Conversion between I/Q axis and r/θ axis.

Step2: Now the information in the signal is still on the coordinate axes, and the next step is to convert the transformed symbols on the r/θ axis into a two-dimensional image. The conversion process can be represented by algorithm 1, where the range of r-axis is represented by r_0, r_1, the range of θ-axis is represented by $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$, and u_r, u_θ represents the image resolution of r-axis and θ-axis respectively.

First, the grid interval Δk_r, Δk_θ of r-axis and θ-axis is set according to the input. Then, the transformation symbols on r-axis and θ-axis are mapped to a point on the image, and [x, y] denotes the coordinates of r-axis and θ-axis, respectively. Finally, if no symbol is mapped to a point in the image, the pixel value of the grid-like image $\dot{\boldsymbol{M}}$ is set to 1.

Therefore, we obtain the binary image $\dot{\boldsymbol{M}}$ with its value set to 0 or 1. Figure 4 (b) shows the transformed image based on the polarity feature.

Step3: After the previous transformation, the information on the polar axis is mapped to a two-dimensional image, but the image is gray. The historical information of the symbols is accumulated by adding a time axis **T**. The accumulated image is converted from gray to color, and different colors indicate that the symbols have different probabilities of appearing at that point. The additive gaussian white noise generated in the signal passing through the channel can be considered a random process, which causes a decrease in the accuracy of the modulation classification, but the symbols with noise will have a high probability of appearing near the original point. Using the information in the accumulated symbols, the higher the probability of appearing in the image, the darker the color will be, as in Figure 4(c). Finally, $\dot{\boldsymbol{M}}$ is the data input to the neural network. In summary, converting the signal to polar coordinate representation and then to gray images finally accumulated into color images improves the classification accuracy and increases the convergence speed, significantly reducing the offline training overhead of deep learning.

```
Algorithm 1 Calculation Methods for Images
Input $r, \theta, r_0, r_1, \theta_0, \theta_1, u_r, u_\theta$
Initialize Image Matrix $\dot{M} = 0$
Calculate grid interval of r axis $(r_1 - r_0)/u_r \leftarrow \Delta k_r$
Calculate grid interval of $\theta$ axis $(\theta_1 - \theta_0)/u_\theta \leftarrow \Delta k_\theta$
for n = 0 : N-1 do
Coordinate conversion r axis $\left[\frac{r[n]-r_0}{\Delta k_r}\right] \leftarrow x$
Coordinate conversion $\theta$ axis $\left[\frac{\theta[n]-\theta_0}{\Delta k_\theta}\right] \leftarrow y$
if polar feature then Calculate pixel value of black and white image $\dot{M} \leftarrow [i, j]$
else if accumulated polar feature then Calculate pixel value of grayscale image $\dot{M}[x, y] + 1 \leftarrow \dot{M}[x, y]$
end if
end for
return $\dot{M}$
```

III. MODULATION CLASSIFICATION SCHEME

The scheme proposed in this paper for the modulation classification of signals can be divided into three steps: The first step, the original signal is preprocessed and converted into an image to represent the signal features. The second step is extracting the classification features using a neural network. The third step is to fuse the features extracted from the two images and then input them to the fully connected layer in the neural network to classify the signal modulation type. The principle of converting the signal to an image has been described in Section II. In the following, the network for feature extraction of the signal and the basic theory of fusion of the two features are described.

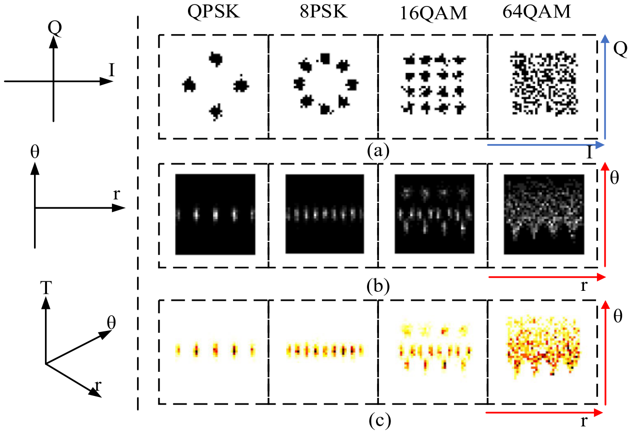


FIGURE 4. Three different conversion images of the four signals when SNR = 18dB. (a)The constellation diagram after the standard I/Q conversion. (b) Grayscale image when the signal is converted to polar coordinates. (c) Color image after adding time accumulation.

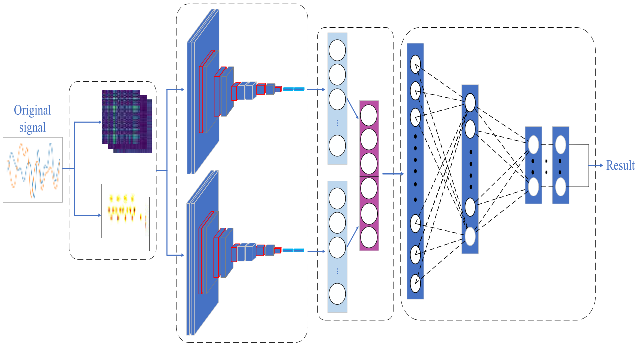


FIGURE 5. Overall block diagram of the proposed solution.

A. SIGNAL FEATURES EXTRACTION

Machine learning is an approach to artificial intelligence that continuously learns and predicts data from reality. Deep learning is a neural network algorithm used by machine learning to simulate the human brain for learning, which requires high power computing power to support. Deep learning-based methods are capable of automatically learning multi-level representations of high-dimensional data. CNN and Transformer have been found to have better results than other models in the field of classification and detection. In this paper, the neural network model uses two different visual neural net-works to recognize the proposed method. One is the Deep Convolutional Neural Network, and the other is the Transformer network model.

Deep convolutional neural networks are mainly modeled by ResNet and (Convolutional Neural networks) ConvNeXt. ResNet proposes residual learning to solve the problem of network training degradation in deep learning. The most significant difference between residual networks and previous networks is that there is an additional network shortcut branch. Because of the existence of this branch, when the network is back propagated, the loss can be directly transmitted to the more forward network through this shortcut. ConvNeXt is a recently proposed network structure with powerful classification capabilities. The ConvNeXt network demonstrates that traditional convolutional networks still have room for exploration in image recognition, and batchnorm is an essential component of ConvNeXt because it improves convergence and reduces overfitting. Using these two convolutional neural networks for feature extraction and classification is suitable for the proposed image feature fusion method.

The Transformer adopted Swin-Transformer and Twins-PCPVT as the main models. Swin-Transformer is a Transformer-based deep learning model that achieves the state of the art performance in vision tasks. Using a hierarchical Transformer and moving windows brings higher efficiency by limiting the use of self-attention within windows. Moving allows interaction between two adjacent windows and cross-window connections between upper and lower layers, thus achieving a global modeling effect in disguise.

B. FEATURES FUSION

In the subject of automatic signal modulation identification, most of the methods usually identify and classify single features, such as cyclic power spectrum features of the signal, constellation diagram of the signal, higher-order accumulation of the signal, time-frequency conversion of the signal into a time-frequency diagram and then feature extraction. However, the effect of poor classification due to the extraction of a single feature is worthy of attention. Multi-feature fusion will achieve complementary advantages between different features and still have good classification results when the channel environment is complex. The literature [36] achieved significant results by converting the original signal into color time-frequency images by SPWVD and BJD time-frequency conversion methods, combining the features of both images. Based on this, in this paper, the MMF method is used for feature fusion of the two images converted from the original signal proposed earlier, which reduces the complexity by connecting all features of all modes and adding penalty terms between each modal feature and the connected features. It con-siders not only the features of individual modes but also the relationship between the two feature images is considered. During training, penalty terms are added between the two image feature labels to reduce the complexity of the network and improve the classification accuracy of the signal by deep learning.

MMF is proposed because the input contains two features of the same information and the penalty between the two predicted labels needs to be considered. A method for probability distributions, called Kullback-Leibler scatter (KL) [28], [29] was first introduced. KL scatter originates in information theory, where the main goal is to quantify how much information is in the data. The most critical metric in information theory is called entropy. Entropy is defined as:

$$H = -\sum_{i=1}^{N} p(x_i) \log p(x_i) \tag{11}$$

The Kullback-Leibler scatter is only a slight modification of our entropy formula. Bringing the actual distribution of the data p and the theoretical distribution of the data q into KL is to derive the asymmetry of the difference between the two probability distributions. Check the difference for each logged value.

$$D_{KL}\left(p \parallel q\right) = \sum\nolimits_{i=1}^{N} p\left(x_i\right)\left(\log p\left(x_i\right) - \log q\left(x_i\right)\right) \quad (12)$$

More common ways to view KL dispersion:

$$D_{KL}\left(p \parallel q\right) = \sum\nolimits_{i=1}^{N} p\left(x_i\right)\left(\log \frac{p\left(x_i\right)}{q\left(x_i\right)}\right) \quad (13)$$

KL scatter has two essential properties: non-negativity and asymmetry. In deep learning, the training distribution q keeps approaching p, and when the fitted distribution tends to 0, the KL scatters to infinity, and the fitted distribution tends to cover all the ranges of the theoretical distribution when the distance between the minimum fitted distribution, and the actual distribution is used for forwarding KL scatter.

The Jensen-Shannon divergence (JS) [30], [31] scattering constructs a symmetric scattering formula that takes values between 0 and 1.

$$\begin{aligned} u &= \frac{p+q}{2} \\ D_{JS}\left(p \parallel q\right) &= \frac{1}{2} KL\left(p \parallel u\right) + \frac{1}{2} KL\left(q \parallel u\right) \end{aligned} \quad (14)$$

Substitution into the above equation yields:

$$\begin{aligned} D_{JS}\left(p \parallel q\right) = &\frac{1}{2} \sum\nolimits_{i=1}^{N} p\left(x\right) \log \frac{p\left(x\right)}{p\left(x\right)+q\left(x\right)} \\ &+ \frac{1}{2} \sum\nolimits_{i=1}^{N} q\left(x\right) \log \frac{q\left(x\right)}{p\left(x\right)+q\left(x\right)} + \log 2 \end{aligned} \quad (15)$$

By denoting z_i^r as the distributed features of the two image features at the moment i, r as the first image feature or the second image feature, and z_i^c as the feature after the concatenation of the two image features at the moment i. $\beta = \{\theta^c, \theta^r\}$ as the model parameters obtained after the neural network training, then the loss function of the whole feature model can be written as:

$$\begin{aligned} L\left(\beta\right) = &\frac{1}{N} \sum\nolimits_{i=1}^{N} JS\left(t_i \parallel p\theta^c\left(z_i^c\right)\right) \\ &+ \frac{1}{N} \sum\nolimits_{r=1}^{2} \sum\nolimits_{i=1}^{N} JS\left(p\theta^c\left(z_i^c\right) \parallel p\theta^r\left(z_i^r\right)\right) \\ &+ \frac{u}{2} \parallel \theta^c \parallel^2 + \frac{\mu}{2} \parallel \theta^r \parallel^2 \end{aligned} \quad (16)$$

where t_i represents the actual probability distribution of the type of signal modulation, N represents the training samples before input to the neural network, and μ is a parameter set before the neural network starts learning, not obtained through network learning. The third and fourth parts of the above equation are the regularization used to avoid overfitting so that each variable has a little effect on the prediction.

Where the probability distribution $p_\theta\left(z_i\right)$ is obtained by bringing into the softmax function, the multilayer neural network in the penultimate layer output is not easy to regularize the results are challenging to handle. Normalization using the softmax function leads to:

$$\begin{aligned} p_\theta\left(z_i\right) = soft \max\left(z_i\right) = &\frac{1}{\sum\nolimits_{k=1}^{K} e^{\theta_K^T z_i}} \\ &\times \left[e^{\theta_1^T z_i}, e^{\theta_2^T z_i}, \cdots, e^{\theta_k^T z_i}\right]^T \end{aligned} \quad (17)$$

K in the above equation represents the sum of the modulation types of the signal. In the process of model parameter learning, the gradient descent method is used to continuously deoptimize and repeatedly update the model parameters until convergence to achieve the purpose of optimizing the model. Putting (16) performing gradient descent method derivation yields.

$$\begin{aligned} \frac{\partial L\left(\beta\right)}{\partial \theta^c} = &\frac{1}{N} \sum\nolimits_{i=1}^{N} \frac{\partial JS\left(t_i \parallel p_\theta c\left(z_i^c\right)\right)}{\partial \theta^c} + \mu\theta^c \\ &+ \frac{1}{N} \sum\nolimits_{r=1}^{2} \sum\nolimits_{i=1}^{N} \frac{\partial JS\left(p_{\theta^c}\left(z_i^c\right) \parallel p_{\theta^r}\left(z_i^r\right)\right)}{\partial \theta^c} \\ \frac{\partial L\left(\beta\right)}{\partial \theta^r} = &\frac{1}{N} \sum\nolimits_{r=1}^{2} \sum\nolimits_{i=1}^{N} \frac{\partial JS\left(p_{\theta^c}\left(z_i^c\right) \parallel p_{\theta^r}\left(z_i^r\right)\right)}{\partial \theta^c} + \mu\theta^r \end{aligned} \quad (18)$$

The solution can be brought in for the calculation of $\frac{\partial JS\left(p_{\theta^c}\left(z_i^c\right) \parallel p_{\theta^r}\left(z_i^r\right)\right)}{\partial \theta^c}$ using the KL formula(19).

$$\begin{aligned} &\frac{\partial \sum_{K=1}^{K} JS\left(p_{\theta^c}\left(k \mid z_i^c\right) \parallel p_{\theta^r}\left(k \mid z_i^r\right)\right)}{\partial \theta_{jl}^c} \\ &\quad = \sum_{k=1}^{K} \ln\left(\frac{2p_{\theta^c}\left(k \mid z_i^c\right)}{p_{\theta^c}\left(k \mid z_i^c\right) + p_{\theta^r}\left(k \mid z_i^r\right)}\right) \frac{\partial p_{\theta^c}\left(k \mid z_i^c\right)}{2\partial \theta_{jl}^c} \end{aligned} \quad (19)$$

where θ_j^c and θ_{jl}^c denote a sub-vector of θ^c and the $l - th$ element of θ_j^c respectively. The model parameters $\beta = \{\theta^c, \theta^r\}$ need to be updated each time the neural network undergoes training iterations to derive them because the communication signal dataset is extensive, and small batches are used to speed up the learning of model parameters. G^1 and G^2 are set as different features of the signal after conversion into an image. In order to obtain the feature fusion results, the two original signal datasets are randomly divided into small batches for network training so that two different features G^1 and G^2 are connected to obtain G^c, first give the model parameters $\beta = \{\theta^c, \theta^r\}$a random value, randomly select small batches of G, use the previously mentioned gradient descent method to continuously update the parameters and loss function (18), as the parameters are constantly updated, (16) is also constantly updated. Finally, the above process is repeated until the optimal solution is reached.

In the prediction process after network training, the final probability distribution can be obtained using the following equation, where p is the expected label distribution.

$$\min_{p} L\left(p|p_{\theta^1}, p_{\theta^2}\right) = \sum\nolimits_{m=1}^{2} \sum\nolimits_{K=1}^{K} p\left(k\right) \ln\left(\frac{p\left(k\right)}{p_{\theta}m\left(k\right)}\right)$$

$$s.t. \sum\nolimits_{K=1}^{K} p\left(k\right) = 1 \tag{20}$$

Finally, the Lagrangian function is constructed by equation (21) to obtain the classification probability of the model for various types of modulated signals.

$$p(k) = \frac{\sqrt{\prod_{m} p_{\theta^m}(k)}}{\sum_{j}^{10} \sqrt{\prod_{m} p_{\theta^m}(j)}} \tag{21}$$

The experimental results show that the recognition effect after feature fusion is significantly higher than the method without feature fusion.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATASET

1) THE GAF IMAGE DATA

The RadioML(RML) 2016.10a dataset was used. The dataset was generated based on the GNU Radio environment and contained 11 modulation signals, each modulation signal includes 20 SNRs. Each SNR has 1000 samples, each sample has two signals: I and Q, and each signal contains 128 samples. The modulated signal volume in RadioML2016.10a is 220,000, including signals with SNR ranging from -20dB to 18dB. Complex channel simulations were performed to resemble the natural channel environment.

The datasets were converted into images using the GAF transformation method, and the training, validation, and test sets were randomly selected in a 7:2:1 ratio under each SNR. All experiments were implemented using an architecture with a Pytorch backbone and trained on a computer with an NVIDIA RTX 3080 GPU. The training phase was set to 100 stages for the first experiment and 200 for the second round of experiments.

In the training of the neural network on the signal transformed image, the epochs are set at 100 or 200 in the beginning. In order to keep the loss function decreasing, the learning rate is multiplied by 0.01 to improve the accuracy of network recognition if there is no decrease in the training set period, and the loss function is kept decreasing. If the loss does not decrease within the set period, training will be stopped. As in standard neural network training, the recognition accuracy of the network is calculated by comparing the actual labels of the signal categories with the predicted labels during the training process.

First, the RML2016a file is converted into a file in .mat format, the original file format of RML2016 is .pkl format. After the conversion is completed, use Matlab to open the mat format file so that you can see the specific values of I-Q two-way signals in each signal at different signal-to-noise ratios and convert the values in each signal into .csv file format. 1000 × 2 × 128 mat format in RML2016 signal, each signal has a total of 128 groups at one signal-to-noise ratio, and each group of signals has 1000 sampling points, divided into two signals. Each signal produces a total of 128 pictures under each signal-to-noise ratio. The signal-to-noise ratio range in the dataset is (-20,18), which means that there are a total of 4864 pictures for each signal, and the RML2016a dataset is a total of 11 signals with a total of 53504 images.

The GAF conversion method requires csv files to be converted so that RML 2016 can be converted into images and then trained with a neural network to detect the images. Figure 6 shows the time domain waveforms in the RadioML 2016.10A dataset at SNR=10dB.

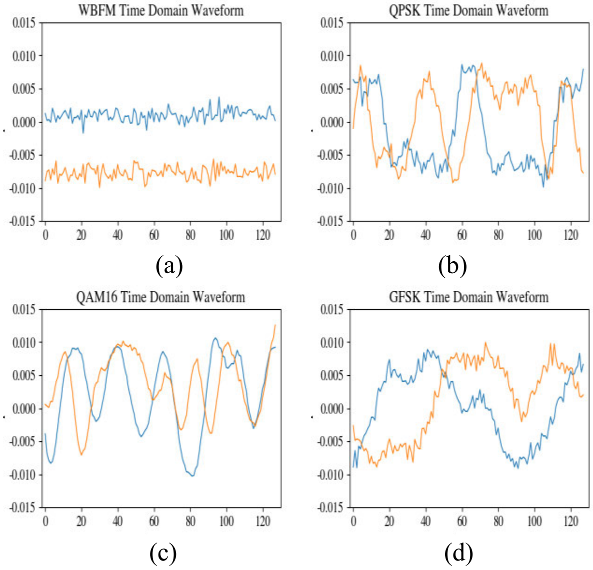


FIGURE 6. Selected signal waveforms of the RML2016.10A dataset. (a)The time domain waveform of WBFM. (b)The time domain waveform of QPSK. (c)The time domain waveform of QAM16. (d)The time domain waveform of GFSK.

When SNR = 12 convert the RML2016a dataset into GAF images as in Figure 7.

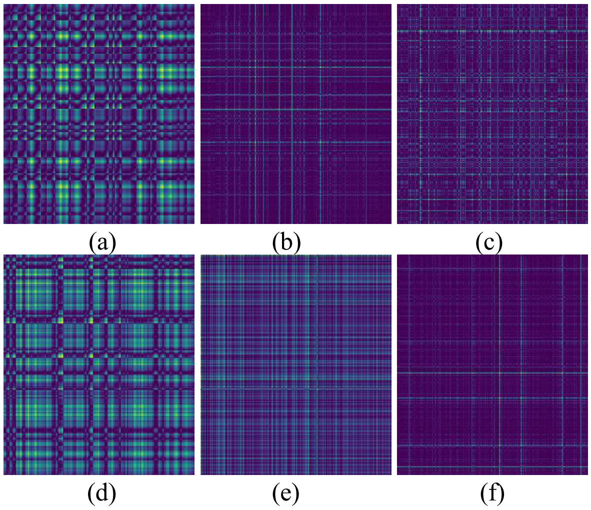


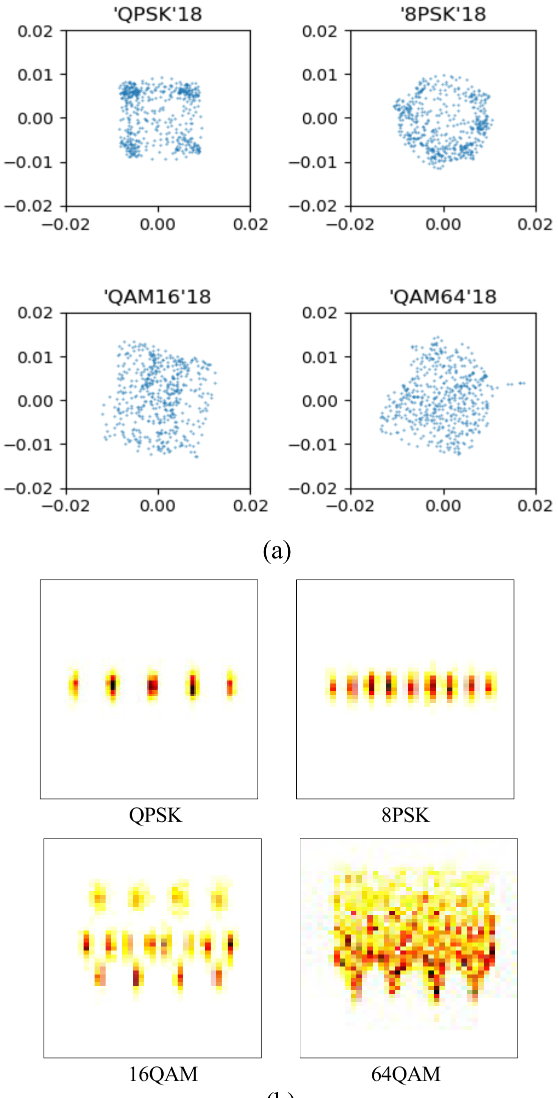
FIGURE 7. Image of data after GAF conversion. (a)The AM-DSB image. (b)The BPSK image. (c)The QAM16 image. (d)The WBFM image. (e)The QPAK image. (f)The PAM4 image.

2) ACCUMULATED POLAR FEATURE CONVERSION

Accumulated Polar Feature plots are first projected from the I-Q data into cumulative polar coordinates. Signal parameters were set as follows: sampling frequency and carrier frequency were the same as in the RML2016 dataset, channel environment was additive Gaussian white noise, selective fading (Rice and Rayleigh distributions), center frequency offset, sample rate offset, and each example contained 512 samples. All signal sources were created in Matlab 2020b. For training, 1000 sets were generated for each modulation type under the same SNR conditions. The process of developing the signals was repeated seven times, where the SNR ranged from −20 to 18 dB with 2 dB intervals.

Using Matlab to create the dataset, a similar RF dataset was generated using Matlab's Communication Toolbox that has the same modulation type in radioML, this time with 2 × 1024 points per data sample, but still roughly the same sign rate, with 1000 samples each (mod, snr).The general workflow for generating I/Q samples is shown below. Raw information: Generate random symbols from 0 to M-1 in a uniformly distributed manner. Convert to I/Q complex form using the built-in modulation function in Matlab Communication Toolbox. Up-sample each symbol to obtain 8 samples. Have a rising cosine filter. Apply channel effects. Normalize and extract 1024 samples.

Figure 8 shows the cumulative plot of the polar domain constellations in the dataset when SNR=18dB.



(a)

QPSK 8PSK

16QAM 64QAM

(b)

FIGURE 8. (a)The RML2016 constellation diagram. (b) Image of data after Accumulated Polar Feature conversion.

B. CLASSIFY SIGNALS USING FOUR NETWORKS

1) USE TWO TYPES OF CONVOLUTIONAL NETWORKS AND TWO TYPES OF TRANSFORMER NETWORKS

ResNet and ConvNeXt have the following network parameters, size = 224, in_channels = 1024, batch_size = 32, Optimizer = SGD, weight_decay = 1e-4, lr = 0.1*32/256. SwinT and TwinsT have the following network parameters, size = 224, in_channels = 768, batch_size = 32, Optimizer = AdamW, weight_decay = 0.005, lr = 5e-4*32/64. Four network models were used to train and evaluate 11 signals at SNR = 18, epoches = 100, two convolutional neural networks, and two Transformer visual neural networks. Initially, four networks are used to train and predict the ten signals in the dataset.

The above pictures and the table clearly show the recognition effect of the neural network after the fusion of the two images. It is easy to see that for AM-DSB, four networks are able to recognize accurately with a recognition rate of up to 100%, followed by a higher recognition of the PAM4 signal and a lower recognition rate of the signals QPSK and CPFSK respectively.

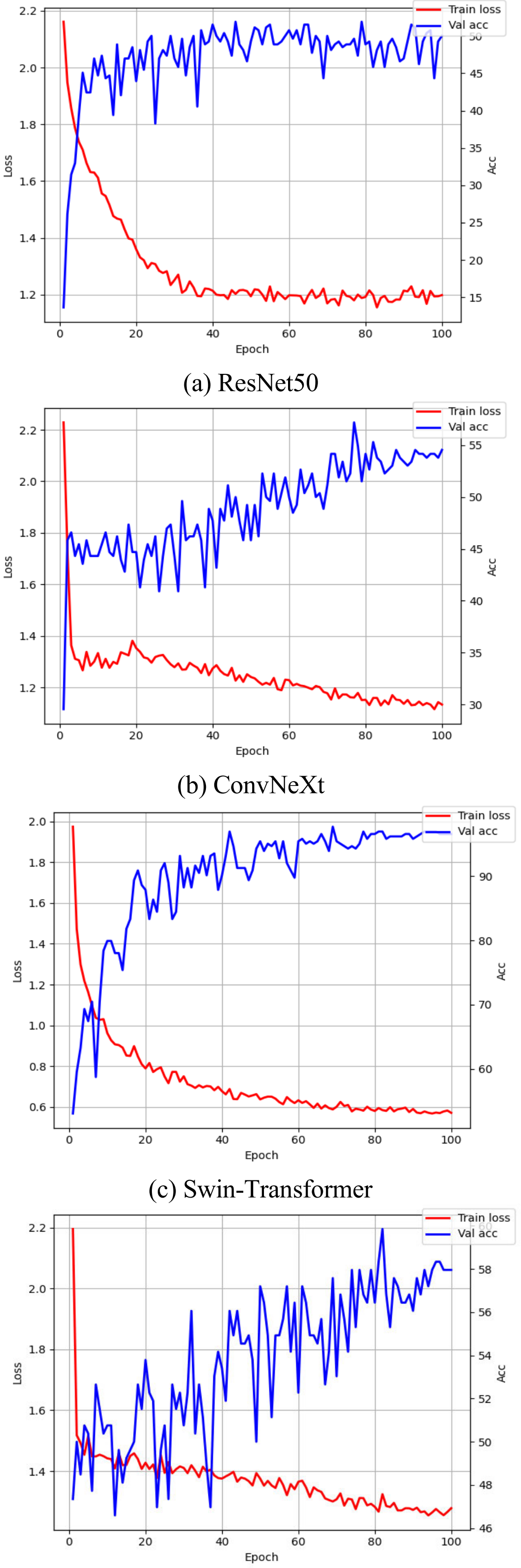
It can be found that Swin-Transformer network model has the best recognition effect. Basically, each signal recognition rate can reach more than 94%, and the multiple signals recognition rate is 100%. In the Twins-T network, three signals have the highest recognition rate, while the QPSK signal has the lowest recognition rate among the four networks, which can be explained by the dataset, whose images are similar to each other, and the main error is the misidentification of QPSK as BPSK.

Classification accuracy improves significantly with the increase of signal-to-noise ratio and achieves similar performance to other methods at high signal-to-noise ratio. Moreover, convolutional recognition results are considerably lower than Transformer vision network structure.

2) USE TWO TYPES OF CONVOLUTIONAL NETWORKS AND TWO TYPES OF TRANSFORMER NETWORKS

Training and evaluation of 11 signals using four network models at SNR = 10, epochs = 200, two convolutional neural networks and two Transformer visual neural networks.

The above graph shows that after more training times than in the previous experiment, the Swin-Transformer network



(a) ResNet50

(b) ConvNeXt

(c) Swin-Transformer

(d) Twins-Transformer

FIGURE 9. Test accuracy and train loss values when Epoch equals 100.

TABLE 1. Prediction accuracy of each signal.

Classes	ResNet Accuracy	ConvNeXt Accuracy	Swin T Accuracy	Twins T Accuracy
8PSK	50%	70%	98.46%	62.5%
AM-DSB	100%	100%	100%	100%
AM-SSB	97.34%	95.48%	98.72%	95.86%
BPSK	60.53%	62.16%	96.88%	72.5%
CPFSK	44%	33.33%	98.04%	40%
GFSK	24%	31.51%	100%	36.36%
PAM4	96.67%	82.35%	98.41%	100%
QAM16	59.09%	74.07%	98.41%	54.84%
QAM64	43.33%	53.33%	95.24%	51.85%
QPSK	24%	33.33%	94.12%	11.76%
WBFM	20.53%	27.59%	46.58%	43.55%

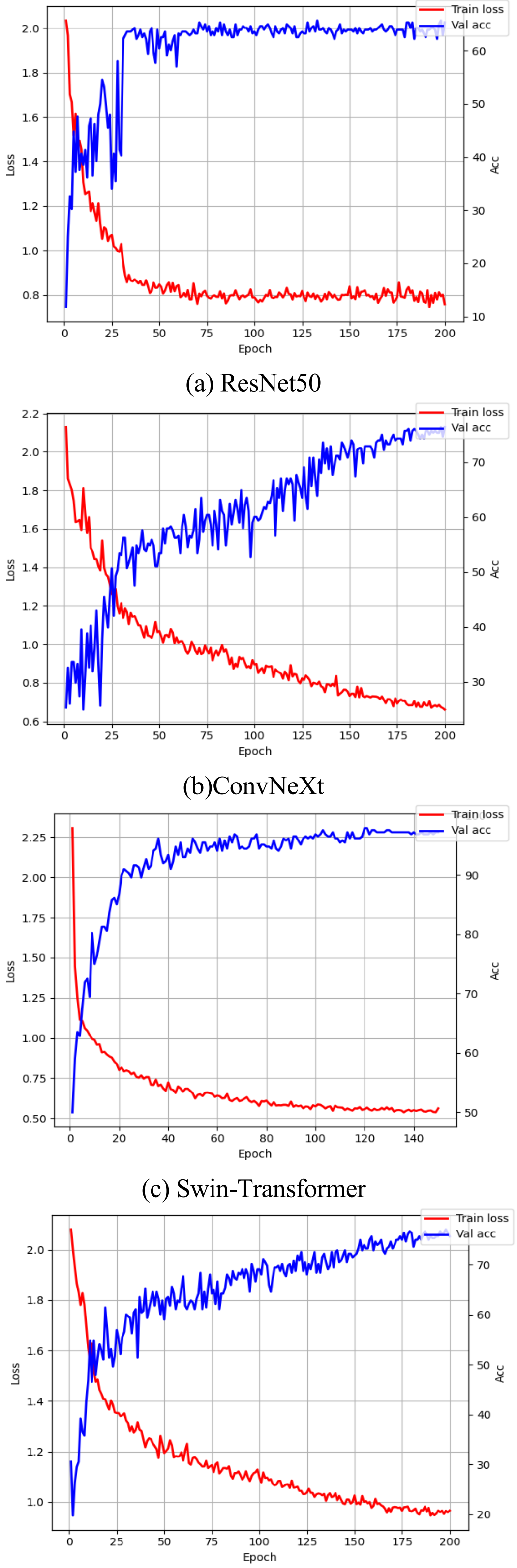
model is still found to perform better, and the recognition effect of ResNet and ConvNeXt is also better than in the previous round of recognition, but it is not as apparent as Swin-Transformer, and Swin-Transformer only uses small size and did not use a larger base or large size. The fusion mechanism is beneficial for the feature extraction of the network model. Except for SwinT, which classifies every signal very well, the remaining three networks are less effective in classifying 8PSK, especially ResNet50, a network structure where the recognition accuracy is only 43.64%. Considering that the size of the SwinT network is larger than the other three networks, it can be inferred that it is the other three networks that do not extract the features of 8PSK deeply enough. It is unrealistic to increase the number of network layers to improve the classification accuracy, and the hardware memory of the signal receiver is minimal.

C. CLASSIFICATION EFFECTS OF SEVEN TRANSFORMER NETWORKS

Since the Transformer optical network performs well, the next step is identifying the signals in the dataset with different Transformer network models at SNR=6. It can be seen that the Swin-Transformer network is still the best performer, and the recognition results are shown in the figure below.

During the experiments, it was found that the recognition effect of the Transformer network is significantly better than that of the convolutional neural network, especially the Swin-Transformer, which is based on the idea of the ViT model and innovatively introduces the sliding window mechanism, allowing the model to learn the information across the window and by down sampling layers, it enables the model to handle super-resolution images, saving computational effort as well as being able to focus on global and local in-formation.

The sliding window operation has a very significant accuracy for feature extraction of both images, including non-overlapping local windows and overlapping cross windows. When computing image attention for GAF images or cumulative polar coordinate images, the computational effort can be



(a) ResNet50

(b)ConvNeXt

(c) Swin-Transformer

(d) Twins-Transformer

FIGURE 10. Test accuracy and train loss values when Epoch equals 200.

TABLE 2. Prediction accuracy of each signal.

Classes	ResNet Accuracy	ConvNeXt Accuracy	Swin T Accuracy	Twins T Accuracy
8PSK	43.64%	48.57%	100%	55.56%
AM-DSB	100%	100%	100%	100%
AM-SSB	100%	100%	100%	100%
BPSK	100%	100%	100%	100%
CPFSK	57.14%	73.08%	100%	83.33%
GFSK	58.97%	88.46%	100%	75%
PAM4	80.49%	91.67%	100%	89.47%
QAM16	37.78%	61.54%	97.14%	77.42%
QAM64	53.85%	82.14%	91.67%	63.83%
QPSK	91.67%	54.05%	97.22%	78.26%
WBFM	39.54%	38.57%	56.25%	44%

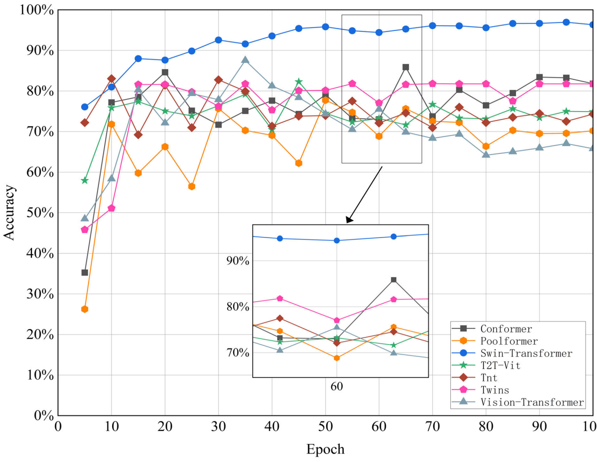


FIGURE 11. The average recognition accuracy of seven Transformer neural networks with SNR =6 for eleven signals after 100 epochs of neural network training.

significantly reduced by restricting to a window of fixed size. The hierarchical design of down sampling gradually increases the perceptual field, thus allowing the attention mechanism to notice global features.

The whole model adopts a hierarchical design, containing four stages, except for the first stage. Each stage will first reduce the resolution of the input feature map through the PatchMerging layer and perform the down sampling operation to expand the field of perception layer by layer like CNN to obtain the global information. This measure has dramatically improved the recognition accuracy of the proposed GAF images and cumulative polar coordinate transformed images.

D. COMPARISON OF THE AVERAGE RECOGNITION ACCURACY OF SWIN-TRANSFORMER AT EACH SIGNAL-TO-NOISE RATIO

In the overall framework of the Swin-Transformer network, firstly, the converted images of both signals are input to the PatchPartition layer for a chunking operation and then sent to the Linear-Embedding module for channel number channel adjustment. Finally, the final prediction results are

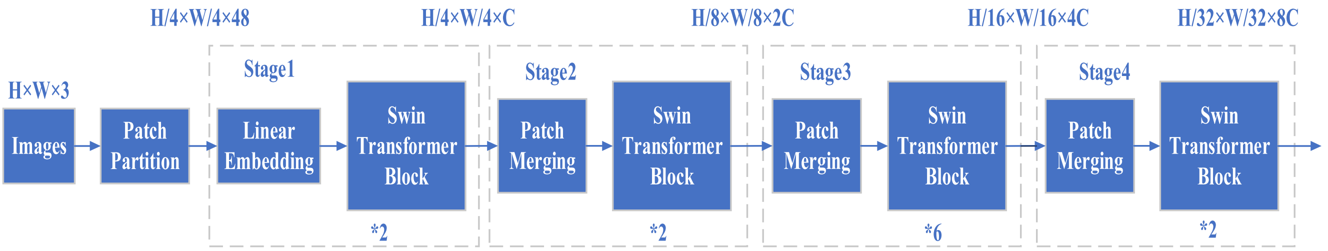


FIGURE 12. Structure of Swin-Transformer network.

obtained through the feature extraction and downsampling of Stage1,2,3,4, and the image size will be reduced to 1/2 of the original size, and the channel will be expanded to twice the original size after each stage, similar to the convolutional neural network. The Swin-Transformer-Block in each Stage consists of two connected Transformer-Blocks, which contain two base networks, Window Multi-head Self-Attention (W-MSA) and Shifted Window Multi-head Self-Attention (SW-MSA), as shown in figure13. These two networks improve the computational performance through the Window mechanism of the visual Transformer and the ShiftedWindow mechanism. The whole Swin-Transformer-Block structure contains two Window-Attention, where the first Attention mechanism is also known as Attention operation within each Window, and the second Window-Attention models the global data to extract features.

In order to reduce the computational complexity, Swin-T divides the input image into non-overlapping windows, and then performs self-attention calculation in different windows, each window contains M×M patches. The W-MSA is divided into several fixed windows, and the pixels in each window can only be inner-product with other pixels in that window to obtain information, which greatly reduces the computational effort and improves the efficiency of the network.

In the SW-MSA module, although the W-MSA reduces the computational effort by dividing the windows, the accuracy of the network is affected by the fact that the windows cannot interact with each other to obtain more globally accurate information.

In order to realize the information interaction between different windows, the windows are slidable, the windows are offset to contain different pixel points, and then the W-MSA calculation is performed, and the results of the two W-MSA calculations are connected to combine the information contained in the pixel points of two different windows to realize the information common between the windows. The SW-MSA mechanism completes the MSA calculation of the pixel points of the offset window and realizes the information exchange between different windows, thus indirectly expanding the "field of perception" of the network and improving the efficiency of information utilization.

The core of the Transformer model is the self-attention mechanism, while the core of the CNN model is convolution and pooling. the Transformer model can learn the correlation between the data, and the images generated by transformation in the dataset RML2016A used in this paper have a strong correlation between the before and after information, and in this particular aspect Swin Transformer is more effective than the normal ResNet. The CLDNN network model is more accurate than CNN or LSTM alone in modulated signal classification using deep learning because the advantages of the latter two networks are fitted together, and Swin Transformer achieves excellent results by combining the advantages of both CNN and Transformer.

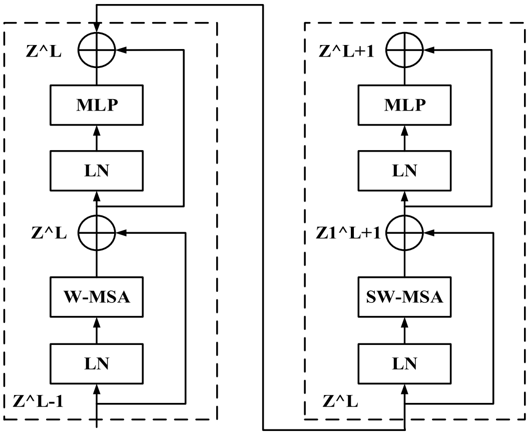


FIGURE 13. Swin-Transformer blocks.

As shown in figure 14, all signal-to-noise ratios of signals in the entire dataset are classified using Swin-Transformer. The results show that the classification accuracy can reach more than 90% at more excellent signal-to-noise ratios than 2 dB. Using the method proposed in this paper can bring the same results as the mainstream methods. However, one of the drawbacks is the large size of the network model, the ample memory space it occupies, and the long classification time of the extensive network model. We have used simple neural networks such as MobileNet and MobileVIT to classify the proposed method, but the results are poor. Maintaining accuracy while keeping the network size small is a problem that we will continue to explore.

The classification accuracy of the Swin-Transformer model was tested using validation samples. 92% accuracy was achieved at SNR=18, and 87% accuracy was achieved at SNR=0, which is still a good performance at a low signal-to-noise ratio. However, the recognition accuracy of WBFM signals is still lacking, and the classification accuracy of this

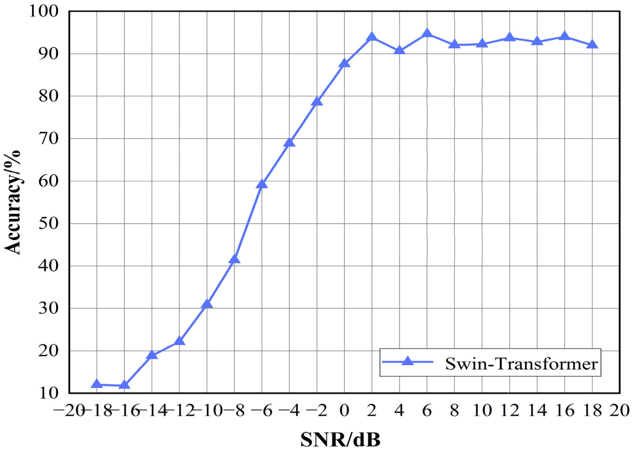


FIGURE 14. The average recognition accuracy of Swin-Transformer for eleven signals with different signal-to-noise ratios.

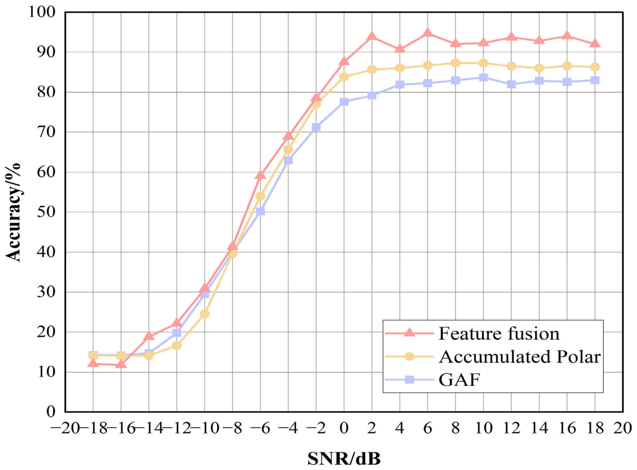


FIGURE 15. Comparison of the average recognition accuracy of eleven signals using feature fusion and without feature fusion.

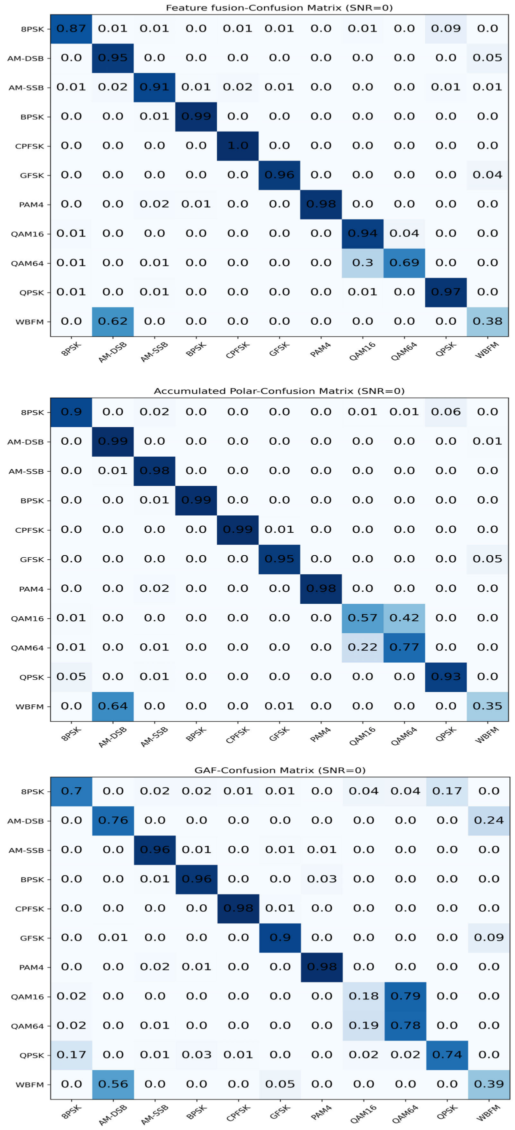


FIGURE 16. Confusion matrix of three methods.

signal type will be improved separately in the subsequent work.

E. PROVE THE EFFECTIVENESS OF FEATURE FUSION

The model proposed in this paper outperforms the other two models, especially at low SNR. Considering the poor signal transmission conditions in real communication environments, it is more valuable to have high classification accuracy at low SNR states. We can also find that Swin-Transformer performs better than other networks in various types of signal classification, especially in DSB, GFSK, PAM4 and BPSK; that is because GAF images and cumulative polar coordinate contain more information than IQ sequences. Moreover, the above results further reflect that GAF is the advantage of GAF in modulated signal characterization.

Signal feature classification using Swin-Transformer for three methods. The figure shows that when SNR=0, the average recognition rate is only 77.59% for GAF images and 83.81% for cumulative polar charts. Still, when the two features are fused and the neural network is used to recognize them, the recognition rate is as high as 87.82%. When SNR=18, the average recognition rate for GAF images is 82.98%, while the average recognition rate for cumulative polar maps is 86.314%, and the recognition rate after the fusion of two features reaches 92.94%. Compared with the results of neural networks for single signal feature classification, the fused image feature classification using neural networks for both signal transformations significantly improves accuracy.

The prediction results of the three classification network models for various types of modulated signals can be visualized in Figure 16. Each column of the confusion matrix represents the true category and each row represents the predicted category. The results show that the Feature fusion classification model has high recognition accuracy for all types of signals and good robustness in a low signal-to-noise environment. The advantages of the proposed method are mainly reflected in two aspects: one is the complementary advantages achieved in the extraction of two image features; the other is the role of the feature fusion mechanism.

The image-based representation method represents the received signal as an image, combines the DL framework for automatic feature extraction, and converts the sequence signal recognition into recognition of the image. The method proposed in this paper fully considers the features of both images, and the experimental results show the value of the method in the field of AMC.

F. COMPARISON WITH EXISTING METHODS

In this section, a series of comparative experiments based on the dataset are performed on modulated signals to evaluate the performance of our method. The method using single features in AMC is analyzed and compared with the dual-mode fusion method proposed in this paper to demonstrate the advantages of feature fusion using GAF images and cumulative polar coordinate images. The fusion mechanism proposed in this paper has apparent advantages over the single-feature AMC method.

As shown in figure 17, the proposed method in this paper is compared with existing methods that perform feature extraction directly on sequences [32]. The proposed method has a clear advantage at low signal-to-noise ratios, with an average of 10% higher accuracy than the network with the worst recognition accuracy and a maximum recognition accuracy at high signal-to-noise ratios.

Sequence representation-based methods are common in AMC, although the sequence representation method directly performs I/Q sequence or AP sequence feature extraction and recognition on the received signal, which has the advantage of being computationally small and fast. However, the method proposed in this paper combines the features of both images to extract complex features of the signal, and the classification effect is better than that of the sequence representation-based method. Therefore, the sequence representation method requires a practical and reasonable CNN network or RNN network according to the signal characteristics. If the network structure is very simple or unreasonable, the modulation classification will obviously show a poor result. The performance of the method decreases at a low signal-to-noise ratio, so different representation methods need to be selected according to the noise environment.

As shown in figure 18, the proposed method is compared with the direct use of deep learning for constellation images. The proposed method has significant advantages in both low and high signal-to-noise ratios [33]. The method using constellation diagrams has existed for a long time, although it is simple, it has obvious drawbacks and weak anti-interference ability in the case of complex channel environments. Based on the constellation diagram representation, the received signal is represented as a constellation diagram and combined with the DL framework for automatic feature extraction. Converting sequence signal identification to constellation diagram has better performance than feature representation and sequence representation methods. However, using individual image methods alone also has limitations because hierarchical image information requires deeper and more complex CNN to achieve the feature extraction task.

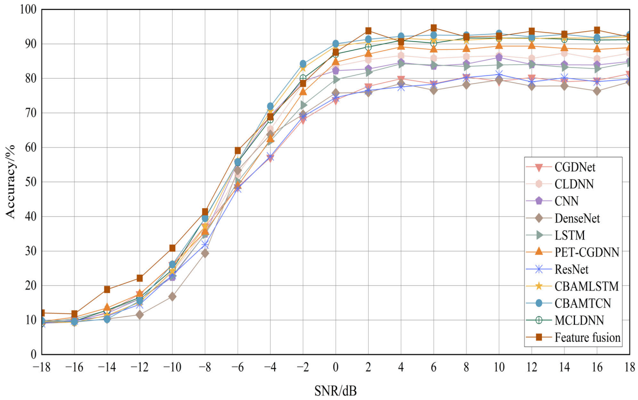


FIGURE 17. Comparison with sequence-based method.

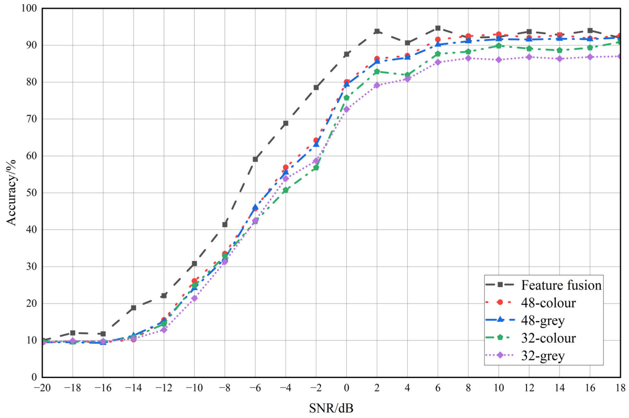


FIGURE 18. Comparison with constellation method. Converting I/Q time series to constellation charts, Generate four types of image datasets with different resolutions.

To solve these problems, feature fusion mechanism is designed to avoid the drawbacks of single feature representation. GAF image and cumulative polar coordinate image extraction features are implemented by Swin-Transformer to reduce the complexity of the network by adding penalty terms between the two features, and JS scatter is used to transform the classification problem mapping to a probability problem. Also, in training a regularization term is added to the loss function to avoid overfitting avoid overfitting to improve the classification model's convergence speed during training.

DL-based modulation classification algorithms represent raw signals in various formats. Obviously, the original modulated signal is represented by a combination of multiple features, images or sequences, which can integrate the advantages of various features and obtain better classification performance. In contrast, classification methods that use a fusion of features from different modalities have a clear advantage in the case of low signal-to-noise ratios. It also shows that the use of uncorrelated and different image combination features will improve the performance of AMC. The classification accuracy of the image feature fusion method is better than that of the sequence feature fusion method. The image feature fusion method combined with Swin-Transformer [34] has better classification performance than ResNet at low SNR. Therefore, it can be assumed that a deeper network structure will lead to better classification performance. In the method of [20], SPVD and BJD images are used for modulated signal characterization, with the disadvantage that the corresponding classification networks are not designed for different types of images.

In summary, the feature combinations need to be selected reasonably according to the fusion mechanism of AMC. The Transformer network has achieved remarkable results in deep learning and has contributed to many researches, and the Swin-Transformer network has also achieved many SOTAs machine vision. The feature fusion method proposed in this paper combines image features with significant differences and complementarities. Compared with other advanced methods, this method shows good robustness in different SNR environments and excellent performance in AMC classification accuracy.

V. CONCLUSION

In this paper, we propose to convert the received raw modulated signals into GAF and cumulative polar coordinate images, input these two images into a neural network for feature recognition, and use a feature fusion method to fuse the extracted features and in-put them into a neural network classifier for modulation classification. The results show that the Swin-Transformer network model has the highest classification accuracy and can achieve more than 90% recognition accuracy at high signal-to-noise ratios. By comparing the method that does not take fusion with the method that uses neural networks to recognize I/Q signals directly, the method proposed in this paper still has advantages.

The method proposed in this paper is valuable in terms of classification accuracy, but it is much slower in terms of detection time than the method that directly performs feature extraction on the signal. The next step is to explore a lightweight network that not only has high accuracy but also has a short detection time to classify signals. In addition, using deep learning is a good solution in the field of automatic modulation classification, how-ever the method requires a large number of high quality signal datasets, based on how to extend the dataset and how to improve the recognition performance by using small samples is a pressing issue.

REFERENCES

[1] O. A. Dobre, "Signal identification for emerging intelligent radios: Classical problems and new challenges," *IEEE Instrum. Meas. Mag.*, vol. 18, no. 2, pp. 11–18, Apr. 2015.

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

[3] J. Zhang, D. Cabric, F. Wang, and Z. Zhong, "Cooperative modulation classification for multipath fading channels via expectation-maximization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6698–6711, Oct. 2017.

[4] 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, Sep. 2018.

[5] O. Ozdemir, R. Li, and P. K. Varshney, "Hybrid maximum likelihood modulation classification using multiple radios," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1889–1892, Oct. 2013.

[6] S. Majhi, R. Gupta, W. Xiang, and S. Glisic, "Hierarchical hypothesis and feature-based blind modulation classification for linearly modulated signals," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11057–11069, Dec. 2017.

[7] T. Huynh-The, C.-H. Hua, T.-T. Ngo, and D.-S. Kim, "Image representation of pose-transition feature for 3D skeleton-based action recognition," *Inf. Sci.*, vol. 513, pp. 112–126, Mar. 2020.

[8] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, Apr. 2019.

[9] A. Iqbal, M.-L. Tham, and Y. C. Chang, "Double deep Q-network for power allocation in cloud radio access network," in *Proc. IEEE 3rd Int. Conf. Comput. Commun. Eng. Technol. (CCET)*, Beijing, China, Aug. 2020, pp. 272–277.

[10] A. Iqbal, M.-L. Tham, and Y. C. Chang, "Double deep Q-network-based energy-efficient resource allocation in cloud radio access network," *IEEE Access*, vol. 9, pp. 20440–20449, 2021.

[11] T. J O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," 2016, *arXiv:1602.04105*.

[12] Z. Liang, M. Tao, J. Xie, X. Yang, and L. Wang, "A radio signal recognition approach based on complex-valued CNN and self-attention mechanism," *IEEE Trans. Cognit. Commun. Netw.*, vol. 8, no. 3, pp. 1358–1373, Sep. 2022.

[13] N. E. West and T. O'Shea, "Deep architectures for modulation recognition," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Mar. 2017, pp. 1–6.

[14] G. J. Mendis, J. Wei, and A. Madanayake, "Deep learning-based automated modulation classification for cognitive radio," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Shenzhen, China, Dec. 2016, pp. 1–6.

[15] S. Peng, S. Sun, and Y.-D. Yao, "A survey of modulation classification using deep learning: Signal representation and data preprocessing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7020–7038, Dec. 2022.

[16] Q. You, J. Luo, H. Jin, and J. Yang, "Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia," in *Proc. 9th ACM Int. Conf. Web Search Data Mining*, San Francisco, CA, USA, Feb. 2016, pp. 13–22.

[17] Z. Zhang, Y. Zou, and C. Gan, "Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression," *Neurocomputing*, vol. 275, pp. 1407–1415, Jan. 2018.

[18] H. Wu, Y. Li, L. Zhou, and J. Meng, "Convolutional neural network and multi-feature fusion for automatic modulation classification," *Electron. Lett.*, vol. 55, no. 16, pp. 895–897, Aug. 2019.

[19] F. Wang, C. Yang, S. Huang, and H. Wang, "Automatic modulation classification based on joint feature map and convolutional neural network," *IET Radar, Sonar Navigat.*, vol. 13, no. 6, pp. 998–1003, Jun. 2019.

[20] Z. Zhang, C. Wang, C. Gan, S. Sun, and M. Wang, "Automatic modulation classification using convolutional neural network with features fusion of SPWVD and BJD," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, no. 3, pp. 469–478, Sep. 2019.

[21] M. Zhang, Y. Zeng, Z. Han, and Y. Gong, "Automatic modulation recognition using deep learning architectures," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Kalamata, Greece, Jun. 2018, pp. 1–5.

[22] S. M. Hiremath, S. Behura, S. Kedia, S. Deshmukh, and S. K. Patra, "Deep learning-based modulation classification using time and stockwell domain channeling," in *Proc. Nat. Conf. Commun. (NCC)*, Bengaluru, India, Feb. 2019, pp. 1–6.

[23] S. Peng, H. Jiang, H. Wang, H. Alwageed, and Y.-D. Yao, "Modulation classification using convolutional neural network based deep learning model," in *Proc. 26th Wireless Opt. Commun. Conf. (WOCC)*, Newark, NJ, USA, Apr. 2017, pp. 1–5.

[24] J. Bai, J. Yao, J. Qi, and L. Wang, "Electromagnetic modulation signal classification using dual-modal feature fusion CNN," *Entropy*, vol. 24, no. 5, p. 700, May 2022.

[25] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 1–7.

[26] C.-F. Teng, C.-Y. Chou, C.-H. Chen, and A.-Y. Wu, "Accumulated polar feature-based deep learning for efficient and lightweight automatic modulation classification with channel compensation mechanism," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15472–15485, Dec. 2020.

[27] A. S. L. O. Campanharo, M. I. Sirer, R. D. Malmgren, F. M. Ramos, and L. A. N. Amaral, "Duality between time series and networks," *PLoS ONE*, vol. 6, no. 8, Aug. 2011, Art. no. e23378.

[28] T. van Erven and P. Harremos, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.

[29] S. Claici, M. Yurochkin, S. Ghosh, and J. Solomon, "Model fusion with Kullback–Leibler divergence," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2038–2047.

[30] T. M. Sutter, I. Daunhawer, and J. E. Vogt, "Multimodal generative learning utilizing Jensen-Shannon-divergence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6100–6110.

[31] S. Xie, Y. Chen, S. Dong, and G. Zhang, "Risk assessment of an oil depot using the improved multi-sensor fusion approach based on the cloud model and the belief Jensen-Shannon divergence," *J. Loss Prevention Process Industries*, vol. 67, Sep. 2020, Art. no. 104214.

[32] F. Zhang, C. Luo, J. Xu, Y. Luo, and F. Zheng, "Deep learning based automatic modulation recognition: Models, datasets, and challenges," 2022, *arXiv:2207.09647*.

[33] Y. Wang, M. Liu, J. Yang, and G. Gui, "Data-driven deep learning for automatic modulation recognition in cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4074–4077, Apr. 2019.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10002.

WENXUAN MA received the B.S. degree in electronic information engineering from the Shandong Huayu University of Technology, in 2021. He is currently pursuing the master's degree in information and communication engineering with the School of Physics and Electronic Information, Yantai University, Yantai, China.

His research interests include deep learning methods, cognitive radio, and compressed sensing.

ZHUORAN CAI received the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology, in 2013.

He is currently an Associate Professor with the School of Physics and Electronic Information, Yantai University, Yantai, China. His research interests include deep learning methods, cognitive radio, and compressed sensing.